From weakly supervised learning to active labeling

Vivien Cabannes

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

From Weakly Supervised Learning to Active Labeling

Soutenue par
Vivien Cabannes
Le 18/07/2022

École doctorale n°386
Science Mathématiques

Spécialité
Apprentissage Statistique

Composition du jury :

Marc Lelarge INRIA / DI ENS / PSL	*Président*
Eyke Hüllermeier Paderborn University	*Rapporteur*
Guillaume Lecué ENSAE	*Rapporteur*
Joan Bruna NYU	*Examinateur*
Francis Bach INRIA / DI ENS / PSL	*Directeur de thèse*
Alessandro Rudi INRIA / DI ENS / PSL	*Directeur de thèse*

1794
ENS
ÉCOLE NORMALE
SUPÉRIEURE

PSL

Inria

À Kwama,

```
@PhDThesis{Cabannes2022,
  author    = {Vivien Cabannes},
  title     = {From Weakly Supervised Learning to Active Labeling},
  year      = 2022,
  school    = {\'Ecole Normale Sup\'erieure},
  type      = {PhD Thesis}
}
```

Contents

Chapter 1

Forewords

1.1 Acknowledgement

This thesis would not have been the same without the brilliant minds of Francis Bach and Alessandro Rudi, who have constantly impressed me by the speed at which they formed deep thoughts on any subject I would bring up in meetings. I consider myself really fortunate to have worked under your direction. Beside your sagacity, you have been a great source of inspiration as a person: always wise and caring, even when I was wondering about life in uncalled-for fashions, I was abusing academic freedom, or I was claiming false statements. My second thoughts go for my colleagues at INRIA Paris. I am particularly grateful to Alex Nowak for our many conversations around structured prediction, and to Yann Labbé for his continuous support. My last year as a PhD student has been highly enjoyable, which correlates with the arrival of Quentin Le Lidec in my office. Finally, I would like to warmly thank Guillaume Lecué and Eyke Hüllermeier for having accepted to review this manuscript, and especially considering how valuable your research hence your time is. In particular, Eyke, you were the first person to give me feedback on my work, which I have experienced as an accolade that marked my entry in the research world.

Une thèse de doctorat représente un accomplissement académique certain. Ma reconnaissance s'exprime également envers toutes les entités et les personnes extérieures qui m'ont permis d'y accéder. Tout d'abord tous les professeurs qui ont su me transmettre leur enthousiasme, en particulier Xavier Lacroix et ses merveilleuses digressions. Rétrospectivement, je m'étonne du dévouement et de la pédagogie remarquable d'un grand nombre d'entre eux, l'exemple le plus frappant étant celui d'Alain Camanès qui a su m'enseigner la rigueur nécessaire pour entrer dans le monde des sciences du supérieur. Enfin, mon parcours n'aurait pas été le même sans la bienveillance d'un certain nombre de personnes qui ont cru en moi plus que moi-même n'y croyait, me laissant nombre de vifs souvenirs. Remercions également les institutions qui font de leur mieux pour permettre ce genre d'entreprises, notamment en les finançant. En particulier, cette thèse a été financée par le gouvernement français via le programme « Investissements d'avenir » de l'Agence Nationale de la Recherche (référence ANR-19-P3IA-0001). Francis et Alessandro sont également soutenus financièrement par l'European Research Council (bourses SEQUOIA 724063 et REAL 94790).

Cet accomplissement s'inscrit également dans une histoire plus intime. Celle-ci s'écrit depuis bien avant ma naissance, et si je peux ici ressusciter les morts pour les remercier, je le ferai pour l'importance qu'ils ont porté à l'éveil, l'instruction et la curiosité, des valeurs qu'à leur suite, mon père et ma grand-mère maternelle ont tout particulièrement souhaite m'inculquer. Je dois également à ma famille de précieux sentiments, nourris par le souvenir, qui me fournissent souvent les aspirations nécessaires pour repartir quand la difficulté se fait trop présente en moi. J'espère qu'en retour, je participe un peu au bonheur et au développement de chacun. Si ce court texte ne permet pas de remercier personnellement toutes les personnes qui m'importent, je n'oublierai ni mon papy qui m'a hébergée pendant ces trois dernières années, ni ma mère qui s'est occupée de moi si longtemps. À la famille s'ajoutent mes amis. J'aimerai les remercier tous, qui sont des éléments constitutifs de ma personne. En particulier, je citerai Thomas Kerdreux pour sa magnanimité et Charles Arnal pour ses pertinents conseils. Tous deux ont une vigueur intellectuelle et physique qui ne cesse de m'impressionner.

1.1.1 Résumé

Les mathématiques appliquées et le calcul nourrissent beaucoup d'espoirs à la suite des succès récents de l'apprentissage supervisé. Dans l'industrie, beaucoup d'ingénieurs cherchent à remplacer leurs anciens paradigmes de pensée par l'apprentissage machine. Étonnamment, ces ingénieurs passent plus de temps à collecter, annoter et nettoyer des données qu'à raffiner des modèles. Ce phénomène motive la problématique de cette thèse: peut-on définir un cadre théorique plus général que l'apprentissage supervisé pour apprendre grâce à des données hétérogènes? Cette question est abordée via le concept de supervision faible, faisant l'hypothèse que le problème que posent les données est leur annotation. On modélise la supervision faible comme l'accès, pour une entrée donnée, non pas d'une sortie claire, mais d'un ensemble de sorties potentielles. On plaide pour l'adoption d'une perspective « optimiste » et l'apprentissage d'une fonction qui vérifie la plupart des observations. Cette perspective nous permet de définir un principe pour lever l'ambiguïté des informations faibles. On discute également de l'importance d'incorporer des techniques sans supervision d'appréhension des données d'entrée dans notre théorie, en particulier de compréhension de la variété sous-jacente via des techniques de diffusion, pour lesquelles on propose un algorithme réaliste afin d'éviter le fléau de la dimension, à l'inverse de ce qui existait jusqu'alors. Enfin, nous nous attaquons à la question de collecte active d'informations faibles, définissant le problème de « catalogage en ligne », où un intendant doit acquérir une maximum d'informations fiables sur ses données sous une contrainte de budget. Entre autres, nous tirons parti du fait que pour obtenir un gradient stochastique et effectuer une descente de gradient, il n'y a pas besoin de supervision totale.

1.1.2 Abstract

Applied mathematics and machine computations have raised a lot of hope since the recent success of supervised learning. Many practitioners in industries have been trying to switch from their old paradigms to machine learning. Interestingly, those data scientists spend more time scrapping, annotating and cleaning data than fine-tuning models. This thesis is motivated by the following question: can we derive a more generic framework than the one of supervised learning in order to learn from clutter data? This question is approached through the lens of weakly supervised learning, assuming that the bottleneck of data collection lies in annotation. We model weak supervision as giving, rather than a unique target, a set of target candidates. We argue that one should look for an "optimistic" function that matches most of the observations. This allows us to derive a principle to disambiguate partial labels. We also discuss the advantage to incorporate unsupervised learning techniques into our framework, in particular manifold regularization approached through diffusion techniques, for which we derived a new algorithm that scales better with input dimension then the baseline method. Finally, we switch from passive to active weakly supervised learning, introducing the "active labeling" framework, in which a practitioner can query weak information about chosen data. Among others, we leverage the fact that one does not need full information to access stochastic gradients and perform stochastic gradient descent.

1.2 Summary of contributions

1. The main contribution of this thesis is to provide a generic framework to deal with partial supervision. Along this framework, we provide a consistent algorithm for any structured prediction problem (Cabannes et al., 2020b); a principled algorithm to disambiguate weak information into full supervision which go beyond the partial supervision setting (Cabannes et al., 2021b), as well as a statistically and computationally efficient way to incorporate Laplacian regularization (Cabannes et al., 2021a).
2. We provide a method to derive fast rates for any discrete output problem (Cabannes et al., 2021c) based on least-squares surrogate. Although geared toward least-squares surrogate with Tikhonov regularization, those derivations can easily be extended to self-concordant losses based on Marteau-Ferey et al. (2019) and to any spectral filtering techniques based on Lin et al. (2020). We hope that this could be useful for statistical learning people, and we wish to see in the future similar results for other losses, as well as a better theory to compare surrogate problems for classification. As a first step in this direction, we provide some derivations for the hinge loss (Cabannes and Vigogna, 2022).
3. We provide a statistically and computationally efficient method to approach Laplacian regularization (Cabannes et al., 2021a). Our method could be useful for a vast number of applications. Our low-rank approximation could help scale up exciting methods for sampling based on Langevin dynamics

(Pillaud-Vivien, 2020a), Gaussian processes with derivatives (Eriksson et al., 2018), sparse models based on regularization with derivatives (Rosasco et al., 2013). The fact that we "kernelized" Laplacian regularization, making it statistically superior to existing local averaging methods, is exciting when considering the impact of "local" methods such as diffusion maps (Coifman and Lafon, 2006).

4. We introduce the active labeling problem, which unify several problems of searching for information through imprecise query, and might be useful to deal with privacy issues. We provide a stochastic gradient descent method that does not require full supervision to tackle this active weakly supervised learning problem that we formalized in Cabannes et al. (2022).

1.2.1 List of publications

On supervised learning, reproduced in Part II.

- (Cabannes et al., 2021c) Vivien Cabannes, Alessandro Rudi, and Francis Bach. Fast rates in structured prediction. In *Conference on Learning Theory*, 2021.
- (Cabannes and Vigogna, 2022) Vivien Cabannes and Stefano Vigogna. A case of exponential convergence rates for SVM. *In preparation* 2022.

On partial supervision, reproduced in Parts III and IV.

- (Cabannes et al., 2020b) Vivien Cabannes, Alessandro Rudi, and Francis Bach. Structured prediction with partial labeling through the infimum loss. In *International Conference on Machine Learning*, 2020.
- (Cabannes et al., 2021a) Vivien Cabannes, Alessandro Rudi, and Francis Bach. Disambiguation of weak supervision with exponential convergence rates. In *International Conference on Machine Learning*, 2021.
- (Cabannes et al., 2021a) Vivien Cabannes, Loucas Pillaud-Vivien, Francis Bach, and Alessandro Rudi. Overcoming the curse of dimensionality with Laplacian regularization in semi-supervised learning. In *Neural Information Processing Systems*, 2021.
- (Cabannes et al., 2022) Vivien Cabannes, Francis Bach, Vianney Perchet, and Alessandro Rudi. Active labeling: streaming stochastic gradients. *In preparation* 2022.

Some artistic reflections formatted for workshops (not reproduced).

- (Cabannes et al., 2019) Vivien Cabannes, Thomas Kerdreux, Louis Thiry, and Tina & Charly. Dialog on a canvas with a machine. In *NeurIPS workshop on Creativity*, 2019.
- (Cabannes et al., 2020a) Vivien Cabannes, Thomas Kerdreux, Louis Thiry. Diptychs of human and machine perception. *In NeurIPS workshop on Creativity*, 2020.

Side works done for French reviews (not reproduced).

- Vivien Cabannes. Perquisition de données sur le cloud : les Etats-Unis en avance, l'Europe à la traîne. In *le Grand Continent*, 2019.
- (Cabannes, 2022) Vivien Cabannes. Le futur du numérique sera-t-il incarné ? *Esprit*, 487:117-125, 2022.

1.3 Research chronology

This section traces back some of my research history, before summarizing the contributions of this thesis. I wish it to be useful in order to understand the thought process behind this thesis and to ease its reading.

The thesis originated from the motivations of my supervisors to confront the interest of Francis for weak supervision with the least-squares surrogate method of Alessandro. My interest was sparked by the fact that I wondered how to better incorporate fuzzy human knowledge into learning processes. Many abstractions can be learned in a hierarchical fashion, *e.g.* first recognizing edges on input images, then aggregating edges into tires, doors, handles, and those last components into cars, fridges, bridges, before understanding the semantic of an image and outputting a wanted label. Often, we have a good understanding of this hierarchical structure, but it is hard to formulate it clearly in order to guide the learning process. In this thesis, we focus on weak supervision, which is understood as specifying crucial elements in the last layer of abstraction before labels, such as specifying some animal attributes that allow us to discriminate the animals that our algorithm should learn to recognize. Other priors on the problem to solve are assumed to be incorporated into the model or into regularizers. Note that priors are distinct from supervision in the sense that they are given once and for all, while supervision can be refined (*e.g.* samples $(X_i, Y_i)_{i \leq n}$ are a function of $n \in \mathbb{N}$ that can be made larger).

From a theoretical point of view, successes of machine learning are understood as successes of supervised learning, which are successes of statistics when given enough clean data. Yet, in many problems, accessing tidy data is too costly, so many heuristics have been developed to deal with settings that do not completely fit in the supervised learning setting. The goal of this thesis is to ground those heuristics into principles. To make those principles clear and rigorous, this work finds its roots in the statistical learning theory. In particular, we assume that there is a unique metric to quantify the quality of a learning algorithm, and that this metric is specified by a loss function.

The first task was to define the setup to work on. Many directions could have been taken. For example, supervision can come as experts specifying what is their guess for the label, and what is their levels of certitude. To avoid dealing with fuzziness, we decided to assume that our supervision provides sure information that allows us to locate the label into a set of label candidates, assuming that errors on the supervision could be captured by the randomness of the label conditionally to the input. This generic yet interesting framework turns out to have already been studied in the literature under the name of partial labeling as well as superset learning. In a first paper (Cabannes et al., 2020b), we described this framework, advocate for combining weak information in a parsimonious way and came up with some theoretical results to strengthen our point. We put on top the framework of Alessandro to exhibit a consistent algorithm with decent convergence rates. While we initially wanted to compare this principle with several heuristics and associated principles, we soon realized that this task was endless, and we reduced the experimental comparison to "comparable" principles.

Several indications point us toward a second paper (Cabannes et al., 2021b). First of all, the algorithm we came up with was dealing with a really big related surrogate problem, hard to understand, hiding in ugly constants in our convergence rates. Fundamentally, we were dealing with functions from input to set of labels, and were paying the price of the combinatorial structure of power sets. To stay in the space of functions from input to labels, we decided to explicitly retrieve labels from weak supervision before learning with a fully supervised dataset. To come up with a clear empirical objective, we thought in terms of distributions rather than functions, leveraging the kernel mean embedding structure beyond Alessandro framework. However, theoretical guarantees were much harder to obtain, missing generic tools to understand deviation between empirical principle and population principle when these principles are expressed on distributions. This forces us to incorporate assumptions to ease the theoretical understanding. Surprisingly, those assumptions were not that strong and led to much better convergence rates, and this was not only the case for our problem but for the whole framework of Alessandro. We published this interesting result in a different paper (Cabannes et al., 2021c), and left open research questions on how far we can push those derivations. We came back to those questions recently, extending the proof mechanism to support vector machines (Cabannes and Vigogna, 2022).

In the first two papers, we built on combining weak information conditionally to a given input. While relation between inputs were captured by the model of functions we learn, the model we used was not smartly leveraging the input distribution. As such, we wanted to strengthen our principle to deal with settings akin to semi-supervised learning. In particular, we wanted to incorporate Laplacian regularization into our framework. Surprisingly, there were no good "kernelized" methods for Laplacian regularization but only "local averaging" methods. Benefiting from the expertise of my supervisors on kernel methods, as well as exchanging with Loucas Pillaud-Vivien, who already realized this fact, we came up with a paper on the matter (Cabannes et al., 2021a). Arguably, this gave a satisfying final piece to the theoretical framework built in this thesis in order to learn from partial supervision.

From there were several natural continuations. Deepen the study of partial supervision by understanding and incorporating more heuristics into our frameworks, such as collaborative filtering. Refine our algorithms with refinement of the structured prediction framework of Alessandro, for example, what can we say about conditional random fields and partial supervision? could we simplify the link between the structure of the supervision and the structure in "structured prediction", as we have implicitly done in the first two papers applications and algorithms? Widen our scope to all weak supervision frameworks, or even all unclutter data problems (*e.g.* missing input data). Ultimately, we decided to go for dataset annotation.

The ultimate goal of this work is to be useful to the practitioner hustling with data. Indeed, our research was not motivated by the fact that many datasets come with weak supervision, but by the idea that weak supervision is easier to collect for the practitioner. As a consequence, if we suppose that the practitioner is annotating data, we might help him to efficiently provide the most useful supervision. This problem has deep links with active learning, sequential design, non i.i.d. statistics and search games; links that were attractive to me. Building on our previous framework, we came up with a simple model of annotation cost and looked at techniques to minimize this cost for a given level of probably approximately correct bound. In

dimension one, our atomic questions are exactly the gradient of the L^1 loss; this pushed us to realize that one does not need full supervision to acquire unbiased stochastic gradients (Cabannes et al., 2022). However, when dealing with highly structured prediction, naive stochastic gradient descent on continuous surrogates suffer from suboptimality issues. Hence, in order to have a finer control on the samples we collected, we are currently switching to a bandit setting, for which, with the help of Vianney Perchet, we hope to come up with creative solutions.

Ethical considerations

This work aims at advancing our understanding of weakly supervised learning. Weakly supervised learning enrolls in the quest of an automated artificial intelligence, free from the need of human supervision. Automation, which is at the basis of computer science (Turing, 1950), is known to increase productivity at a reduced human labor cost, and is associated with several political/societal issues.

Part I

Introduction

Chapter 2

The Machine Learning Paradigm

In 2021, the International Data Corporation estimates that by 2025, the yearly global spending on artificial intelligence (AI) will overrun $204 billions (Glennon and Shirer, 2021). To give an order of magnitude, this corresponds to the gross domestic product of countries such as Greece, New Zealand or Peru.[1] It means that if this business was concentrated in Peru, its 33 millions inhabitants would only produce goods and services related to AI. Nowadays, the core engine behind AI is machine learning, that is the learning of rules and algorithms by computers. In this chapter, we provide a subjective exposition of what machine learning is about. It is written to be accessible for a large public beyond the machine learning community.

2.1 From algorithmics to machine learning

In this section, we present machine learning as the evolution of algorithmics.

In a prosaic fashion, an algorithm can be thought of as a cooking recipe. It is a sequence of instructions that, given some inputs, produces an output. For example, the inputs could be some apples, lard, eggs, salt and flour, the sequence of instructions be the one of an apple pie recipe, and the output be a delicious pie. This can be formalized mathematically, an algorithm being understood as a sequence of instruction I, leading to a mapping $f : \mathcal{X} \to \mathcal{Y}$, which, given an input x belonging to an input space $\mathcal{X}$, outputs $y = f(x)$ belonging to an output space $\mathcal{Y}$.

Example 1 (Cooking recipe)**.** *$\mathcal{X}$ is a set of potential ingredients, $\mathcal{Y}$ is a set of potential pies. For some ingredients x,* e.g. *x = (5 apples, 100g of lard, 200g of flour, 1 egg, 1 pinch of salt), $y = f(x)$ describes the pie obtained by following the instructions I of a recipe.*

Example 2 (Sorting algorithm)**.** *$\mathcal{X} = \mathbb{R}^n$ is the set of lists of size n, $\mathcal{Y} = \{(y_1, \cdots, y_n) \in \mathbb{R}^n \mid y_1 \leq \cdots \leq y_n\}$ is the set of sorted lists of size n. Any sorting algorithm corresponds to $f : \mathcal{X} \to \mathcal{Y}$ that maps a list $x = (x_i)_{i \leq n}$ to its sorted version $y = (x_{\sigma(i)})_{i \leq n} \in \mathcal{Y}$ with $\sigma \in \mathfrak{S}_n$ a permutation.*

Example 3 (House pricing)**.** *$\mathcal{X} = \mathbb{R}_+^2$ represents real estate properties with two features: the living space of the house, the outdoor space of the garden, both in meter square. $\mathcal{Y} = \mathbb{R}_+$ represents the price of the house in US dollars. The mapping f is a model for house prices.*

Example 4 (Classifying cats and dogs)**.** *$\mathcal{X} = [255]^{d_1 \times d_2 \times 3}$ is a set of images of cats and dogs. Those images are represented by $d_1 \times d_2$ pixels, with the three RGB channels. $\mathcal{Y} = \{-1, 1\}$ with $f(x) = 1$ if x is an image of cats, and $f(x) = -1$ if x is an image of dogs.*

The problematics arising in our examples are all slightly different and will be helpful to illustrate differences between algorithmics, statistics and machine learning.

Algorithmics. In classical computer science, algorithms are implemented on m bits of memories. As such, the sequence of instructions can be written as $I = (I_t)_{0 \leq t \leq T+1}$, with $I_0 : \mathcal{X} \to \{0,1\}^m$ an encoding of $\mathcal{X}$ on m bits, $I_{T+1} : \mathbb{R}^m \to \mathcal{Y}$ a decoding of m bits to $\mathcal{Y}$, and, for $t \in [T]$, $I_t : \{0,1\}^m \to \{0,1\}^m$ a basic bitwise operation. In problems such as sorting lists, we know exactly which f we want to achieve, but we do not

[1] According to the World Bank Open Data.

know how this mapping can be decomposed into a set of basic operations which could be implemented practically on a computer. In those algorithmic problems, an algorithm is thought of as the sequence of instructions I,and its quality is discussed under the light of two notions.

- Correctness: does this sequence of instructions correspond to the desired mapping f? In other terms, can we guarantee the equality $f = I_{T+1} \circ I_T \circ \cdots \circ I_0$?
- Complexity: how much basic operations T and computer memory m does it take to perform the full sequence of instruction I?

In the cooking example 1, the first question can be rephrased as: does our recipe, given a specification of ingredients, always lead to the same pie? The second question is related to the equipment and the time required by the recipe. When it comes to sorting lists, example 2, deriving and implementing a sequence of basic operations to solve this task is a reasonable idea. For example, think of how you sort a deck of cards, write rules about it, and implement those rules in your favorite programming language that will convert it into a sequence of bitwise instructions for the computer. When it comes to understanding images, deriving and implementing a sequence of basic operations to differentiate cats and dogs have not been proven a successful way to proceed. Indeed, even for language translation, efforts, made by linguists to describe translation as a set of simple syntactic rules, have not led to efficient translation algorithms.

Learning optimal sequence of instructions. Rather than designing an algorithm from a set of humanly-thought rules, machine learning suggests designing an algorithm by quantifying a notion of optimality, or equivalently a notion of error, cost or risk $\mathcal{R}$, and looking among all potential sequences of instructions $I \in \mathcal{I}$, in a class $\mathcal{I}$ of potential sequences of instruction, for the one that minimize the risk $\mathcal{R}(I)$. For example, in the cooking example, we could consider $\mathcal{I}$ as all potential combinations of basic instructions such as "breaking eggs", "mixing some ingredients", "frying/baking a batter" and so on. Optimality, and the risk $\mathcal{R}$, could be designed by some experts testing the pie, or through molecular gastronomy considerations. To find the best recipe, that is the best sequence of basic instructions, a naive approach consists in trying all potential combinations of basic instructions I, and quantifying the result quality through $\mathcal{R}(I)$, and considering the sequence I that achieves the minimal risk. Of course, there are zillions of ways to combine basic instructions, and baking zillions of pie, and having them tested by experts is not a viable option. But on a computer that can perform millions of operations by seconds, the picture is quite different. Applied to the list sorting problem, such a procedure leads to a completely different approach from the classic algorithmic one. Roughly speaking, the machine learning paradigm is to try all potential algorithms and take the best one, where "the best" is quantified through the risk $\mathcal{R}$.

2.2 The emergence of statistics

In this section, we get more specific on what machine learning is today, that is, statistics in the era of powerful computers.

While machine learning could be seen as the art of automatically designing algorithms, nowadays, it is rather a new take on statistics in the era of powerful computers. As such, it is less focused on having a "good" correct sequence of instruction I to perform a task, but rather in finding a mapping f that is optimal for a task, such as pricing houses or distinguishing cats and dogs. In such a setting, the risk $\mathcal{R} : \mathcal{Y}^{\mathcal{X}} \to \mathbb{R}_+$ associates a score $\mathcal{R}(f)$, quantifying a notion of risk, cost or error, to a function $f : \mathcal{X} \to \mathcal{Y}$. Ideally, we would like to retrieve an optimal mapping $f^* : \mathcal{X} \to \mathcal{Y}$ such that

$$\mathcal{R}(f^*) = \mathcal{R}^*, \qquad \text{where} \qquad \mathcal{R}^* = \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f). \tag{2.1}$$

The problematics of machine learning are slightly different from the ones of algorithmics. The quality of a procedure, that allows obtaining automatically a mapping $f : \mathcal{X} \to \mathcal{Y}$ to solve a specified task, is discussed under the light of three notions.

- Optimality: is the risk $\mathcal{R}(f)$ of the obtained mapping f close to $\mathcal{R}^*$?
- Inference complexity: given an input $x \in \mathcal{X}$, is it easy to compute $f(x)$?
- Training complexity: is it computationally easy to obtain the mapping f?

The first two issues are the direct translation of the algorithmic issues of correctness and complexity, while the third one is specific to machine learning, and has strong links with the field of optimization.

Remark 1 (Correctness and Optimality)**.** *In classical algorithms, there is the idea that an algorithm is perfect/correct, that we know exactly what it is doing. In machine learning, the idea of f being correct is dropped by the notion of f being optimal or quite optimal. For critical applications such as flying planes, replacing the notion of correctness brought in by classical algorithms, by a notion of quasi-optimality is not innocuous. Yet, once a mapping f is retrieved by an artificial intelligence system, it can be tested for correctness through formal proof management systems, or tested for statistical quality through an extensive number of testings. While replacing "being correct" by "being statistically good" can offend purists, note that, in medicine, drugs are more often approved based on statistical studies, rather than on a precise description of their mechanisms and a complete understanding of their interactions with the body.*

If we consider the cooking example 1, we could translate those three issues into prosaic terms. Suppose that you have a way to design new recipes. Optimality translates into "are the new dishes you make good?". Inference complexity translates into "are the recipes easy to reproduce?". Optimization complexity translates into "is your procedure for designing new recipes difficult, is it time and money consuming?".

2.2.1 Supervised learning

A simple idea to learn a mapping f between inputs x and outputs y is to consider a set of n examples $(x_i, y_i)_{i\leq n} \in (\mathcal{X}\times\mathcal{Y})^n$, where $f(x_i)$ should be y_i, and try to infer from those examples a general law between x and y. That is, for house pricing, example 3, consider n houses, report their characteristics $(x_i)_{i\leq n}$ and their prices $(y_i)_{i\leq n}$, and try to find a law that links house characteristics x to their prices y. In particular, a real estate agent can try to fit a linear model $y = w^\top x$, for $w \in \mathbb{R}^d$ a weight vector, the house characteristics x assumed to belong to $\mathcal{X} \subset \mathbb{R}^d$, and y the price of the house. It is usual to fit w by minimizing the least-squares error

$$\hat{w} \in \operatorname*{arg\,min}_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \left\|w^\top x_i - y_i\right\|^2 = (\sum_{i=1}^{n} x_i x_i^\top)^\dagger \sum_{i=1}^{n} y_i x_i,$$

where, for a vector $x \in \mathbb{R}^d = \mathbb{R}^{d\times 1}$, $x^\top \in \mathbb{R}^{1\times d}$ designs its transpose, and, for a matrix A, $A^\dagger$ its pseudo-inverse. When needing to price a new house, the real estate agent could report its characteristics x and price it at $y = \hat{w}^\top x$.

Such a model raises many questions. Which features x govern the price of a house? Can the price of the house be deduced as a linear combination of those features? How many houses n should I consider as training examples $(x_i, y_i)_{i\leq n}$ to retrieve a $\hat{w}$ such that the value $\hat{w}^\top x$ does reflect well the market price of a new house characterized by x? How much should we expect the transaction price of a house to deviate from our model of its market price? Those questions are addressed by statisticians and machine learning practitioners. They are respectively linked with features engineering, model selection, data collection, and generalization guarantees.

Perhaps the most well known result of statistics is the large law number. In essence, it tells you that if you repeat the same experiment, what you see as the result of this experiment, reflects what you are likely to see if you repeat this experiment one more time. For example, if you have rolled a die zillion times, and you have gotten six forty-five percent of the time, the next time you roll this die, you can expect to get a six with forty-five percent of chance. In the past two centuries, statisticians have developed many concentration inequalities that quantify, given the number of experiments you have run so far, how much the statistics collected from those experiments are likely to reflect the outcome of future experiments. This provides tools to give generalization guarantees on models learned from data. But beyond those results are some axioms on the experiments you run. In particular, statistics are grounded in probability theory, assuming that the experiment's outcomes are governed by some probabilistic process that is stable over time, allowing prediction of future outcomes from prior ones. Typically, when talking about dice, we assume that you roll the same dice, in the same fashion, in the same environment.

Learning theory, and in particular supervised learning, builds on this idea of understanding observable phenomena by assuming an underlying probabilistic model. This thesis is set in this framework. It assumes the existence of a joint probability distribution $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ over $\mathcal{X}$ and $\mathcal{Y}$, that have generated the dataset $\mathcal{D}_n = (X_i, Y_i)_{i\leq n} \sim \rho^{\otimes n}$ in an *independent identically distributed* fashion. Moreover, it assumes that the measure of error results from the integration of a pointwise measure of error $\ell(f(X), Y)$, between a prediction $f(X)$ made at the point X and its observed label Y, for $\ell \in \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ a loss function, in the sense that for $f : \mathcal{X} \to \mathcal{Y}$,

$$\mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\rho}\left[\ell(f(X), Y)\right]. \tag{2.2}$$

As one has no access to the distribution ρ but only to samples $(X_i, Y_i)_{i\leq n}$, it is natural to substitute the risk with its empirical counterpart $\mathcal{R}_{\mathcal{D}_n} : \mathcal{Y}^{\mathcal{X}} \to \mathbb{R}$, defined as

$$\mathcal{R}_{\mathcal{D}_n}(f) = \frac{1}{n}\sum_{i=1}^{n} \left[\ell(f(X_i), Y_i)\right]. \tag{2.3}$$

A popular approach to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ is to consider some model of functions $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ and to perform empirical risk minimization over this class of functions, leading to the estimate

$$f_n \in \arg\min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{D}_n}(f).$$

Yet, how optimal is this mapping f_n? In other terms, does it exhibit a low risk $\mathcal{R}(f)$? In the machine learning community, this question refers to generalization properties of f_n: does what we have learned based on n examples generalize to new situations characterized by new inputs? Statistical learning theory provides some answers to this question.

Remark 2 (Why probability?)**.** *Arguably, we live in a deterministic world. Hence, if everything has a reason, why introduce randomness in our model? Think about rolling dice, the face that pops up is a pure deterministic function of the landscape of the support the dice are rolled on, as well as the initial moment, impulsion and position of the dice, which themselves depend on the mood of the person rolling the dice. But as we cannot model those factors, we suppose them to be random, leading to the random distribution on the rolling dice, at the end of the causal chain from mood to initial physical values to dice output. The same applies to house price, many idiosyncratic factors come into play when a buying offer is made on a house. When tracking those explanation variables is too difficult, statisticians deal with this fact by adding randomness into the model.*

2.2.2 Generalization bounds

In this subsection, we mimic typical derivations provided by statistical learning theory which are used in this thesis. Recall the house pricing example 3. Supervised learning consists in assuming the existence of some abstract distribution $\rho \in \mathcal{X} \times \mathcal{Y}$, and that pairs of house characteristics and their prices $(x_i, y_i)_{i\leq n}$ are actually sampled from this distribution $(x_i, y_i) = (X_i, Y_i) \sim \rho^{\otimes n}$ – using capital letters to denote the randomness in the sampling. Trying to predict house prices from a linear combination of house characteristics leads to the model of functions $\mathcal{F} = \left\{x \to w^\top x \,\middle|\, w \in \mathbb{R}^d\right\}$, assuming $x \in \mathcal{X} \subset \mathbb{R}^d$. The best linear model, based on the mean-squares error, is defined by $f^*_{\text{lin}} : x \to x^\top w^*$, with the weighting scheme

$$w^* \in \arg\min_{w\in\mathbb{R}^d} \mathbb{E}_{(X,Y)\sim\rho}\left[\left\|X^\top w - Y\right\|^2\right] = \left(\mathbb{E}_{X\sim\rho_{\mathcal{X}}}\left[XX^\top\right]\right)^\dagger \mathbb{E}_{(X,Y)\sim\rho}\left[YX\right].$$

Here $\rho_{\mathcal{X}}$ designs the marginal of ρ over $\mathcal{X}$, that is, the distribution of X according to ρ. In practice, we do not have access to ρ but to the samples (X_i, Y_i). As a consequence, we can replace ρ by $\hat{\rho} = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}$, which leads to the empirical risk minimizer $f_n : x \to x^\top \hat{w}$ with

$$\hat{w} \in \arg\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \left\|X_i^\top w - Y_i\right\|^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top\right)^\dagger \left(\frac{1}{n}\sum_{i=1}^{n} Y_i X_i\right).$$

As the quality of a function f is measured through the risk $\mathcal{R}$, it is natural to measure the quality of f_n in terms of the excess of risk $\mathcal{R}(f_n) - \mathcal{R}^*$, with respect to the minimum achievable risk $\mathcal{R}^*$. This error can be split in the following way

$$\mathcal{R}(f_n) - \mathcal{R}^* = \underbrace{\mathcal{R}(f_n) - \mathcal{R}(f^*_{\text{lin}})}_{\text{estimation error}} + \underbrace{\mathcal{R}(f^*_{\text{lin}}) - \mathcal{R}^*}_{\text{approximation error}}. \tag{2.4}$$

This split separates an error due to the need for more data so that $\hat{w}$ estimates w^* correctly and a part due to the fact that the best linear predictor f^*_{lin} might not be the best pricing model. The decomposition with estimation and approximation errors, is sometimes called bias-variance decomposition. The variance, or

estimation error, is due to the randomness of the samples, while the bias, or approximation error, is due to the gap between the average (or best) estimate we expect and the solution.

The approximation error is due to the fact that our model might be inaccurate. This error can be controlled by assuming some sort of density of $\mathcal{F}$ in $\mathcal{X} \to \mathcal{Y}$ regarding the topology inherited from $\mathcal{R}$, or by assuming that the best predictor f^* does belong to our class of functions $\mathcal{F}$ (or is well approximated by it with respect to the topology associated with $\mathcal{R}$).

The estimation error is due to the fact that we have estimated a quantity defined through the distribution ρ thanks to sample $(X_i, Y_i)_{i \leq n}$. Similarly to rolling dice or tossing coins, when knowing ρ, statistics and probability can quantify, for a given number of training examples n, what is the distribution of $\mathcal{R}(f_n) - \mathcal{R}(f^*_{\text{lin}})$. This distribution should be seen as the pushforward of $\rho^{\otimes n}$ regarding the process that leads to the estimate f_n from samples $(X_i, Y_i)_{i \leq n}$. When not knowing ρ, this pushforward cannot be estimated, but with few hypotheses on ρ, such as control of extreme behaviors, or equivalently control of high moments, it is possible to control deviation of $\mathcal{R}(f_n)$. This is exactly what provide concentration inequalities, such as Hoeffding or Bernstein inequalities, that allows upper-bounding the deviations between the empirical means $\frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ and $\frac{1}{n}\sum_{i=1}^n Y_i X_i$ and their population counterpart $\mathbb{E}[XX^\top]$ and $\mathbb{E}[YX]$, and therefore the deviation between $\hat{w}$ and w^*.

With assumptions to control the approximation and estimation errors, it is possible to get convergence results of the type,

$$\forall\, t > 0, \qquad \mathbb{P}_{\mathcal{D}_n}(\mathcal{R}(f_n) - \mathcal{R}^* > t) \leq \exp(-\sigma^{-2} n t^2).$$

This specific result tells you that if you have trained your model with n independent identically distributed samples $(X_i, Y_i)_{i \leq n}$, the risk of the learned mapping f_n is comparable to a random number that follows a Gaussian distribution $\mathcal{N}(0, \sigma/\sqrt{n})$, the randomness being due to the randomness of the samples. Such a bound is called a generalization bound as it tells us what risk or error we can expect on generic examples, which is a measure of how well our predictor generalizes to new situations. When an algorithm benefits from such a bound, we call it probably approximately correct (Valiant, 2013), in the sense that with high probability (the randomness being due to the sample population), it approximately minimizes the risk. It is worth noting that, in practice, we only have one realization of $\mathcal{D}_n$, and that this result does not tell us that the risk $\mathcal{R}(f_n)$ is small, it only tells us that we would have been really unlucky if it was not the case.

2.2.3 Unsupervised learning

In the preceding sections, we have presented machine learning as concerned with the learning of a mapping from an input space $\mathcal{X}$ to an output space $\mathcal{Y}$. As such, it differs from statistics that are rather concerned with inferring properties of a distribution $\rho_{\mathcal{X}}$ from samples $(X_i)_{i \leq n}$. Indeed, many papers published as machine learning papers are concerned with this setting, usually referred to as unsupervised learning. There are two major motivations for unsupervised learning.

First, statistical problems such as density estimation, eventually done through maximum likelihood estimation, are instances of problems where we would like to infer properties of a distribution $\rho_{\mathcal{X}}$ through samples $(X_i)_{i \leq n}$. The usage of a computer, and of optimization routines, is crucial to deal with a large number of data in this setting. Indeed, many of the computational and statistical problematics arising from those problems are shared with standard supervised learning, justifying results on those problems to be published in machine learning conferences, and justifying the machine learning community to welcome statisticians.

Second, learning a mapping from inputs to outputs, highly benefits from features engineering, which might be done without access to the supervision $(Y_i)_{i \leq n}$. From a machine learning perspective, the main goal of unsupervised learning is to discover structure beyond input data, assuming that this structure will help to find laws that correlate those inputs to potential outputs. For example, when described as an array of pixels, images are understood as points in $\mathbb{R}^d$ for d of order 10^9. Yet, we expect natural images to have a strong structure that forbid them to be any points in $\mathbb{R}^d$. Indeed, we can expect natural images to be concentrated close to a low-dimensional manifold, potentially parametrized in $\mathbb{R}^m$ for m much smaller than d, describing the many parameters of freedom of a natural image. Finding such a small representation would ease "downstream" tasks, such as learning to recognize the content of an image. Having been approached with many perspectives, unsupervised learning has been divided into many sub-problems, such as clustering, manifold learning, sparse dictionary learning, or self-supervised learning.

2.3 A tour of problems

In this section, we review diverse practical problems that are arguably understood as machine learning problems. We review machine learning to create artificial intelligence with human-alike capability. We showcase business motivations and driving forces beyond machine learning. We explain how data science can fit into intelligent systems, and discuss how machine learning can help science and how its scope is growing behind statistical learning. We do not aim for completeness, but for simple examples.

2.3.1 Seeing, hearing and talking

For non-experts of machine learning, probably the most important result of this last decade was the result achieved by deep learning trained on GPU on the ImageNet challenge of 2012 (Krizhevsky et al., 2012). In essence, for the first time, a computer was able to showcase human-like performance when recognizing simple objects in images. Similarly, we can dictate to our electronic devices and have those devices writing down our thoughts as if they were secretaries. Moreover, those devices can emit sounds articulated into words and sentences to answer a question we have eventually asked. All as if machines were to see, hear and speak. This echoes a long fantasy of machines bearing signs of human intelligence.[2] Interestingly, this "artificial intelligence" replicates tasks that many animals do automatically and instantaneously, concerned with processing simple information. As such it is more concerned with intelligence as collecting pieces of information (such as "military intelligence"), rather than the more deep thought and contemplative process captured by the French notion of *intelligence*. Those recent successes are somewhat based on supervised learning, providing strong evidence of the usefulness of this theoretical framework, especially when compared to previous attempts to build computer vision, audio and natural language processing from so-called "expert systems". Solving those problems has required machine learning to scale with millions of training examples n, as well as millions of input dimensions d, which has strongly pushed to advance calculations engineering (see, *e.g.*, Chowdhery et al., 2022).

2.3.2 Playing games

Another highly publicized result of machine learning was the beating of all humans at Go, an abstract strategy board game known for the richness of all potential strategies (Silver et al., 2018). Games such as Go or chess do not fit exactly the framework of supervised learning. Those games are characterized by a position, or state, which mainly corresponds to the disposition of pieces on the game board, that is changed by moves, or action, made iteratively by two players. The game is over when a terminal state is reached. Each terminal state is associated with a game output, could it be a draw, a win of player one or a win of player two. What needs to be learned is a strategy, or policy, that can be seen as a function that maps a game state to an action or move. The goal is to come up with a winning strategy whatsoever the opponent strategy. These problems can be approached with reinforcement learning (Sutton and Barto, 2018), whose goal is to navigate an environment over time in order to maximize some reward functions. The wording "reinforcement" came from behavioral psychology and describes the fact that the strategy is learned by trials and errors. In the case of games, the computer can simulate many games against itself before finding a winning strategy.

2.3.3 Recommender systems

Far away from the fantasy of creating superhuman creatures, a consequent amount of machine learning development is the fruit of businesses trying to better identify potential customers and prospects of bestseller products. A particularly hot topic is recommender systems, whose goal is to propose to a given customer characterized by some features x, the most adequate content y. Recommender systems are now everywhere to suggest personalized content, could it be for browsing music and movies, or for targeted advertising.

In 2009, Netflix ran a competition to predict how many stars over five a user might rank a movie. Their goal was clear : have a system that finds the most relevant movies to recommend to the user. They offered $1,000,000 to anyone coming up with the best performance on this prediction problem (Bennett and Lanning, 2007; Lohr, 2009). A specificity of this problem was the absence of features to describe movies or users. Among suggestions of many participants, the best solution was provided by

[2] Indeed, some researchers use concepts of human psychology to describe their algorithms such as attention and saliency maps (Cabannes et al., 2020a).

"collaborative filtering". This technique consists in trying to factorize the note given to a movie by a user as a product $u^\top m$, where u characterizes the user, and m characterizes the movie with the smallest (in terms of number of coordinates) vectors u and m. The goal is to find few characteristics for each user that respond to the same number of few characteristics for each movie in order to explain given scores and to predict future scores – think of $u = (\text{how much I like romantic movie}, \text{how much I like action movie})$ and of $a = (\text{how much is it a romantic movie}, \text{how much is it an action movie})$. Formally this technique is formalized by trying to factorize the table $S = (\text{score}(\text{user}, \text{movie}))_{\text{user,movie}}$ as $S = UM$ on observed scores for U and M of smallest rank. While the rank is not a continuous function, making this problem hard to solve, it is possible to relax it into a concave function thanks to the nuclear norm and get nicely behaving solutions.

2.3.4 Intelligent infrastructure

The machine learning of today is the science of data. The competency to manage massive amounts of data opens the way for massive systems of information. Moreover, advances in sensors and measurement tools as well as robots allowing vast automation of tasks create optimistic perspectives in fields such as farming or medicine. The big dreamer might fall for the internet of things governing hydrometry sensors, cameras recognizing aphids, automatic release of pesticides, analysis of the best companion plants and so on. But without going that far, simply putting in place numerical tools to easily perform large scale cross-sectional and longitudinal studies would highly accelerate medicine – recall debates on Hydroxychloroquine as a COVID treatment based on small studies such as Gautret et al. (2020). Arguably, the revolution promised by machine learning is not the revolution of machine learning per se, nor the creation of supra-humanoid, but the future deployment of massive active monitoring systems (Jordan, 2019).

To give a personal touch to this discussion, in 2019, with two friends, we created an “artificial intelligent system” to interact with artists during their creation process (Cabannes et al., 2019). In this project, much of our work was engineering to make sure that the process was innocuous for the artists. The machine learning algorithm running in the middle was only a part of the complete picture.

2.3.5 Making sense of scientific data

While we have described the application of machine learning for non-science purposes, in particular for marketing and agriculture, there is traction for machine learning as science for science. Of course, supervised learning techniques can be applied to scientific problems, such as the protein folding problem (Jumper et al., 2021). But machine learning techniques are also handy to make sense of data coming from massive scientific experiments such as the quest for the Higgs boson at the CERN (Larkoski et al., 2020), or the analysis of gravitational waves (Gebhard et al., 2019). In those last two examples, the goal is to retrieve a signal that has caused some observations given a potential high-level of noise.

2.3.6 Signal recovery and inverse problems

Signal recovery from noisy observations is an active field of research with many appreciable results. The goal of compression is to, given a signal x that is heavy to describe, compress it into a signal $y = f(x)$ much lighter to describe, store, transmit, *etc.* such that given the knowledge of f and a prior on x, one can retrieve the signal x from the observation y. Historically, this has been studied for telecommunications with several results coming from the Bell Labs, such as the Nyquist-Shannon theorem (Shannon, 1949; Nyquist, 1928): assuming x to be a function with bounded frequencies, it can be reconstructed by interpolation (Whittaker, 1915). More recently, Candès et al. (2006) and Donoho (2006) looked at the transmission of a vector $x \in \mathbb{R}^n$ known to be sparse, *i.e.* having only $s << n$ non-zero components, under the form $y = Ax \in \mathbb{R}^m$ for a chosen $A \in \mathbb{R}^{m \times n}$. They showed that by taking the row of A sufficiently “incoherent” it is possible to recover x with $m \approx 2s$. Moreover, when the coefficients of A are random, normally distributed, the “incoherence” property holds with high probability. This allows to drastically reduce compression, as well as acquisition, of sparse signal, and had successful applications in medical imaging (Lustig et al., 2008; Sidky and Pan, 2008). Think of the impact of compressed sensing when it permits reducing X-ray exposure during tomography for the same amount of reconstruction fidelity.

Signal recovery is not the primary focus of machine learning, but as the field is gaining traction, its tools and scope of applications are growing, and similarly to compressed sensing papers that find their place in machine learning journals and conferences, other imaging techniques might as well. To conclude on those

inverse problems beyond machine learning, let us mention the work of Fink (1997), which leverages wave reversibility and information preservation to design time-reversal mirrors in order to locate a wave source and recreate the emitted wave. In the same vein, techniques such as sonar or medical ultrasonography are based on propagating waves into a medium and deduce a topography of the medium from the waves echo. Among others, they have been used to locate oil in the ground, after creating a wave with dynamite. Quite remarkably, recent research has shown those explosions to be a waste of resources as it is possible to leverage ambient noise of the Earth's crust in order to image it (Garnier and Papanicolaou, 2016).

2.4 A tour of models

In this section, we focus on the supervised learning settings, and we discuss classical models to learn functions. Our classification into local averaging, reproducing kernel and deep learning methods is aligned with the class taught by Francis, on learning from first principles (Bach, 2023).

2.4.1 Local averaging methods

If you are a real estate agent pricing houses, and you encounter a new house and need to price it, a natural idea is to look for similar houses that have been sold recently, and infer a price from those similar examples. This is exactly what captures local averaging methods. Suppose that you have collected examples $(X_i, Y_i)_{i\leq n}$ with features X_i characterizing a house indexed by $i \in [n]$ and its price Y_i. Given a new house with characteristics x, you can look at the k-th nearest neighbors $\mathcal{N}_k(x)$ that contains k elements in $[n]$ and such that for $i \in \mathcal{N}_k(x)$ and $j \in [n] \setminus \mathcal{N}_k(x)$, $d(x, X_i) \leq d(x, X_j)$ according to a distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. This defines the nearest neighbors predictor $f_n : \mathcal{X} \to \mathcal{Y}$ as $f_n(x) = \frac{1}{k}\sum_{i\in\mathcal{N}_k(x)} Y_i$. Eventually, you might want to give more importance to houses that are really similar to the one you are trying to price, and less to the k-th nearest neighbor. To do so, you can introduce a system of weights $w_i = w_i^{(n)} : \mathcal{X} \to \mathbb{R}$ that are positive and sum to one, and refine f_n as

$$f_n(x) = \sum_{i=1}^{n} w_i(x) Y_i = \sum_{i=1}^{n} w_i(x) f^*(X_i) + \sum_{i=1}^{n} w_i(x)\varepsilon_i$$

with $\varepsilon_i = Y_i - f^*(X_i)$ linked with the noise of having observed Y_i when we would have preferred to observe $f^*(X_i)$.

Universal consistency. While this pricing method sounds really sensible, the job of statisticians is to make formal statements to prove its soundness. The most classical statement is to prove that when the number of examples goes to infinity, we retrieve the optimal pricing rule. This property is known as *consistency*. Consistency is usually defined formally through convergence in probability (although some prefer to define it with convergence almost surely) of the risk $\mathcal{R}(f_n)$, which is seen a random variable depending on the sampling of the dataset $\mathcal{D}_n = (X_i, Y_i)_{i\leq n} \sim \rho^{\otimes n}$, toward the infimum risk $\mathcal{R}^*$. Methods that are consistent without assumptions on the solution f^* are known as *universally consistent*.

Consistency of local averaging methods has first been derived by Stone (1977). In essence, it consists in assuming that the noises $(\varepsilon_i)_{i\leq n}$ are well-behaved, such that $\sum_{i=1}^n w_i(x)\varepsilon_i$ cancels out as the number of samples n goes to infinity, and assuming the weighting scheme concentrates locally around x such that, for $\rho_{\mathcal{X}}$-almost all x, informally,

$$\lim_{n\to+\infty} f_n(x) = \lim_{n\to+\infty} \sum_{i=1}^{n} w_i(x) f^*(X_i) = \lim_{r\downarrow 0} \frac{1}{\rho_{\mathcal{X}}(B(x,r))} \int_{B(x,r)} f^*(t)\rho_{\mathcal{X}}(\mathrm{d}t) = f^*(x).$$

The last inequality being due to Lebesgue differentiation theorem (the derivative of the integral of a function equals the original function), which is true when f^* is bounded.

Curse of dimensionality. Once universal consistency has been derived, statisticians focus on asymptotic rates of convergence, as well as non-asymptotic generalization bounds. Those results are more informative than universal consistency, similarly to the fact that concentration inequalities are arguably stronger than the central limit theorem which is stronger than the law of large numbers. When providing non-asymptotic

generalization bounds, we should specify a loss function. When we output continuous values, $\mathcal{Y} = \mathbb{R}$, for algebraic reasons, people tend to use the mean-squares loss, defined as $\ell(z, y) = \|z - y\|^2$. In this setting, $f^*(x) = \mathbb{E}_\rho [Y \mid X = x]$, and under the condition that f^* is Lipschitz plus some mild condition on ρ and classical assumptions on the weighting scheme (Györfi et al., 2002), it is possible to get convergence rates of the type

$$\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq cn^{-\frac{2}{d+2}}, \tag{2.5}$$

for c a constant depending on the hardness of the problem, and d the dimension of the input space $\mathcal{X}$. The dependence in d is explained by the fact that to cover the space $[0, 1]^d$ with precision ε in order to have meaningful neighbors to predict the value of a Lipschitz function, one need ε^{-d} points. Meaning that as the dimension increases, one will need exponentially more points in order to guarantee a given risk level, a phenomenon which is referred to as the *curse of dimensionality*. The curse of dimensionality is troublesome for machine learning problems where the input dimension can be really large such as on images with millions of pixels.

Leveraging smoothness and higher-order information. It is possible to break the curse of dimensionality in (2.5) by leveraging additional assumed structure of f^*. In particular, if f^* admits smooth derivatives, one can use Taylor formula, and try to infer derivatives values by finite differences. This will lead to a much more precise estimate f_n of f^*. In particular, if f^* is $s + 1$ times differentiable in every direction, when a zeroth order Taylor expansion makes a local error of order ε, an s-th order expansion will make an error of order ε^s. As such, it is natural to hope for an estimator f_n that achieve the following generalization bound,

$$\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq cn^{-\frac{2s}{d+2s}}, \tag{2.6}$$

This is exactly the rates that local polynomial methods are achieving. We refer to Tsybakov (2009) for additional details on the matter. Note that higher-order derivatives need more points in the neighborhood of x in order for finite difference methods to provide a good estimate of those derivatives. As such, we might expect a transitory regime, before the number of samples n gets sufficiently large, where there are not enough points in the neighborhood of x to benefit from high-order smoothness.

2.4.2 Reproducing kernel methods

Suppose that you are the same real agent, and that you encounter a new house that is really different from the one you have in your database, so that local averaging prediction is meaningless. For example, suppose that the house is huge, but it is not close to any good schools, and while you have already sold big houses, or houses without good nearby schools, you have no records for big houses without good nearby schools. Then a natural idea is to proceed with factor analysis, trying to figure how the price of a house is driven by its size, and by good nearby schools. Simple factor analysis model supposes that characteristics $x \in \mathbb{R}^d$ combine linearly to form the output $y \in \mathbb{R}$. This leads to the search of an estimate f_n as a linear model $f_n(x) = w^\top x$, for $w \in \mathbb{R}^d$ to be determined in order to ensure that $f_n(X_i) \approx Y_i$ on your training dataset $(X_i, Y_i)_{i \leq n}$.

Packing features. Consider the case where $\mathcal{X} = \mathbb{R}$ and when plotting $(X_i, Y_i)_{i \leq n}$ seems to indicate that a quadratic relationship links x to y rather than a linear one, as illustrated on Figure 2.1. In this case, one can enrich the input x with the features $\varphi(x) = (1, x, x^2) \in \mathbb{R}^3$, and perform a linear regression with the data $(\varphi(X_i), Y_i)_{i \leq n}$. When it is hard to visualize the relationship between x and y, one might be tempted to concatenate a lot of features in the vector $\varphi(x)$ in order to make sure to be able to predict y from $\varphi(x)$. However, such a technique is prone to learn spurious correlation between features and outputs of the training data, and not to generalize well to unseen I/O pairs. This phenomenon is called overfitting.

Avoiding overfitting. If we directly find the predictor $f_n(x) = w^\top \varphi(x)$ by minimizing $\sum_i \ell(w^\top \varphi(X_i), Y_i)$, our method will likely overfit the training data. Overfitting is often the fruit of a linear combination of features $\omega^\top \varphi(X) \approx 0$ for $\omega \in \mathbb{R}^d$ (assuming $\varphi(X) \in \mathbb{R}^d$) that is noise spuriously correlated with Y on the training data $(X_i, Y_i)_{i \leq n}$, leading to $w = C \cdot \omega$ for a big C. As a consequence, a way to avoid overfitting is to constraint w to have a small norm, or to look for the minimizer of the regularized objective $\frac{1}{n} \sum_i \ell(w^\top \varphi(X_i), Y_i) + \lambda \Omega(w)$, for $\Omega : \mathbb{R}^d \to \mathbb{R}$ a regularizing function and $\lambda \in \mathbb{R}_+$ a regularizing parameter. To ease this minimization, it is possible to use $\Omega(w) = \|w\|_2^2$. An interesting alternative is to use the ℓ_1-norm $\Omega(w) = \|w\|_1$ as it pushes w

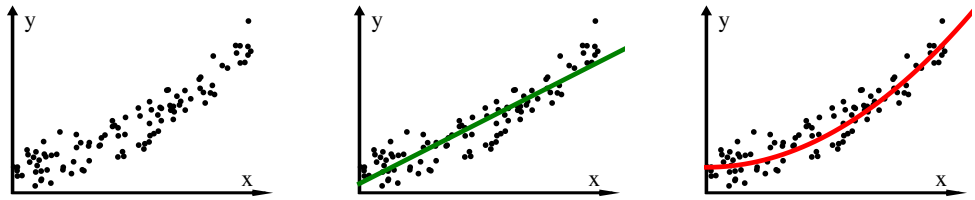

Figure 2.1: Plotting $(X_i, Y_i)_{i\leq n}$ (left) indicates that a quadratic regression (right) is better suited to derive y from x than a linear regression (middle). However a quadratic regression is nothing but a linear regression with features $(1, x, x^2)$.

to be sparse (Tibshirani, 1996), which makes the predictor $f_n(x) = w^\top \varphi(x)$ relatively simple to interpret. Sparsity is a rich concept (Bach et al., 2012), yet it was not a focus of this thesis. In the following, we will consider $\Omega(w) = \|w\|_2^2$.

Kernel methods. When minimizing $\sum_i \ell(w^\top \varphi(X_i), Y_i) + n\lambda \|w\|_2^2$, it is clear that $w^* \in \mathrm{Span}(\varphi(X_i))_{i\leq n}$ as any part of w supported on the orthogonal of the span of the $(\varphi(X_i))_{i\leq n}$ will not change the value of any $w^\top \varphi(X_i)$ but will higher $\|w\|_2$, a fact that is referred as the representer theorem (Scholkopf and Smola, 2001). Looking for $w \in \mathbb{R}^d$ under the form $w = \sum_i \alpha_i \varphi(X_i)$ with $\alpha \in \mathbb{R}^n$ leads to the new problem $\sum_i \ell([K\alpha]_i, Y_i) + \lambda \alpha^\top K \alpha$, with $K = (\langle \varphi(X_i), \varphi(X_j) \rangle)_{i,j\leq n} \in \mathbb{R}^{n\times n}$, $[K\alpha]_i$ the i-th entry of the vector $K\alpha \in \mathbb{R}^n$ and $f_n(x) = \sum_i \alpha_i \langle \varphi(X_i), \varphi(x) \rangle$. As one can notice, everything can be expressed through the scalar product $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+; (x, x') = \langle \varphi(x), \varphi(x') \rangle$, forgetting about the features map $\varphi : \mathcal{X} \to \mathbb{R}^d$. As such, rather than making explicit the features, one can make explicit the kernel k, under the sole condition that for any $m \in \mathbb{N}$ and $(x_i)_{i\leq m} \in \mathcal{X}^m$, $k(x_i, x_j)_{i,j\leq m}$ is symmetric semidefinite positive. In particular, when $\mathcal{X}$ is a collection of structured objects, it can be easier to describe how similar are two x through k rather than deriving features φ.

When $\ell(z, y) = \|z - y\|^2$, the empirical risk minimization admits a closed form solution

$$f_n(x) = K_x^\top (K + n\lambda I)^{-1} \mathbf{Y},$$

with $\mathbf{Y} = (Y_i)_{i\leq n} \in \mathbb{R}^n$ and $K_x = (k(X_i, x))_{i\leq n} \in \mathbb{R}^n$. As n, the number of data, goes to infinity, this converges to $f_\lambda(x) = \bar{K}_x(\bar{K} + \lambda I)^{-1} f^*$ where $\bar{K}$ operates on functions as $\bar{K}(f) = x \to \int k(x, x') f(x') \, \mathrm{d}\rho_{\mathcal{X}}(x')$ and $\bar{K}_x(f) = \bar{K}(f)(x)$. And as λ goes to zero, under mild reachability assumption, f_λ converges to f^* (Caponnetto and De Vito, 2006). An advantage of kernel methods is that it reduces the search of $w \in \mathbb{R}^d$ to the search of $\alpha \in \mathbb{R}^n$ which is valuable when $d >> n$. This is known as the "kernel trick".

Reproducing kernel Hilbert space (RKHS). Kernel methods are backed-up by a rich mathematical concept, linked with well-behaved classes of functions. It can be introduced in a completely different fashion, as it has been historically by Aronszajn (1950). Consider a scalar product of functions mapping $\mathcal{X}$ to $\mathcal{Y} = \mathbb{R}$ and the class of functions $\mathcal{F} = \left\{ f : \mathcal{X} \to \mathcal{Y} \,\middle|\, \|f\| = \sqrt{\langle f, f \rangle} < +\infty \right\}$. Suppose also that the evaluation maps $L_x : \mathcal{F} \to \mathbb{R}; f \to f(x)$ are bounded linear operators. This implies the existence of representer $k_x \in \mathcal{F}$ such that the "reproducing property" $L_x(f) = \langle k_x, f \rangle$ holds for any $x \in \mathcal{X}$ and $f \in \mathcal{F}$. It can be shown that $\mathcal{F}$ is the completion of the linear combination of $(k_x)_{x\in\mathcal{X}}$, $\mathcal{F} = \overline{\mathrm{Span}(k_x)_{x\in\mathcal{X}}}^{\|\cdot\|}$. One inclusion is due to the fact that $k_x \in \mathcal{F}$ and that $\mathcal{F}$ is close by linear combination and completion; the other due to the fact that if $f \in \mathcal{F} \cap (k_x)_{x\in\mathcal{X}}^\perp$ than $f(x) = \langle k_x, f \rangle = 0$. Hence, there is a unique association between the kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}; (x, x') \to \langle k_x, k_{x'} \rangle$ and the so-called reproducing kernel Hilbert space $\mathcal{F}$ (Mercer, 1909).

Smoothness adaptability. Under the knowledge that the target function f^* belongs to a RKHS $\mathcal{F}$, using the associated kernel k generates good learning rates, akin to linear regression. Moreover, when using the mean-squares error, one can guarantee universal consistency of the method described above, under the assumption that $\mathcal{F}$ is dense in $L^2(\rho_{\mathcal{X}})$, a property that is verified for many usual kernels (Micchelli et al., 2006). But what about the rates in this case, do they deteriorate badly when $f^* \notin \mathcal{F}$? How does kernel methods adapt to the regularity of the function f^*? Indeed the operator $\bar{K}$ introduced above provides the

answer. In fact, it is possible to show that $L^2(\rho_{\mathcal{X}}) = \operatorname{im} \bar{K}^0$ and $\mathcal{F} = \operatorname{im} \bar{K}^{1/2}$ (we provide many details on this operator in sections 5.C and 9.C), and that by only adapting the regularization parameter λ (which is usually done through cross-validation) the excess of risk for the least-squares error of the estimator introduced above is $\mathcal{O}(n^{-2q/(2q+1)})$ for the biggest $q \in (0, 1)$ such that $f^* \in \operatorname{im} \bar{K}^q$ (Caponnetto and De Vito, 2006).

Transitory regimes. While research papers tend to showcase rates that are best in terms of exponents raising the number of samples, those rates are often not observed in practice before accessing a high number of samples. As such, practitioners might prefer to write

$$\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \inf_{0 \leq t \leq q} c_t n^{-\frac{2t}{2t+1}}, \tag{2.7}$$

for the biggest $q \in (0, 1)$ such that $f^* \in \operatorname{im} \bar{K}^q$, and some constants $(c_t)_{0 \leq t \leq q}$. The transitory regime corresponds to the case where the best bound is not achieved for $t = q$, and which is when, if the kernel k is leveraging smoothness, points are too far apart to benefit from higher-order information.

2.4.3 Neural networks and deep learning

Most of the state-of-the-art algorithms derived through machine learning are currently based on deep learning. Deep learning consists in long artificial neural networks. Artificial neural networks are a cascade of linear operators and non-linearities that are optimized through gradient descent on empirical risks. This model appeared in the middle of the 20$^{\text{th}}$ century (McCulloch and Pitts, 1943; Rosenblatt, 1958), and were advocated by a few researchers in the 90s (LeCun and Bengio, 1995). It rose to dominance when implemented on graphical processing units, achieving astonishing results on image recognition (Krizhevsky et al., 2012).

Cascade of linear operation. A neural network $f_n : \mathbb{R}^d \to \mathbb{R}^m$ with three hidden layers is parametrized with four matrices $W_1 \in \mathbb{R}^{h_1 \times d}$, $W_2 \in \mathbb{R}^{h_2 \times h_1}$, $W_3 \in \mathbb{R}^{h_3 \times h_2}$, $W_4 \in \mathbb{R}^{m \times h_3}$ and reads

$$f_n(x) = W_4 \sigma(W_3 \sigma(W_2 \sigma(W_1 x))),$$

with $h_1, h_2, h_3 \in \mathbb{N}$ the number of neurons in the first, second and third hidden layers, $\sigma : \mathbb{R} \to \mathbb{R}$ a non-linearity, *e.g.* $\sigma(x) = \max(x, 0)$, and the convention $\sigma((x_1, \cdots, x_n)^\top) = (\sigma(x_1), \cdots, \sigma(x_n))^\top$. We illustrate it in Figure 2.2. Given a loss ℓ, some data $\mathcal{D}_n = (X_i, Y_i)_{i \leq n}$, a regularizer term $\lambda\Omega(W_1, W_2, W_3, W_4)$ and the consequent regularized empirical risk $\mathcal{R}_{\mathcal{D}_n}$, the parameters W_1, W_2, W_3, W_4 can be optimized in order to minimize $\mathcal{R}_{\mathcal{D}_n}(f_n)$. This minimization is usually done with gradient descent, or stochastic gradient descent (Robbins and Monro, 1951), which is better suited for large scale learning problems (Bottou and Bousquet, 2007). Gradients are obtained thanks to the chain rule, and while the computation of $f(x)$ can be seen as forward propagation from x to $W_1 x$ to $W_2 \sigma(W_1 x)$ and so on, the chain rule naturally leads to *backward propagation* or *backpropagation*, which is the name used by researchers to refer to the derivation of gradients with neural networks. Astonishingly, while such a procedure is supposed to stall into undesirable local minima, it has not constrained the recent successes of deep learning.

Hierarchy and convolution. In the last paragraph, we have described the most basic architecture of a neural network, only made of fully connected layers. There are many architectures that refine this basic neural network, the most well known are recurrent neural networks (Williams et al., 1986) that allow dealing with time series of different length and convolutional neural networks that were proven successful to understand images (LeCun et al., 2015). We refer the interested reader to this last paper to get a more precise idea on what convolutional neural networks are. In substance, a one dimensional convolutional filter replaces the linear mapping $\mathbb{R}^{h_i} \to \mathbb{R}^{h_o} x \to Wx$ with $W \in \mathbb{R}^{h_0 \times h_i}$ by the linear mapping $\mathbb{R}^{h_i} \to \mathbb{R}^{h_i - p}; x \to w * x$ with $w \in \mathbb{R}^p$ and $(w * x)_i = \sum_{j=1}^p w_i x_{i+j}$. This is useful when x has a translation-invariant structure (Fukushima, 1980). Rather than learning to recognize the same pattern at different places by tuning similarly different columns of W, it allows for parameter sharing. As one filter allows for the recognition of one pattern, a convolutional layer is made of several filters, whose responses are concatenated along a new dimension in order to form the layer output. After passing through a filter bank, the response is usually pruned by keeping only local maxima, an operation known as max-pooling. This subsampling operation builds a hierarchical structure, allowing for further filters to learn broader patterns (Behnke, 2003). Translation invariance and hierarchy are two important concepts (Simon, 1962; Mallat, 1989) that have been used to design others "hand-made" methods that were used in the past to analyze images (Lowe, 1999; Dalal and Triggs, 2005).

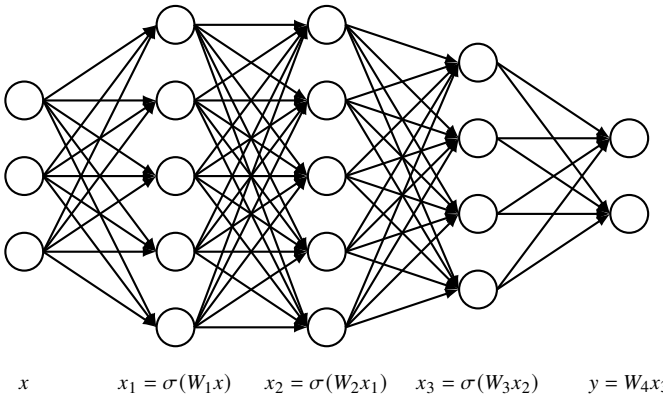

Figure 2.2: Representation of a neural network with three hidden layers and parameters $(d, h_1, h_2, h_3, m) = (3, 5, 5, 4, 2)$, as a weighted directed acyclic graph. The input $x = (x_{(i)}) \in \mathbb{R}^3$ is passed through the first set of edges, leading to the first layer with value $x_1 = (x_{1,(i)}) \in \mathbb{R}^5$ specified by $x_{1,(i)} = \sigma(\sum_{1 \leq j \leq 3} W_{1,(i,j)} x_{(j)})$ based on the set of weights $W_1 = (W_{1,(i,j)}) \in \mathbb{R}^{5 \times 3}$ and the non-linearity $\sigma : \mathbb{R} \to \mathbb{R}$, and so on until reaching the output $y \in \mathbb{R}^2$. The wording "deep learning" refers to learning with neural networks that have a really large number of hidden layers, which can lead to hundreds of billions of parameters (see *e.g.* Brown et al., 2020).

Plethora of architectures. While convolutional neural networks are the model of choice to work with images, they do not exhaust what deep learning is about. Recently, transformer models (Vaswani et al., 2017) have become the model of choice to deal with time series in natural language processing, with amazing realizations in language understanding (Devlin et al., 2019) and text generation (Brown et al., 2020), and generative adversarial networks (Goodfellow et al., 2014) introduced earlier have also led to impressive applications such as "deepfakes" (Karras et al., 2019), which are synthetic images that look real and might be used for malicious purposes. The rapid spread of deep learning is partly due to its ease to implement through differential programming libraries that implement automatic differentiation and backpropagation when forward propagation is specified (Abadi et al., 2016; Paszke et al., 2019). This allows practitioners to easily develop and test new ideas and architectures, even though training neural networks is known to be quite unstable and greedy in terms of time and energy (see *e.g.* García-Martín et al., 2019).

Statistical properties. While neural networks work better than local averaging and kernel methods, statisticians have not arrived at a consensus in order to describe theoretically why neural networks work so well. In particular, it is currently hard to state generalization bounds akin (2.5), (2.6) and (2.7). Because of their wide-spread use, many are trying to derive a better understanding of those models. Some are trying to come up with similar models based on concepts that are better understood such as wavelets (Bruna and Mallat, 2013), kernels (Mairal et al., 2014) or sparse coding (Papyan et al., 2017). Others are looking at the properties of neural networks once the training has taken place (Mahendran and Vedaldi, 2015; Selvaraju et al., 2017), notably discovering the possibility to fool them with imperceptible noise (Szegedy et al., 2014; Ilyas et al., 2019). Finally, statisticians are trying to explain some of their characteristics with different tools, could it be some measure of complexity (Bartlett et al., 2017; Zhang et al., 2021), group theory and harmonic analysis (Mallat, 2016), statistical physics (Spigler et al., 2019) or asymptotic limit (Jacot et al., 2018; Chizat et al., 2019; Mei and Montanari, 2022).

2.5 The practice

In this section, we discuss the daily job of data scientists before confronting this reality with the prior theory in order to discuss the limitations of the supervised learning framework.

2.5.1 The job of data scientists

There are arguably four steps in a project of a data scientist when it comes to machine learning. While we present those tasks in a downstream fashion, in practice, it is important to get a big picture and a rough sketch

of each step first in order to ease the transitions between steps and avoid going back too often to prior steps. All along the process, intuition can be found by reflecting on how the work would be accomplished manually, that is without learning perspectives.

1. Define the problem. First of all, data scientists should be clear about the task to solve, look at the current solution, and see how it can be improved with machine learning. In order to prove the usefulness of the new method, they should define some clear measures of success, and benchmark simple baselines. At this point comes the definition of the output space $\mathcal{Y}$ and of the loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that should be defined in coherence with the precedent measures of success. Once the target space $\mathcal{Y}$ is well-defined, the next step is to find a set of relevant features that defines the input space $\mathcal{X}$ and allows prediction of targets. Of course, the features that can be accessed, hence the definition of $\mathcal{X}$, depend on the data one can get.

2. Get clean data. Once $\mathcal{X}$ and $\mathcal{Y}$ are well-defined, it is time to find a set of I/O examples $(X_i, Y_i)_{i \leq n}$. This is often the most time-consuming task. How many samples are needed regarding the statistical hardness of the task at hand? Where to scrap data? How to easily associate outputs to inputs? Once enough samples have been collected, it is usual to set aside a part of those to form a test dataset, that would only be looked at before deployment to validate the working status of the built system. The remaining data form the train set. At this point, data is often very cluttered, implying a consequent amount of cleaning before feeding a machine learning model. In particular, when collected from different sources, there might be some coordinates missing for various input samples. For example when working with countries, it will be hard to get an indication of the number of people below poverty line in North Korea. A decision should be made with those missing data, could it be filling with the feature average, or dropping completely the corresponding sample. Similarly, the practitioners might have to deal with outliers or data sources that got stalled on wrong values.

3. Train a model. Once clean data have been collected, machine learning models are ready to be trained. In order to help models, some features engineering might be useful, such as centering and normalizing the variance of the data, or applying more advanced transformations to get samples that are well-behaved, *e.g.* following a unit Gaussian distribution. It is usual to combine different models together, when those models capture different relationships between inputs and outputs, in order to boost the overall performance, a procedure known as boosting (Schapire, 1990). Hyperparameters are generally tuned with cross-validation. Cross-validation consists in splitting the samples into a given number of folds, using all the folds but one to train the models and getting a measure of success on the last fold before averaging this measure over the different folds that can be retained for testing. This associates a score to a method with a given set of hyperparameters. Best hyperparameters are chosen as the ones leading to the highest score. Practitioners should be careful not to try too many models or hyperparameters as the more they try, the more likely will be learned spurious correlations that only explain the training samples I/O relationship.

4. Deploy the model. Once a model has been learned to predict output from input, one should excavate the test dataset, and see if the method performs better than baselines. If so, it is time to benefit from this new method and put it into production. At this point, it is important to monitor the quality of new input data, to make sure that it is not corrupted, and that its distribution is similar to the distribution of training data. There might also be an emphasis on engineering problems, such as managing databases and computing architectures.

For the curious reader, there are many good references to prepare future data scientists for their job (see *e.g.* Géron, 2017).

2.5.2 Framework limitations

Loss design. In the formalization of supervised learning that we gave earlier, we assumed that the loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is clearly defined and that the risk $\mathcal{R}$ to minimize is nothing but the average over samples of the error made by a predictor measured with this loss. In some critical applications, such as medical one, one might prefer to build $\mathcal{R}_{\mathcal{D}_n}$ as the worst case of $(\ell(f(X_i), Y_i))$ or build $\mathcal{R}$ as some quantile of the pushforward distribution $\ell(f(\cdot), \cdot)_{\#}\rho$ rather than building them as their respective means. Moreover, in practice, it might be hard to define clearly the loss ℓ. For example, it is the case when it comes to style transfer, that is taking two images and outputting an image with the content of the first one and the style of

the second, or to super-resolution, that is enhancing the resolution of an image (Johnson et al., 2016). The same is true with our cooking example, how to access the quality of a recipe derived by a computer? Should it be the most healthy, the most tasty or somewhat healthy and tasty? Can we derive such losses based on chemical considerations? Or should we only measure success based on people reviews, meaning high costs to evaluate the risk $\mathcal{R}$?

Sometimes, losses can be partially or totally designed by leveraging the structure of the problem. For example, on problems where there is a desired invariance defined by a set G of functions $g : \mathcal{X} \to \mathcal{X}$ such that we should have $f \circ g = f$, if this invariance is not built into the model, it can be built into the loss by replacing $\ell(f(X), Y)$ with $\sum_{g \in G} \ell(f(g(X)), Y)$, which is notably the idea of "data augmentation". Another example is when we have a parametric model of the relation between X and Y, $Y = g(X, \theta, \varepsilon)$ for g a known function, θ an unknown parameter to optimize and ε some random noise. The parameter θ can be fitted by maximizing the likelihood of having observed $\mathcal{D}_n = (X_i, Y_i)_{i \leq n}$ under the model. Assuming the samples are independent, this probability factorizes as the product of the probability of observing each (X_i, Y_i). As a consequence, maximizing the likelihood, or equivalently the log-likelihood, is similar to minimizing the risk $\mathcal{R}$ defined through the loss $\ell(X, Y, \theta) = -\log \mathbb{P}(Y \mid X; \theta)$ with the last term designing the probability of observing Y given the observation of X under the model parametrized by θ. When the parametric model is an exponential family, this procedure is to be linked with generalized linear models (Nelder and Wedderburn, 1972).

Independent identically distributed variables? Suppose you want to predict annual growth from indicators, such as the gross domestic product, regarding a country. You might collect a dataset $(X_{c,t}, Y_{c,t})_{c \in C, t \in T}$ with many countries, indexed by a set C, and many years, indexed by a set T, features X and corresponding growth Y, but this dataset will violate the independent identically distributed assumption of statistical learning. For example, an oil crisis happening in year t will have a repercussion on each $(Y_{c,t})_{c \in C}$, and if one try to add oil prices in the features X to enforce the independence of the conditional variables $\left(Y_{c,t} \,\middle|\, X_{c,t}\right)_{c \in C}$, than the $(X_{c,t})_{c \in C}$ will not be independent as this factor will be constant over countries. Similarly, for a given country, indicators are arguably forming a non-stationary process that depends on politics in place, hence $(X_{c,t})_{t \in T}$ is not a family of independent random variables. However, many data scientists might be tempted to still use the many resources provided by supervised learning, without trying to leverage the underlying processes.

More importantly, suppose that one trains a model on western countries, benefiting from higher quality data, before deploying the model to help some decision-making in developing countries. There might be some important cofactors such as quantitative easing policies, or trust of populations in money, that change completely the picture when it comes to predicting growth from country indicators. As such, the testing distribution is not the same as the training distribution, a phenomenon known as *covariate shift* (Heckman, 1979; Shimodaira, 2000). Recently, this problem has come back to denounce the risk of automatic systems that reproduce human-biases (Caliskan et al., 2017), which has found an echo in the civil society (Benjamin, 2019).

Interestingly, while data scientists often try to artificially squeeze their problems into the framework of supervised learning, many researchers are trying to leverage specific settings, such as the fact that data might be coming from different datasets (Arjovsky et al., 2019).

Practice and theory regarding generalization. In theory, generalization error might be given by derivations allowing to get results such as (2.5), (2.6) and (2.7). In practice, those bounds might be hard to compute because of unknown constants, yet they provide precious insights on the number of samples to consider in order to achieve statistical significance. For example, bounds reading d/n suggests that the number of samples n should be bigger than the number of dimensions d of the input space $\mathcal{X}$, while bounds in $n^{-1/d}$ suggests that the number of samples should be exponentially bigger than the number of dimensions. Anyhow, data scientists mostly compute generalization scores by computing the empirical risk of a predictor f_n on a fresh test set of data. According to this practice, researchers might study the benefits of the test set validation procedure. For example, they may leverage the distribution of the empirical test error in order to transform a predictor $f_n : \mathcal{X} \to \mathcal{Y}$ into a confidence interval predictor $f_n : \mathcal{X} \to 2^{\mathcal{Y}}$ with a given probability to be valid (Vovk and Shafer, 2008).

Chapter 3

Learning Rates with Discrete Outputs

Understanding the reliability of machine learning is crucial to deploy learned models in the wild. It was also an important point of this thesis that focuses on algorithms for weakly supervised learning with appreciable theoretical guarantees. This chapter elaborates on such guarantees and discusses some of our work (Cabannes et al., 2021c; Cabannes and Vigogna, 2022), which is reproduced entirely in part II. For simplicity, it is set in the supervised learning formalization made in the precedent chapter. In particular, we consider $\mathcal{X} \subset \mathbb{R}^d$ an input space, $\mathcal{Y}$ a discrete output space, ρ a joint distribution on $\mathcal{X} \times \mathcal{Y}$, and ℓ a loss function. Our goal is to minimize the risk $\mathcal{R} : \mathcal{Y}^{\mathcal{X}} \to \mathbb{R}$ defined as

$$\mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\rho} \left[\ell(f(X), Y) \right] . \tag{3.1}$$

3.1 Statistical learning theory

In this section, we discuss results offered by statistical learning theory to get insights on what can be learned from data. In particular, we get more precise about classical derivations of generalization bounds.

3.1.1 Insights from information theory

Artificial intelligence designs a putative system created by humans that would display signs of intelligence, whatsoever this means. Machine learning consists in implementing such a system with a machine that learns, *i.e.* finds in a sort of autonomous fashion, how to behave in order to produce a desired output given some input parameters. Statistical learning consists in presenting the machine with a set of training examples that show desired outputs for different inputs in order for the machine to infer a rule to produce outputs from inputs. Statistical learning theory borrows many tools from information theory.

Information theory is concerned with transmitting information. Think that we want to describe a function $f : \mathcal{X} \to \mathcal{Y}$ to a friend, with only a finite number of information. If we know that f belongs to a finite set of functions $\mathcal{F}$, based on dichotomy, we only need $\log_2(|\mathcal{F}|)$ bits to encode this function among all functions in $\mathcal{F}$. Now, if $\mathcal{F}$ is continuous, we cannot encode all functions on a finite number of bits. Yet, if we only want to unravel a signal $f \in \mathcal{F}$ up to a precision ε, given a certain error metric[1] L (typically $L(\hat{f}, f) = \mathbb{E}\left[\ell(\hat{f}(X), Y) - \ell(f(X), Y)\right]$), we could do so with an encoding on less than $\lceil \log_2(\mathcal{N}(\varepsilon)) \rceil$, with $\mathcal{N}(\varepsilon)$ the minimum number of balls of size ε (with the metric L) to cover $\mathcal{F}$. To quantify the complexity of the set $\mathcal{F}$, it is natural to look at the behavior of the number of bits we need to encode any function in $\mathcal{F}$ up to an error ε as ε get to zero. As a consequence, it is useful to define a notion of size of $\mathcal{F}$ (with respect to the error metric L) as the superior limit

$$C_{\mathcal{F}} = \limsup_{\varepsilon \to 0} - \log_2(\mathcal{N}(\varepsilon)) / \log_2(\varepsilon).$$

Note the analogy with statistical mechanics, where signals are microscopic arrangements of unit elements, and the set $\mathcal{F}$ is the corresponding macroscopic system. Assuming that all arrangements have the same probability to appear, the Boltzmann entropy is exactly, up to the Boltzmann constant, the logarithm in base

[1] Here, the word "metric" is used in a prosaic fashion since L might not be a distance.

two of the cardinality of $\mathcal{F}$. This explains the wording Kolmogorov entropy for the logarithm in base two of the covering number $\mathcal{N}$.

To make the preceding paragraph more concrete, suppose that we want to transmit a function $f : [0,1]^d \to \mathbb{R}$ to our friend. Typically, we would give to our friend m bits of discrete information, plus some information on the function class (*e.g.* f is a polynomial of order d, or, thinking in terms of Fourier transform, it has a low energy on high harmonics). For example, assume that f belongs to the Sobolev space $H^m(\mathrm{d}x)$, *i.e.* it is m times differentiable, with derivatives up to order m square integrable against the Lebesgue measure. Suppose that our friend knows that $\|f\|_{H^m} \leq 1$, and suppose that we want to minimize the L^2 error. Then, we can give n binary information to localize this function in a covering of $\mathcal{F} = \{f \,|\, \|f\|_{H^m} \leq 1\}$ with 2^n L^2-balls.[2] At best, our friend will be able to localize the signal in an L^2-ball of radius $\varepsilon(2^n)$, thus making at most an error $\varepsilon(2^n)$ on the reconstructed signal, with $\varepsilon(n) = \inf[\varepsilon \,|\, \mathcal{N}(\varepsilon) \leq n] \approx n^{-C_{\mathcal{F}}}$, and $C_{\mathcal{F}}$ the size of $\mathcal{F}$ with respect to the L^2-metric. We refer the interested reader to the seminal paper of Kolmogorov and Tikhomirov (1959) for precise quantification of space sizes.

3.1.2 Vapnik-Chervonenkis theory

Machine learning is concerned with inferring information. In this setting, we want to learn a task $f : \mathcal{X} \to \mathcal{Y}$, and we do so by collecting examples of the solved instance of this task $(x_i, y_i = f(x_i))$. In the current machine learning paradigm, we do not choose the bits of information to transmit, but we are given some points $(x_i, y_i)_{i \leq n}$ that are assumed independently sampled according to the distribution that matters to us. Let us review the picture offered by the classical statistical learning theory for the sake of completeness. Recall that we want to study the excess of risk. We have, with $f^*_{\mathcal{F}}$ and f_n the respective minimizers of the population and empirical risks over $\mathcal{F}$,

$$\mathcal{R}(f_n) - \mathcal{R}(f^*_{\mathcal{F}}) \leq \mathcal{R}(f_n) - \mathcal{R}_{\mathcal{D}_n}(f_n) + \mathcal{R}_{\mathcal{D}_n}(f^*_{\mathcal{F}}) - \mathcal{R}(f^*_{\mathcal{F}}).$$

For any function $f \in \mathcal{F}$, we can use concentration inequality (such as Bernstein inequality) to control the difference between the empirical risk $\mathcal{R}_{\mathcal{D}_n}(f)$, which inherits its randomness from $\mathcal{D}_n$, and its average $\mathcal{R}(f)$. This works well for the second part of the equation, due to $f^*_{\mathcal{F}}$. Sadly, as f_n depends on $\mathcal{D}_n$, we can not apply concentration inequality directly to control the deviation between its empirical and population risk. The classical way to proceed is to find a uniform concentration bound over $\mathcal{F}$, which we can do by controlling the following random quantity

$$\sup_{f \in \mathcal{F}} \mathcal{R}(f) - \mathcal{R}_{\mathcal{D}_n}(f).$$

When $\mathcal{F}$ is finite, we can control this supremum with a union bound, joining together all individual concentration inequality. Similarly, when $\mathcal{F}$ is continuous, we could do similar things with a well-specified ε-cover of $\mathcal{F}$, which can be refined based on chaining techniques, to avoid redundancy of events when performing union bound for supremum. In the statistical learning literature, it is classical to rather use a symmetrization trick to relate $\mathcal{R}(f) - \mathcal{R}_{\mathcal{D}_n}(f)$ to $\mathcal{R}_{\mathcal{D}'_n}(f) - \mathcal{R}_{\mathcal{D}_n}(f)$ for $\mathcal{D}'_n$ another dataset, and study the Rademacher complexity defined as $\mathbb{E} \sup_{f \in \mathcal{F}} n^{-1} \sum_{i=1}^n \varepsilon_i f(X_i)$ for $\mathbb{E}$ the expectation taken over the ε_i defined as variables taking value one or minus one with probability one half.

In the case of binary classification with the 0-1 loss, *i.e.* $\mathcal{Y} = \{-1, 1\}$, $\ell(y, y') = \mathbf{1}_{y \neq y'}$, Vapnik and Chervonenkis proposed slightly different derivations leading to the following bound.

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*_{\mathcal{F}}) \leq c \sqrt{\frac{V_{\mathcal{F}}}{n}}, \tag{3.2}$$

for c a universal constant, and $V_{\mathcal{F}}$ the so-called VC dimension of $\mathcal{F}$, that relates to the average number of points that the class $\mathcal{F}$ is able to shatter (Vapnik, 1995).

A natural question with generalization bounds is how well they quantify the behavior of the excess of risk $\mathcal{R}(f_n) - \mathcal{R}^*$. Indeed, the estimation of the excess of risk given by (3.2) is known to be optimal, in the sense that, for any class of functions $\mathcal{F}$, it is possible to find a specific distribution ρ such that this bound is a lower bound up to a multiplicative constant. This behavior is referred to as *minimax optimality*, meaning that it is not possible to find a better bound, given the class $\mathcal{F}$ of estimators considered.

[2] Note that such a covering of $\mathcal{F}$ can be taken as it is compact with respect to the L^2-topology - even though it is not compact with the H^m-topology, H^m being infinite dimensional.

3.2 Surrogate methods

In this section, we discuss the difficulty of learning discrete-valued functions, and the possibility to tackle the learning problem through surrogate problems that consist of learning continuous-valued functions.

3.2.1 Practical limitations of VC theory

While VC theory was a keystone in machine learning, it does not exhaust practitioner issues, as this theory does suffer from some limitations.

Approximation/estimation trade-off. First of all, VC theory controls the estimation error in (2.4), that is the error in terms of population risk $\mathcal{R}$ between f_n the minimizer of the empirical risk in a hypothesis class $\mathcal{F}$ and $f^*_{\mathcal{F}}$ the minimizer of the population risk in the class $\mathcal{F}$. In particular, equation (3.2) does not tell anything about the approximation error, that is the gap between $\mathcal{R}(f^*_{\mathcal{F}})$ and $\mathcal{R}^*$. To control the approximation error, we need to make assumptions on how good our model $\mathcal{F}$ is to minimize the risk $\mathcal{R}$. There is a trade-off between choosing a big model $\mathcal{F}$, so that the optimal risk $\mathcal{R}^*$ can be achieved by $\mathcal{R}(f^*_{\mathcal{F}})$ inside the hypothesis class, and choosing a small model, so that its capacity, captured by $C_{\mathcal{F}}$ or $\mathcal{V}_{\mathcal{F}}$, is small.

Optimization issues. Finally, if we were to apply VC theory to learn from a continuous input space $\mathcal{X} \subset \mathbb{R}^d$, to a discrete output space $\mathcal{Y}$, we could consider a model $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$. Those functions could be characterized through their decision regions and decisions boundaries, defined for $y, z \in \mathcal{Y}$ as

$$f^{-1}(y) = \{x \in \mathcal{X} \mid f(x) = y\} \subset \mathcal{X}, \qquad \Delta_{y,z}(f) = \overline{f^{-1}(y)} \cap \overline{f^{-1}(z)} \subset \mathcal{X},$$

where the bar stands for the closure of the set. As such, a model $\mathcal{F}$ could be defined as a collection of putative decision regions. The region of disagreement between two functions f and g for a prediction $y \in \mathcal{Y}$ can be expressed through the symmetric difference $f^{-1}(y) \triangle g^{-1}(y)$. When using the 0-1 loss, the excess of risk between f_n and $f^*_{\mathcal{F}}$ is controlled by the measure of the union of disagreement regions, that is $\rho_{\mathcal{X}}(\cup_{y \in \mathcal{Y}} (f_n)^{-1}(y) \triangle (f^*_{\mathcal{F}})^{-1}(y))$. It is then possible to retake the VC theory, using the covering number of $\mathcal{F}$ with respect to this pseudo-distance (Mammen and Tsybakov, 1999). Sadly, even for a simple model $\mathcal{F}$, such as binary classification with half-plane decision regions, minimizing the empirical risk is NP-hard (Arora et al., 1997). This echoes optimization issues in machine learning. While VC theory asks to consider the empirical risk minimizer f_n, it might be really hard to find it in practice, and one might have to settle with an estimate $\hat{f}$ of f_n. This adds a third term to the risk decomposition that is an optimization error term

$$\mathcal{R}(\hat{f}) - \mathcal{R}^* = \underbrace{\mathcal{R}(\hat{f}) - \mathcal{R}(f_n)}_{\text{optimization error}} + \underbrace{\mathcal{R}(f_n) - \mathcal{R}(f^*_{\text{lin}})}_{\text{estimation error}} + \underbrace{\mathcal{R}(f^*_{\text{lin}}) - \mathcal{R}^*}_{\text{approximation error}} .$$

3.2.2 Related continuous surrogate problems

When learning with discrete outputs, direct risk minimization leads to many combinatorial difficulties that translate into computational intractability. This is related to the difficulties of studying and optimizing discrete-valued functions. One technique to overcome this issue consists in learning a discrete-valued function, by learning a surrogate continuous-valued function and thresholding its output in order to make it discrete. Consider binary classification, that is $\mathcal{Y} = \{-1, 1\}$ and $\ell(y, z) = \mathbf{1}_{y \neq z}$. In this setting, the optimal predictor $f^* : \mathcal{X} \to \mathcal{Y}$ is defined as

$$f^*(x) = \text{sign}(g^*(x)), \qquad \text{where} \qquad g^*(x) = \mathbb{E}_{(X,Y)\sim\rho} [Y \mid X = x] .$$

This suggests learning the discrete-valued function $f^* : \mathcal{X} \to \mathcal{Y}$ by learning the continuous-valued function $g^* : \mathcal{X} \to \mathbb{R}$. To an estimate g of g^*, we associate the estimate $f = \text{sign}(g)$ of f^*. One method to learn the conditional expectation g^* is through its characterization with the least-squares error

$$g^* \in \operatorname*{arg\,min}_{g:\mathcal{X}\to\mathbb{R}} \mathcal{R}_S(g) := \mathbb{E}_{(X,Y)\sim\rho} \left[|g(X) - Y|^2 \right] .$$

Given data (X_i, Y_i), one can then consider a model of real-valued functions $\mathcal{G} \subset \mathbb{R}^{\mathcal{X}}$ and the empirical risk minimization

$$\hat{g} \in \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} |g(X_i) - Y_i|^2 .$$

This estimate is cast as an estimate of f^* through the decoding $\hat{f} = \operatorname{sign} \hat{g}$. It is natural to wonder how a good estimate of g^* translate into a good estimate of f^*. This question is answered by calibration inequalities that relate the excess of risk on the surrogate problem with the excess on risk of the original problem (Bartlett et al., 2006). For example, for the least-squares surrogate considered here, we have

$$\mathcal{R}(\operatorname{sign} g) - \mathcal{R}(f^*) \leq \sqrt{\mathcal{R}_S(g) - \mathcal{R}_S(g^*)}.$$

While we only present the least-squares surrogate for binary classification, other surrogates such as the hinge loss, leading to support vector machine, the logistic loss, leading to softmax regression, can be considered. Similarly, surrogate methods can be extended beyond binary classification, for example, in multiclass problem with the $0-1$ loss, that is $\mathcal{Y} = \{1, \cdots, m\}$ and $\ell(y, z) = \mathbf{1}_{y \neq z}$, we can consider $g^* : \mathcal{X} \to \mathbb{R}^{\mathcal{Y}}$, defined as

$$g^*(x) = (\mathbb{P}(Y = y \mid X = x))_{y \in \mathcal{Y}} = \mathbb{E}\left[(\mathbf{1}_{Y=y})_{y \in \mathcal{Y}} \,\middle|\, X = x \right],$$

and the decoding

$$f^*(x) = \arg\max_{y \in \mathcal{Y}} g^*_y(x).$$

In such a setting, g^* is often referred to as a *score* function.

Remark 3 (The pros of predicting scores)**.** *From a formal perspective, if we are only interested in the optimal mapping $f : \mathcal{X} \to \mathcal{Y}$, learning g can be seen as a waste of resources. In essence, this waste of resources is similar to the one when learning the full probability function $(p(y))_{y \in \mathcal{Y}}$ for some $p \in \Delta_{\mathcal{Y}}$ while we only care about the argmax $y^* \in \arg\max_{y \in \mathcal{Y}} p(y)$. In practice, however, it might be of interest to get an estimate of $g^*(x) = \mathbb{P}(Y = y \mid X = x)$, as it might provide important information about how specified is $f^*(x)$ and how we could confidently discard other potential labels y for the input x.*

We refer the interested reader to Nowak-Vila (2021) for deeper reflections about learning with surrogate methods.

3.3 Fast rates derivation

In this section, we sketch roughly the underlying machinery beyond Cabannes et al. (2021c). While writing this section, we gave a fresh look at it to avoid redundancy with results already published and reproduced in Chapter 5. Confronting this machinery with SVM has led us to Cabannes and Vigogna (2022) which is reproduced in Chapter 6.

Regularized risk minimization. For simplicity, consider a surrogate problem consisting of learning a real-valued function; with a linear parametric model of functions $\mathcal{G} = \{g_\theta : x \to \langle \theta, \varphi(x) \rangle \mid \theta \in \Theta\} \subset \mathbb{R}^{\mathcal{X}}$, parametrized by some Hilbert space Θ, and some features $\varphi : \mathcal{X} \to \Theta$. Consider the regularized empirical risk minimization

$$\theta_n \in \arg\min_{\theta} \mathcal{R}_{S,\mathcal{D}_n}(g_\theta) + \lambda_n \|\theta\|^2 .$$

Here, λ_n is a regularization parameter that goes to zeros as the number of samples n goes to infinity and the risk of overfitting vanishes, and $\mathcal{R}_{S,\mathcal{D}_n}$ is the empirical surrogate risk computed from the dataset $\mathcal{D}_n$. Eventually, we could also consider $\hat{\theta}$ an approximation of θ_n related to some optimization techniques. Finally, it is useful to introduce the bias estimate $\theta_\lambda \in \arg\min_\theta \mathcal{R}_S(g_\theta) + \lambda \|\theta\|^2$.

Classical versus new convergence study. Classically, to prove that $\hat{f}$, the decoding of $g_{\hat{\theta}}$, will minimize the original risk $\mathcal{R}$ as n goes to infinity and to give rates of convergence, one can use capacity/estimation assumptions to state that $\hat{\theta}$ concentrate around θ_λ, as well as source/approximation assumptions to state that $\mathcal{R}_S(g_{\theta_\lambda})$ will convergence to $\mathcal{R}_S(g^*)$ as λ goes to zero. Finally, using calibration inequalities, one can relate concentration in θ to concentration in $\mathcal{R}_S$, and then to concentration in $\mathcal{R}$. The problem with this

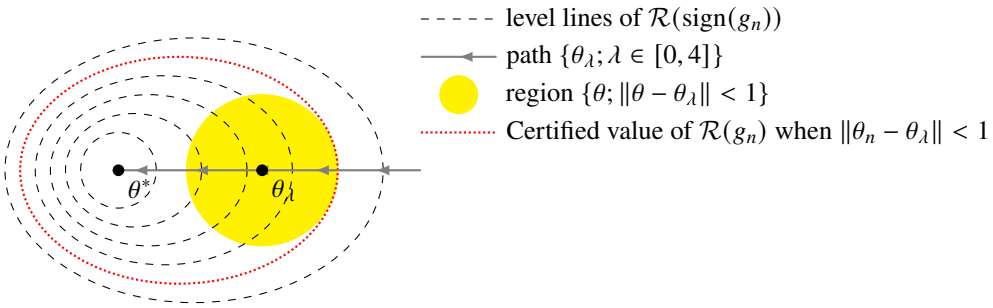

Figure 3.1: Our new convergence analysis consists in relating natural concentration given by the surrogate method to the original excess of risk without passing by the surrogate excess of risk. As the drawing shows, concentration in parameter space Θ can be cast as deviation on the original excess of risk. Yet, this casting relation depends on the geometry of this picture, which itself depends on which surrogate is used, what is the function to learn, how the bias estimator approaches it, and how our empirical estimate concentrates around the bias estimator.

technique is that cascading calibration inequalities can lead to quite suboptimal rates, as shown by the work of Audibert and Tsybakov (2007), which bypassed this procedure, and, under a simple well-thought margin condition, derived dramatically faster convergence rates than classical ones. In our view, this work can be generalized by deriving directly calibration inequality that relates the original excess of risk to concentration in the parameter space. The specificity of those new calibration inequalities is that they are not universal, but depends on some approximation hypothesis that quantifies how hard or easy it is to solve the original problem based on the parametrized surrogate problem. The most well-known such hypothesis is the Tsybakov margin condition, but others were proposed (for example by Steinwart and Scovel, 2007).

Special case of exponential convergence rates. To give more light on the previous paragraph, consider the binary classification problem. Suppose that there exists λ such that $\text{sign}(g_{\theta_\lambda}) = \text{sign}(g^*)$. Suppose also that there exists $\varepsilon > 0$ such that $\|\theta - \theta_\lambda\| \leq \varepsilon$ implies that $\text{sign}(g_\theta) = \text{sign}(g_{\theta_\lambda})$. With those two approximation hypothesis, we have the following calibration inequality

$$\mathcal{R}(\text{sign}(g_\theta)) - \mathcal{R}(f^*) \leq \mathbf{1}_{\|\theta - \theta_\lambda\| > \varepsilon}.$$

As a consequence, $\mathbb{E}_{\mathcal{D}_n}\left[\mathcal{R}(\text{sign}(g_{\theta_{\mathcal{D}_n}}))\right] - \mathcal{R}(f^*) \leq \mathbb{P}_{\mathcal{D}_n}\left(\left\|\theta_{\mathcal{D}_n} - \theta_\lambda\right\| > \varepsilon\right)$. Assuming that this last $\theta_{\mathcal{D}_n}$ concentrated around θ_λ with sub-Gaussian tails, that is $\mathbb{P}_{\mathcal{D}_n}\left(\left\|\theta_{\mathcal{D}_n} - \theta_\lambda\right\| > \varepsilon\right) \leq \exp(-cn\varepsilon^2)$, for a constant c, we get exponential convergence rates.

3.4 Discussion

In this section, we discuss the practical usefulness of generalization bounds, and the different paradigms to define and quantify learnability.

Practical bounds on expected error. In theory, generalization bounds provide guarantees on how much error we might expect when deploying a model learned from data. In practice, those bounds are rather taken as indications that the learning methods are sound, but rates are not reported in order to get confidence on the learned models. This might have to do with the fact that bounds often depend on parameters or constants that are hard to know in practice. As such, practitioners often prefer to derive error indications from test samples. In this line, research on *conformal prediction* is trying to leverage test samples to provide useful confidence information on learned models.

Assumptions to quantify learnability. Given data, it is highly valuable to get an idea of what can be learned from it. Of course, learnability depends on regularity assumptions of the problem at hand. Under such assumptions, lower bounds given by minimax rates are supposed to give optimal baselines for learning rates, thus to quantify learnability. In practice, those lower bounds are derived by considering the most degenerate function respecting those assumptions. Sometimes a simple well-thought additional assumption

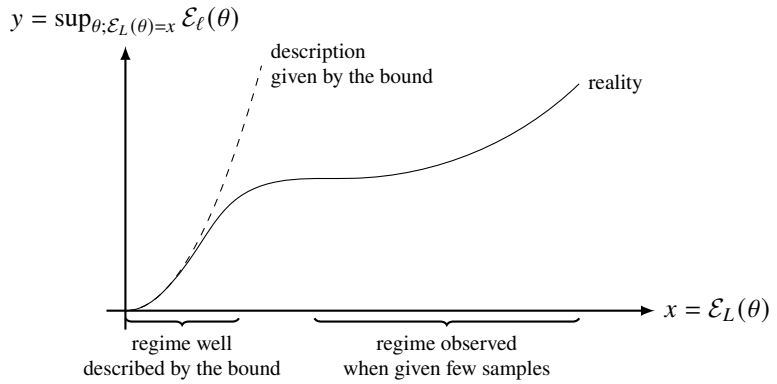

Figure 3.2: The behavior described by generalization bound might be relevant only when accessing an indecently large number of samples, thus being of little use in practice.

can lead to much better bounds. As such, is there a natural paradigm of assumptions on the function to learn to discuss its learnability? For example, assumptions on the surrogate solution g^* are not intrinsic to the original problem but depends on the surrogate problem that is considered. On the contrary, for a fixed decoding d, one can quantify the regularity of f through the most regular function g such that $f = d(g)$ (think of the binary classification case with $d = \text{sign}$), this will correspond to the regularity of the decision frontier, rather than the one of the conditional expectation $x \to \mathbb{E}[Y \mid X = x]$. We refer to the paragraph "Beyond least-squares" in Section 9.A.1 for a practical example of such considerations. This echoes the difference between realizable-consistency and Fisher-consistency of surrogate methods (Long and Servedio, 2013).

Non-asymptoticness of finite-sample bounds? There are two points to mention about probably approximately correct bounds. First they are derived by considering worst-cases, and as we said in the previous paragraph, simple assumptions often allow refining those bounds by removing pathological cases. Second they tend to describe the behavior of how risks decrease with the number of samples when accessing numerous samples. As such, while those bounds are non-asymptotic in the sense that they are valid for any number of samples n, the described regime might not take place without accessing an indecent number of data, thus being of little use in practice.

To get more concrete, consider a loss ℓ and a surrogate loss L. Suppose that we want to bound the excess of risk $\mathcal{E}_\ell(\theta)$ on the original problem with the excess of risk $\mathcal{E}_L(\theta)$ on the surrogate problem, where θ is a learned parameter. The goal of calibration theory is to find the best function ξ such that for any θ, $\mathcal{E}_\ell(\theta) \leq \xi(\mathcal{E}_L(\theta))$. In practice, people try to find ξ concave (because the inequality can be derived for a single x and propagated to the whole space thanks to Jensen inequality), and will look for it under the form $\xi(x) = cx^\alpha$ with $c > 0$ and $\alpha \in (0, 1]$, trying to get the highest α – that is trying to optimize the behavior of ξ close to zero. This will lead to somewhat tight bounds only in the regime where $\mathcal{E}_L(\theta)$ is small, a regime which might not be reached without an indecent number of samples. We illustrate the regime where constants matter more than exponents on Figure 3.2.

In practice, it would be more useful to take into consideration the number of samples first, before looking for a function ξ that verifies $\mathcal{E}_\ell(\theta) \leq \xi(\mathcal{E}_L(\theta))$ for any θ, and minimizes $\max \xi([a, b])$ for $[a, b]$ the range of value that we expect the surrogate risk to take given our fixed number of samples.

Chapter 4

Partially Supervised Learning

Many data scientists spend more time scrapping, annotating and cleaning data than fine-tuning models. This motivates the following question: can we derive a more generic framework than the one of supervised learning in order to learn from cluttered data? In this thesis, this question is approached through the lens of weakly supervised learning, assuming that the bottleneck of data collection lies in data annotation. This chapter summarizes our main contributions to weakly supervised learning. It is based on our published work Cabannes et al. (2020b, 2021b,a), which is reproduced in part III.

4.1 Navigating frameworks of weakly supervised learning

In this section, we give a brief introduction to weakly supervised learning.

Weakly supervised learning is concerned with the setting where it is easy to get n inputs $(X_i)_{i\leq n}$, but it is hard to get the corresponding outputs $(Y_i)_{i\leq n}$, a setting which is relevant for many applications. For example, when dealing with images, one can easily scrap zillions of images on the web, which provides many input data, but getting the corresponding outputs demands laborious annotations of data. The situation is similar when it comes to medical applications, such as understanding drug interactions (Davis et al., 2012) or recognizing cancers with radiography. While one can access databases from hospitals to get many X-ray images, recognizing cancers on those requires the expertise of a radiologist, which is expensive to access. Indeed, annotation costs can critically add up when one plans to use powerful models that require millions of data points to be trained. As a counter example, notice that weakly supervised learning does not help when input data are features coming from heterogeneous sources leading to missing data. We shall discern three types of weak supervision, the last two being not completely disjoint.

1. Global statistics on groups of inputs. This setting consists in accessing global information on bags of samples – *e.g.* knowing that half of the $(Y_i)_{i\in I}$ are zeros for a given $I \subset [n]$. Examples of global statistics supervision include multiple-instance learning (Dietterich et al., 1997) and learning from label proportion (Quadrianto et al., 2009). We refer to those articles for motivations, descriptions and applications of those two problems.

2. Weak classifiers. A second approach consists in assuming the access to many weak classifiers $\varphi_j : \mathcal{X} \to \mathcal{Y}$ for $j \in [p]$ that weakly correlate with f^*. Those classifiers might model labelers from a crowdsourcing platform, experts or noisy measurements. More generally, they might be provided by side information, coming from *distant supervision* (Mintz et al., 2009; Craven and Kumlien, 1999) or *transferred* from algorithms that have been designed for a related task (Pan and Yang, 2010). Recently, Ratner et al. (2020) have implemented a software interface based on this approach.

3. Incomplete annotation. Finally, weak supervision might be understood as the access to partial knowledge on what Y_i is for each $i \in [n]$. In some instances, partial observation on Y_i can be cast as a set of potential labels $S_i \subset \mathcal{Y}$ that are compatible with this partial observation, which is the setting of partial supervision (Cour et al., 2011). Partial supervision is a generalization of semi-supervised learning, which has

been the classical approach to overcome the bottleneck of data annotation (Chapelle et al., 2006) and consists in learning from a set of labeled data, $S_i = \{Y_i\}$ for $i \in L \subset [n]$, and of unlabeled data, $S_i = \mathcal{Y}$ for $i \in [n] \setminus L$.

Beyond those three settings, limitations that motivate weakly supervised learning might be tackled by leveraging human knowledge under the form of priors (Mann and McCallum, 2010) or of function architectures (reviving old approaches of artificial intelligence such as Muggleton and de Raedt, 1994).

4.2 The importance to create consensus

In this section, we review our first approach to the problem of learning from partial supervision. We first formalize this setting before introducing a variational objective to minimize and providing guarantee regarding this approach.

4.2.1 Partial supervision setting

In this thesis, hence in the following, we focus on partial supervision, which is also known as *superset learning*. In other terms, we assume the access to weak supervision for each label y under the form of a set $s \subset \mathcal{Y}$ that contains the true labels $y \in s$. To formalize this problem, we introduce the set $\mathcal{S} \subset 2^{\mathcal{Y}}$ of closed subsets of $\mathcal{Y}$ and the distribution $\tau \in \Delta_{\mathcal{X}\times\mathcal{S}}$ generating samples $(X_i, S_i)_{i\leq n} \sim \tau^{\otimes n}$. The goal is still to minimize the risk

$$\mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\rho}\left[\ell(f(X), Y)\right], \tag{4.1}$$

for a known loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and a distribution on I/O pairs $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$. In contrast with supervised learning, ρ will never be accessed, nor any samples $(X_i, Y_i) \sim \rho$.

In order to motivate our theoretical setting, we now discuss simple instances of partial supervision.

Example 5 (Classification with attributes)**.** *Suppose that you want to learn fine-grained classes on images,* e.g. *to precisely distinguish "caracals" and "domestic cats", as well as "hedgehog mushrooms" and "shiitakes". It will probably be hard for commoners hired on crowdsourcing platforms to label images with precise classes. Yet, this commoner might easily label attributes, such as "tufted ears" characterizing "caracals" among "felines", or "spines rather gills underside of the cap" and "tan irregular cap" characterizing "hedgehogs" among "mushrooms". Those partial labels can easily be cast as sets of labels: "feline" would be {"lion", "panther", ...} and "tufted ears" be {"great horned owl", "Araucana chicken", ...}. The key idea here is that it is much cheaper to get a set of many partial labels that is informative enough to learn f^* than a set of few complete labels enabling the same learning.*

Example 6 (Ranking with partial ordering)**.** *Consider a ranking problem where, given a user characterized by some features x, the goal is to learn their preference over m flowers, defining the output $y \in \mathfrak{S}_m$ as an ordering of m elements. Once again, getting the full y is a laborious task, as one might have a hard time to rank all m flowers. In contrast, one might easily provide partial ordering information, such as "I prefer roses to lilies", or "tulips are my favorite", which can be cast as the set of total orderings that verify those partial orderings.*

Example 7 (Regression with censored data)**.** *For many regression problems, it is not possible to get exact labels but only possible to access bins or censored labels $[a, b] \ni y$. This could be due to the use of a measuring scale, or, if retaking example 3, due to some uncertainty regarding a price at which a house has been sold.*

4.2.2 Solution definition through the infimum loss

Currently, the problem (ℓ, τ) is ill-defined since we cannot define clearly the goal (4.1) from those two objects only – ℓ being a loss function on pairs of outputs and τ a distribution on (X, S). One way to redefine a solution in this context is to keep the precedent variational point of view and define a loss $L : \mathcal{Y} \times 2^{\mathcal{Y}} \to \mathbb{R}$ that given a prediction $z \in \mathcal{Y}$ and an observation $S \subset \mathcal{Y}$, provides a compatibility or performance score. Hence the new objective

$$\mathcal{R}_S(f) = \mathbb{E}_{(X,S)\sim\tau}\left[L(f(X), S)\right].$$

Yet, how to derive the loss $L(z, S)$ for $z \in \mathcal{Y}$ and $S \subset \mathcal{Y}$ when the original task was specified by the loss $\ell(z, y)$ with $y \in S$? Arguably there are three possibilities.

Figure 4.1: Two different figures of partial supervision in $\mathcal{Y} = \mathbb{R}^2$ without input (or $\mathcal{X}$ being a singleton). The loss is given by the mean square error, *i.e.* the usual Euclidean geometry. On both figures, we represent $\mathcal{Y}$ with two sets, S_1 in light gray and S_2 in gray, and the estimates $z = \arg\min_{z \in \mathcal{Y}} \{L(z, S_1) + L(z, S_2)\}$ provided by the infimum loss (z^i) and the average loss (z^a). The point y_1^a represents the disambiguation of S_1 by the average loss. On the left figure, although S_1 intersects S_2, only the infimum loss verifies $z^i \in S_1 \cap S_2$.

1. Bounding the original excess risk from above with the supremum loss $L(z, S) = \sup_{y \in S} \ell(z, y)$. Suprema have been used in statistics for robustness purposes (Wald, 1945). The idea here is that if we prevent ourselves against the worst possible candidates (in terms of error given a prediction z), we will prevent ourselves against whatever is the ground truth label in S. Yet, as we will discuss later, this approach is too conservative in our setting.
2. Matching a bit of all the elements in the set S with the average loss $L(z, S) = \frac{1}{|S|} \sum_{y \in S} \ell(z, S)$. This loss has the benefit of being agnostic on what should be the true $y \in S$ regardless of what the prediction z is. Under symmetry assumptions of the original loss, for example when ℓ is the 0-1 loss in classification, averaging candidates is a reasonable solution. When the loss is not symmetric, it can insidiously bias the solution as we describe on Figure 7.1. In many cases, it is possible to correct for this asymmetry, which techniques implicitly implied by Proposition 59 and illustrated with the filling with zero techniques in section 8.5.4.
3. Making sure that the prediction z matches at least one element with the infimum loss $L(z, S) = \inf_{y \in S} \ell(z, S)$. When ℓ is seen as a distance, L is its natural extension to sets. Hüllermeier (2014) has referred to it as the optimistic loss, since given a prediction z, it somehow disambiguates $y \in S$ by considering the best label in order to minimize $\ell(z, y)$ under the constraints $y \in S$.

Let us take a step back and recall the commoner providing the set "feline" that is $S_1 = \{$"cat", "lion", $\dots\}$ and the set "tufted ears", that is $S_2 = \{$"owl", "Araucana chicken", $\dots\}$, when annotating the picture of a "caracal". Naturally, we would like to output a prediction $z \in S_1 \cap S_2 = \{$"caracal"$\}$ of a feline with tufted ears. In other terms, we want to create consensus between observations, which is what the infimum loss provides. The supremum loss is outcast from the good solutions, since it might disambiguate $y_1 \in S_1$ as "leopard" and $y_2 \in S_2$ as "owl", as well as the average loss as we illustrated on Figure 4.1.

4.2.3 Theoretical guarantees

With the infimum loss, we switch to the original supervised learning problem (4.1) to the new problem defined through the risk

$$\mathcal{R}_S(f) = \mathbb{E}_{(X,S)\sim\tau} \left[\inf_{y \in S} \ell(f(X), y) \right]. \tag{4.2}$$

How are those two problems related? In particular, we would like to make sure that solving the partially supervised problem does help to tackle the fully supervised one. Following the formalism of the previous chapter, we can try to derive a calibration inequality that relates the excess of the original risk (4.1) to the excess of the "surrogate" risk (4.2). To derive such a generic inequality, we need strong assumptions, which are usually given by ensuring non-ambiguity of the partial supervision (Definition 34), that is for each x, the sets in the support of the conditional distribution $(S \mid X = x)$ intersect into a singleton $\{y_x\} = \{f^*(x)\}$. Such an inequality was first provided by Cour et al. (2011) in the case of classification with the 0-1 loss, and is generalized to generic problems by Proposition 37. This technique allows us to derive rates on a structured prediction algorithm learning from partially supervised data in Theorem 14. In practice, given samples (X_i, S_i), the practitioner can use their favorite algorithm to translate this population principle into an empirical objective to be optimized in order to get an estimate f_n of the risk minimizer.

Remark 4 (A false case against partial supervision)**.** *Partially supervised learning and the infimum loss are often criticized because of the strength of the non-ambiguity assumption. This should rather be seen as a limitation of theoretical analysis, rather than a limitation of the framework. Actually, a more careful analysis based on the idea provided in the previous chapter shows that the non-ambiguity assumption is akin to the Massart noise condition, and can be used to derive much faster rates without modifying algorithms as shown by Theorem 15. This suggests that there are many interesting behaviors to characterize beyond the non-ambiguity assumption.*

4.3 Label disambiguation to complete supervision

In this section, we characterize the solution provided by the infimum loss as a disambiguation strategy, that consists in retrieving full supervision first, before learning from it. This leads to a different perspective in terms of learning and practical implementations.

4.3.1 Expected testing distribution

Let us roll back a little and think back on which solution to look for when in presence of partial supervision. The ultimate goal is to find a mapping that minimizes the generalization error on the testing distribution $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$. Yet which distribution ρ should we expect given the loss ℓ and the distribution $\tau \in \Delta_{\mathcal{X}\times 2^{\mathcal{Y}}}$? The only clearly defined constraints is that the distribution should be compatible with the weak information we have seen on the data. The notion of compatibility can be formalized through the following definition, which we introduced in Cabannes et al. (2020b).

Definition 5 (Compatibility)**.** *Two distributions $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ and $\tau \in \Delta_{\mathcal{X}\times\mathcal{S}}$ are said to be* compatible *if there exists an* underlying distribution *of probability $\pi \in \Delta_{\mathcal{X}\times\mathcal{Y}\times\mathcal{S}}$ such that ρ and τ are the respective marginals of π according to $\mathcal{X}\times\mathcal{Y}$ and $\mathcal{X}\times\mathcal{S}$ and such that for any $(x, y, s) \in \operatorname{supp}\pi$, we have $y \in s$. This compatibility property will be denoted by $\rho \vdash \tau$.*

To discriminate the expected ρ among all the distributions compatible with τ, we need a principle. It is natural to look for a distribution that satisfies some properties. As mentioned in the last section, we would like to go for a distribution that is "consensual", *i.e.* we would like to reduce the noise of the conditional $(Y \mid X)$. In other terms, we would like those conditional distributions to be as deterministic as possible. As a consequence, we define the expected testing distribution as

$$\rho^\star \in \operatorname*{arg\,min}_{\rho\vdash\tau} \mathcal{E}(\rho),$$

for $\mathcal{E} : \Delta_{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R}$ a measure of determinism. A well-known measure of determinism is the entropy $\mathcal{E}(\rho) := \mathbb{E}[-\log(\rho(X, Y))]$. This measure is independent of the loss ℓ, which might be a desirable property if one would like to reuse the same supervised distribution for different tasks linked with different losses. Yet, when there is a structure on $\mathcal{Y}$ such that its explicit dimension is much bigger than the intrinsic dimension of this space seen through the loss ℓ, the entropy will scale with the explicit dimension. For example, in ranking problems where the goal is to output an ordering over m items, which is similar to a permutation in $\mathfrak{S}_m$, the output space is of cardinality $m!$, while the intrinsic dimension is m. For those problems, it is wise to incorporate the loss in the objective, in particular, we suggest the following loss-based variance $\mathcal{E}(\rho) := \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_\rho[\ell(f(X), Y)]$. Once the distribution ρ is recovered, a function f can be learned in a supervised learning fashion

$$f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,Y)\sim\rho^\star}[\ell(f(X), Y)], \qquad \text{with} \qquad \rho^\star = \operatorname*{arg\,min}_{\rho\vdash\tau} \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_\rho[\ell(f(X), Y)]. \tag{4.3}$$

Remarkably, this second solution to the partially supervised learning problem is exactly the same as the one provided by the infimum loss. This is based on the fact that, under really mild definition assumptions

$$\inf_{\rho\vdash\tau} \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_\rho[\ell(f(X), Y)] = \inf_{f:\mathcal{X}\to\mathcal{Y}} \inf_{\rho\vdash\tau} \mathbb{E}_\rho[\ell(f(X), Y)] = \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_\tau[\inf_{y\in S} \ell(f(X), y)].$$

A nice property of this new characterization of the solution f^* is that it goes beyond the partial labeling problem. As a matter of fact, (4.3) can be applied more generally to any weakly supervised learning paradigm where the weak information acquired on the problem can be translated into hard constraints, *e.g.* in the case of label proportion.

4.3.2 Optimization issues

Before going any further, let us point out that any objective $\mathcal{E}(\rho)$ that is minimized for deterministic distributions is intrinsically not-convex. Consider the case where there is no input space (or $\mathcal{X}$ is a singleton), $\mathcal{Y}$ is discrete, and $\mathcal{E}$ is a function taking as argument elements of the simplex $\Delta_{\mathcal{Y}}$ and outputting a real-valued score. If $\mathcal{E}$ is minimized for deterministic distributions, it is minimized on extremities of the simplex. In other terms, $\mathcal{E}$ should really be thought of as a concave function, making its minimization a non-convex optimization problem. We picture level lines of such an objective on Figure 8.4.

Regarding the constraints, it should be noted that $\{\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}} \mid \rho \vdash \tau\}$ is a convex set. So we are trying to minimize a concave function over a convex set. In Cabannes et al. (2021b), we suggested two different heuristics to approach this minimization procedure. The first one consists in starting with a good "center point" and finding an extremity of the definition domain after gradient descent. Our notion of center takes into account the loss to avoid insidious bias (similarly to what we discussed the average loss). The second one is based on the Diffrac algorithm (Bach and Harchaoui, 2007), and consists in convexifying the objective by adding a quadratic term that is constant on extremities, perform gradient descent, and get a solution that is a convex combination of extreme points. We refer to sections 8.5 and 8.B for further discussions on those techniques.

4.3.3 Empirical objective

Given data $(X_i, S_i)_{i\leq n} \sim \tau^{\otimes n}$, how to translate the disambiguation principle (4.3) into an empirical objective allowing to disambiguate the sets $(S_i)_{i\leq n}$ into labels $(\hat{Y}_i)_{i\leq n}$? To answer this question, we need to translate an objective on distribution into an objective on samples. The usual translation of risk minimization into empirical risk minimization can be thought as the approximation $\rho \approx \hat{\rho} = \frac{1}{n}\sum_{i\leq n} \delta_{X_i} \otimes \delta_{Y_i}$ plus some constraints $f \in \mathcal{F}$ on the functions that we minimize. As such, we could go for

$$\min_{(y_i\in S_i)_{i\leq n}} \inf_{f\in\mathcal{F}} \sum_{i=1}^{n} \ell(f(X_i), y_i).$$

This will be really hard to solve in practice. The solution we suggested is based on kernel mean embedding (Muandet et al., 2017), which is a natural way to understand the surrogate approach of Ciliberto et al. (2016). It consists in incorporating directly the hypothesis on the solution of the problem in the estimated distribution with $\hat{\rho} \approx \frac{1}{n}\sum_{i,j\leq n} \delta_{X_i} \otimes \alpha_j(X_i)\delta_{Y_j}$, where $\alpha : \mathcal{X} \to \mathbb{R}^n$ is a weighting scheme that states how to diffuse information $(Y_i)_{i\leq n}$ seen at $(X_i)_{i\leq n}$ to any point $x \in \mathcal{X}$. In particular, those weights might be used to estimate the conditional distribution $(Y|X = x)$ as $\sum_{j\leq n} \alpha_j(x)\delta_{Y_j}$. This leads to the following empirical objective

$$(\hat{y}_i)_{i\leq n} \in \operatorname*{arg\,min}_{(y_i\in S_i)_{i\leq n}} \inf_{(z_i\in\mathcal{Y})_{i\leq n}} \sum_{i,j=1}^{n} \alpha_i(X_j)\ell(z_i, y_j), \tag{4.4}$$

In the case of kernel ridge regression, those weights are given by $\alpha(x) = (K + \lambda n)^{-1}K_x$ with the notations of section 2.4.2. But they could also be given by similarity metrics, or derived with more advanced unsupervised techniques such as the one we present in Cabannes et al. (2021a) as we detail in section 9.A.

4.3.4 Theoretical guarantees

The convergence of an estimate f_n learned on the dataset $(X_i, \hat{y}_i)_i$ with $(\hat{y}_i)_{i\leq n}$ recovered through (4.4) cannot easily be performed with the tools presented so far. Indeed, it is hard to understand how the different disambiguation $(y_i \in S_i)$ interacts when studying (4.4). This relates to the non-concavity of this objective: a small change in the optimal $(z_i)_{i\leq n}$ can lead to completely different optimal $(y_i)_{i\leq n}$. While there are probably good tools to think in terms of distribution and derive a consistent empirical estimate of (4.3), in Cabannes et al. (2021b), we used the fact that usual learnability assumptions for the partial labeling problem are actually quite strong, which allows us to prove impressively fast rates.

To derive consistency of algorithm learning with partially supervised data, it is classical to assume non-ambiguity of the distribution $(S \mid X = x)$ as well as some regularity of the function $\mathcal{X} \to \Delta_{\mathcal{Y}}; x \to (Y \mid X = x)$. Those assumptions actually imply that the data are clustered by groups where labels are all equal. When this class separation hypothesis holds true, one can leverage this structure by considering a weighting scheme

based on nearest neighbors. As we discussed in more details in section 8.4, this endows the estimate f_n with exponential convergence rates, meaning that the risk $\mathcal{R}(f_n)$ converges toward the minimum risk exponentially fast as the number of samples n goes to infinity. This is the statement of Theorem 15.

4.3.5 Can collaborative filtering help us to deal with weak supervision?

There is a natural link between completion of weak information and the collaborative filtering problem described in section 2.3.3. To make this link we will need to introduce more concepts.

Assume that the loss $\ell : \mathcal{Y} \to \mathcal{Y} \to \mathbb{R}$ can be written as $\ell(z, y) = -\langle \psi(z), \varphi(y) \rangle$ with $\psi, \varphi : \mathcal{Y} \to \mathcal{H}$ two embeddings into a Hilbert space $\mathcal{H}$, encoding the structure of the loss. For example, the Kendall loss in ranking corresponds to the correlation measure $\varphi = -\psi = (\mathbf{1}_{\sigma(i)>\sigma(j)})_{i,j\leq m}$ for $\sigma \in \mathcal{Y} = \mathfrak{S}_m$ a permutation. We refer to Nowak-Vila et al. (2019) for more examples. Assume also that the weak supervision correspond to observing $A_i\varphi(Y_i)$ instead of $\varphi(Y_i)$ for a known masking operator $A_i : \mathcal{H} \to \mathcal{H}$. For example, observing a partial ordering that σ should verify is equivalent to observing some coordinates of the vector $\varphi(\sigma)$. The disambiguation problem can be reduced to the matrix completion problem of retrieving $(\varphi(y_i))_{i\leq n}$ from the observations $(A_i\varphi(Y_i))_{i\leq n}$. Our algorithm suggests solving the minimization problem

$$\begin{array}{ll} \text{minimize} & \left\| \begin{pmatrix} | & & | \\ \psi(z_1) & \cdots & \psi(z_n) \\ | & & | \end{pmatrix} \begin{pmatrix} & \alpha_i(X_j) & \end{pmatrix} \begin{pmatrix} | & & | \\ \varphi(y_1) & \cdots & \varphi(y_n) \\ | & & | \end{pmatrix}^\top \right\|_F^2 \\ \text{subject to} & A_i\varphi(y_i) = A_i\varphi(Y_i). \end{array}$$

In the case where $\psi = -\varphi$ and there is no context variables, hence $\alpha_i(X_j) = 1$, the slightly different measure of determinism exposed in section 8.B leads to the problem

$$\begin{array}{ll} \text{minimize} & -\left\|\sum_{i=1}^n \varphi(y_i)\right\|^2 \\ \text{subject to} & A_i\varphi(y_i) = A_i\varphi(Y_i). \end{array}$$

For correlation losses, it is usual for $\|\varphi(\cdot)\|$ to be constant, so to minimize the last objective, one should try to align all $\varphi(y_i)$ in one direction. This differs from the collaborative filtering solution which completes the $(\varphi(y_i))$ by minimizing the cardinality of the set $\{\varphi(y_i) \mid i \in [N]\}$. Formally, collaborative aims at solving the following problem

$$\begin{array}{ll} \text{minimize} & \operatorname{rank} \begin{pmatrix} | & & | \\ \varphi(y_1) & \cdots & \varphi(y_n) \\ | & & | \end{pmatrix} \\ \text{subject to} & A_i\varphi(y_i) = A_i\varphi(Y_i), \end{array}$$

which is practically implemented with the nuclear norm instead of the rank. In essence, the infimum loss builds an understanding of $(Y \mid X)$ by collapsing all observations onto a single vector in $\varphi(\Delta_\mathcal{Y})$, while collaborative filtering builds such an understanding by minimizing the number of vectors in $\varphi(\mathcal{Y})$ to complete observations without discrepancy. Bining observations into several groups makes sense when context variables do not allow for inputs to characterize outputs univocally.

Further investigations of the link between our work and collaborative filtering is a research direction that we left open. We conjecture that this link might prove useful to incorporate context variables into collaborative filtering techniques.

4.4 Leveraging input structure

In this section, we discuss unsupervised learning techniques that could be incorporated in the frameworks described previously in order to leverage unlabeled input data.

4.4.1 The case of semi-supervised learning

The infimum loss and the disambiguation framework described previously are based on a pointwise principle that discriminates a distribution $(Y \mid X)$ from a weakly supervised distribution $(S \mid X)$. Such a principle fails to provide any guideline for semi-supervised learning, which is a specific instance of partial supervision where S is either a singleton or the full set $\mathcal{Y}$. In particular for unsupervised parts of the input space, which

correspond to points in the support of $(X \mid S = \mathcal{Y})$ that do not belong to the support of $(X \mid |S| = 1)$, this principle does not provide any information on what $f^*(x)$ should be.

Eventually, the solution on those points could be inferred from the solution on the other points by looking for the simplest function completion with respect to some criterion quantifying how "simple" a function is. For regression problems, if one searches to minimize the empirical risk in a Banach space of function, the norm of this Banach space provides a simple criterion for "simplicity". More generally, in metric space of function, this norm can be replaced more generally as the distance between a neutral function (*e.g.* the null function in a vector space) and the function itself.

4.4.2 Laplacian regularization

The classical approach to solve semi-supervised learning regression problem is to choose the criterion of "simplicity" as the Dirichlet energy $f \to \mathbb{E}_{\rho_{\mathcal{X}}}[\|\nabla f(X)\|^2]$. This regularization is similar to the square norm of the Hilbert space of functions $W^{1,2}(\rho_{\mathcal{X}})$, which is the weighted Sobolev space of functions with weak derivative endowed with the probability measure $\rho_{\mathcal{X}}$. Considering a square of the norm rather than the norm itself is classical in machine learning, making analysis as well as computations easier. Measuring regularity through the probability measure $\rho_{\mathcal{X}}$ rather than the Lebesgue measure is crucial in order to leverage the unsupervised data.

This regularization has been introduced in the seminal paper of Zhu et al. (2003) which estimates the Dirichlet energy through finite difference methods akin to the Nadayara-Watson estimator. This regularization formalized many intuitions. For regression problems, it is natural to assume that the target function does not vary too much on densely populated input regions, or said otherwise, that its variations concentrate on regions without too much data. For classification problems, this echoes the low-density separation hypothesis, assuming that decision frontiers, that is frontier between classes in the input space, lies on sparsely populated regions. In Cabannes et al. (2021a), we proposed a "kernelized" version of this estimation in order to bypass the curse of dimensionality under regularity assumptions, together with some clever low-rank factorization in order to reduce computation time to a decent amount. It should be noted that Laplacian regularization has also a natural interpretation in terms of diffusion, which is captured by the Langevin dynamic, and explains the wording "label propagation" found in semi-supervised literature.

4.4.3 Complete framework

With Laplacian regularization, one can deepen the *lex parsimoniae* of the precedent sections when this rule does not discriminate a unique distribution by looking for the distribution that will minimize the Dirichlet energy of its (or a continuous surrogate) risk minimizer. In particular, this approach is interesting as it could remove the usual non-ambiguity assumption usually made to define and study solutions of the partial supervision problems (Cour et al., 2011; Luo and Orabona, 2010; Liu and Dietterich, 2014; Cabannes et al., 2020b, 2021b). Yet, removing such an assumption will come at the price of adding assumptions on the surrogate problem we are solving that might be harder to verify in practice.

Finally, it should be noted that Laplacian regularization can be introduced as soon as during the preprocessing/feature engineering part of a practical machine learning pipeline. In particular, it provides principled guidelines to retrieve and parametrize a small manifold on which the input data might lie, and might prove itself useful to strengthen self-supervision techniques. At this point, it is less a theory than experiments that could confirm those intuitions.

4.5 Active labeling

Weakly supervised learning is concerned with retrieving a target function under weak observations. It is often motivated by the bottleneck of annotating data. Yet it fails to answer the most simple question practitioners might ask themselves: how to collect the most discriminative dataset to learn this function under some cost constraints. Hence, the last part of this thesis is devoted to the "active" collection of weak supervision. Active refers to the fact that we will iteratively and adaptively search for annotations, in contrast with passive settings in which annotations are given once and for all by an exterior mechanism.

The cost of dataset annotation depends on different task specifications. Hence, a model of annotation cost can only describe a limited number of situations. As well as we split our understanding of weakly

supervised learning in three parts (group statistics, weak classifiers, and labels corruption), we could split dataset annotation with the same three categories. In practice, it is frequent to annotate dataset by grouping input data according to a classifier output before identifying the most present class in each group and filtering mistakes, which allows annotating large chunks of data at once. In a similar fashion, ImageNet was collected by searching for different categories on image search engines, before spotting outliers in a batch of images resulting from a single query (Deng et al., 2009).

The final part of this thesis touches upon the active variant of partially supervised learning (Cabannes et al., 2022). In particular, we assume that given data $(X_i)_{i\leq n}$, one can query any information of the type $\mathbf{1}_{Y_{i_t}\in S_t}$ for $i_t \in [n]$ an index to choose and $S_t \in \mathcal{S} \subset 2^{\mathcal{Y}}$ a set to choose among certain subsets of labels. We refer the reader to Part IV for additional considerations on the matter.

Part II

Considerations on Learning Theory

Chapter 5

Fast Rates for Structured Prediction

The following is a reproduction of Cabannes et al. (2021c).

Discrete supervised learning problems such as classification are often tackled by introducing a continuous surrogate problem akin to regression. Bounding the original error, between estimate and solution, by the surrogate error endows discrete problems with convergence rates already shown for continuous instances. Yet, current approaches do not leverage the fact that discrete problems are essentially predicting a discrete output when continuous problems are predicting a continuous value. In this paper, we tackle this issue for general structured prediction problems, opening the way to "superfast" rates, that is, convergence rates for the excess risk faster than n^{-1}, where n is the number of observations, with even exponential rates with the strongest assumptions. We first illustrate it for predictors based on nearest neighbors, generalizing rates known for binary classification to any discrete problem within the framework of structured prediction. We then consider kernel ridge regression where we improve known rates in $n^{-1/4}$ to arbitrarily fast rates, depending on a parameter characterizing the hardness of the problem, thus allowing, under smoothness assumptions, to bypass the curse of dimensionality.

5.1 Introduction

Machine learning is raising high hopes to tackle a wide variety of prediction problems, such as language translation, fraud detection, traffic routing, speech recognition, self-driving cars, DNA-binding proteins, *etc*. Its framework is appreciated as it removes humans from the burden to come up with a set of precise rules to accomplish a complex task, such as recognizing a cat on an array of pixels. Yet, it comes at a price, which is of forgetting about algorithm correctness, meaning that machine learning algorithms can make mistakes, *i.e.*, wrong predictions, which can have dramatic implications, *e.g.*, in medical applications. This motivates work on generalization error bounds, quantifying how often one should expect errors.

Many of the problems discussed above are of discrete nature, in the sense that the number of potential outputs is finite, or infinite countable. To learn such problems, a classical technique consists in defining a continuous surrogate problem, which is easier to solve, and such that:

(1) an algorithm on the surrogate problem translates into an algorithm on the original problem;

(2) errors on the original problem are bounded by errors on the surrogate problem.

The first point refers to the concept of plug-in algorithms, while the second point to the notion of calibration inequalities. For example, binary classification can be approached through regression by estimating the conditional expectation of the output Y given an input X (Bartlett et al., 2006).

On the one hand, continuous surrogates for discrete problems are interesting, as they benefit from functional analysis knowledge, when discrete problems are more combinatorial in nature. On the other hand, continuous surrogates can be deceptive, as they are asking to solve for more than needed. Considering the example of binary classification, where $Y \in \{-1, 1\}$, one only has to predict the sign of the conditional expectation, rather than its precise value. Interestingly, without modifying the continuous surrogate approach, this last remark can be leveraged in order to tighten generalization bounds derived through calibration inequalities (Audibert and Tsybakov, 2007). In this work, we extend those considerations, known in binary classification (*e.g.*, Koltchinskii and Beznosova, 2005; Chaudhuri and Dasgupta, 2014), to generic discrete supervised learning problems, and show how it can be applied to the kernel ridge regression algorithm introduced by Ciliberto et al. (2016).

5.1.1 Contributions

Our contributions are organized in the following order.

- In Section 5.2, we consider the general structured prediction from Ciliberto et al. (2020) based on Lipschitz-continuous losses and derive refined calibration inequalities to leverage the fact that learning a mapping into a discrete output space is easier than learning a mapping into a continuous space.
- In Section 5.3, we show how to exploit exponential concentration inequalities to turn them into fast rates under a condition generalizing the Tsybakov margin condition.
- In Section 5.4, we apply Section 5.3 to local averaging methods with the particular example of nearest neighbors. This leads to extending the rates known for regression and classification to a wide variety of structured prediction problems, with rates that match minimax rates known in binary classification.
- In Section 5.5, we show how Section 5.3 can be applied to kernel ridge regression. This allows us to improve rates known in $n^{-1/4}$ to arbitrarily fast rates depending on the hardness of the associated discrete problem.

5.1.2 Related work

Surrogate framework. The surrogate problem we will consider to tackle structured prediction finds its roots in the approximate Bayes rule proposed by Stone (1977), analyzed through the prism of mean estimation as suggested by Friedman (1994) for classification, and analyzed by Ciliberto et al. (2020) in the wide context of structured prediction. In particular, we will specify results on two classes of surrogate estimators: local averaging methods, or kernel ridge regression.

Local averaging methods. Neighborhood methods were first studied by Fix and Hodges (1951) for statistical testing through density estimation. Similarly, Parzen–Rosenblatt window methods (Parzen, 1962; Rosenblatt, 1956) were developed. Those methods were cast in the context of regression as nearest neighbors (Cover and Hart, 1967) and Nadayara-Watson estimators (Watson, 1962; Nadaraya, 1964). Stone (1977) was the first to derive consistency results for a large class of localized methods, among which are nearest neighbors and some window estimators (Spiegelman and Sacks, 1980; Devroye and Wagner, 1980). Rates were then derived, with minimax optimality (Stone, 1980; Yang, 1999). Several reviews can be found in the literature, such as Györfi et al. (2002); Tsybakov (2009); Biau and Devroye (2015); Chen and Shah (2018).

Reproducing kernel ridge regression. The theory of real-valued reproducing kernel Hilbert spaces was formalized by Aronszajn (1950), before finding applications in machine learning (*e.g.*, Scholkopf and Smola, 2001). Minimax rates for kernel ridge regression were achieved by casting the empirical solution estimate as a result of integral operator approximation (Smale and Zhou, 2007; Caponnetto and De Vito, 2006), allowing to control convergence through concentration inequalities in Hilbert spaces (Yurinskii, 1970; Pinelis and Sakhanenko, 1986) and on self-adjoint operator on Hilbert spaces (Minsker, 2017). First derived in L^2-norm, rates were cast in L^∞-norm through interpolation inequalities (*e.g.*, Fischer and Steinwart, 2020; Lin et al., 2020).

Tsybakov margin condition. Learning a mapping into a discrete output space is indeed easier than learning a continuous mapping, as, for binary classification for example, one typically only has to predict the sign of $\mathbb{E}[Y|X]$ rather than its precise value. As such, calibration inequalities that relate the error on a discrete structured prediction problem to an error on a smooth surrogate problem are often suboptimal. This phenomenon was exploited for density discrimination, a problem consisting of testing if samples were drawn from one or the other of two potential distributions, by Mammen and Tsybakov (1999), and for binary classification by Audibert and Tsybakov (2007). Those works introduce a parameter $\alpha \in [0, \infty)$ characterizing the hardness of the discrete problem, and leverage concentration inequalities to accelerate rates known for regression by a power $\alpha + 1$ (Audibert and Tsybakov, 2007), while rates plugged-in directly through calibration inequalities only present an acceleration by a power $2(\alpha+1)/(\alpha+2)$ (see, *e.g.* Boucheron et al., 2005; Bartlett et al., 2006; Bartlett and Mendelson, 2006; van Erven et al., 2015; Nowak-Vila et al., 2019).

5.2 Structured prediction with surrogate control

In this section, we introduce the classical supervised learning problem, and a surrogate problem that consists of conditional mean estimation. We recall a calibration inequality relating the original problem to the surrogate one. We mention how empirical estimations of the conditional means usually deviate from the real means following a sub-exponential tail bound, similarly to bounds obtained through Bernstein inequality. We end this section by providing refined surrogate control, that is the key toward "superfast" rates, that is, rates faster than $1/n$.

5.2.1 Surrogate mean estimation

Consider a classic supervised learning problem, where given an input space $\mathcal{X}$, an observation space $\mathcal{Y}$, a prediction space $\mathcal{Z}$, a joint distribution $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ and a loss function $\ell : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}_+$, one would like to retrieve $f^* : \mathcal{X} \to \mathcal{Z}$ minimizing the risk $\mathcal{R}$.

$$f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Z}} \mathcal{R}(f) \qquad \text{with} \qquad \mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\rho}\left[\ell(f(X), Y)\right].$$

In practice, $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ and ℓ are givens of the problem, while ρ is unknown, yet partially observed thanks to a dataset $\mathcal{D}_n = (X_i, Y_i)_{i\le n} \sim \rho^{\otimes n}$, with data (X_i, Y_i) sampled independently from ρ. Note that in fully supervised learning, the observation space is the same as the prediction space $\mathcal{Y} = \mathcal{Z}$, yet we distinguish the two for our results to stand in more generic settings, such as instances of weak supervision (Cabannes et al., 2020b). In the following, we consider $\mathcal{Z}$ finite. In several cases, solving the supervised learning problem can be done through solving a surrogate problem that is easier to handle. Ciliberto et al. (2016) provides a setup that reduces a wide variety of structured prediction problems (ℓ, ρ) to a problem of mean estimation. It works under the following assumption.

Assumption 1 (Bilinear loss decomposition)**.** *There exists a Hilbert space $\mathcal{H}$ and two mappings $\psi : \mathcal{Z} \to \mathcal{H}$, $\varphi : \mathcal{Y} \to \mathcal{H}$ such that*

$$\ell(z, y) = \langle \psi(z), \varphi(y) \rangle .$$

We will also assume that ψ is bounded (in norm) by a constant c_ψ.

This assumption is not really restrictive (Ciliberto et al., 2020). Among others, it works for any losses on finite spaces, usually with spaces $\mathcal{H}$ whose dimensionality is only polylogarithmic with respect to the cardinality of $\mathcal{Z}$ (Nowak-Vila et al., 2019). Under Assumption 1, solving the supervised learning problem can be done through estimating the surrogate conditional mean $g^* : \operatorname{supp} \rho_{\mathcal{X}} \to \mathcal{H}$, defined as

$$g^*(x) = \mathbb{E}_{Y\sim\rho|_x}\left[\varphi(Y)\right], \tag{5.1}$$

where we denote $\rho|_x$ the conditional law of $(Y \mid X)$ under $(X, Y) \sim \rho$.

Lemma 6 (Ciliberto et al. (2016))**.** *Given an estimate g_n of g^* in* (5.1)*, consider the estimate $f_n : \mathcal{X} \to \mathcal{Z}$ of f^*, which is obtained from "decoding" g_n as*

$$f_n(x) = \operatorname*{arg\,min}_{z\in\mathcal{Z}} \langle \psi(z), g_n(x) \rangle . \tag{5.2}$$

Then the excess risk is controlled through the surrogate error as

$$\mathcal{R}(f_n) - \mathcal{R}(f^*) \le 2c_\psi \left\| g_n - g^* \right\|_{L^1(\mathcal{X},\mathcal{H},\rho)} . \tag{5.3}$$

Inequalities relating the original excess risk $\mathcal{R}(f_n) - \mathcal{R}(f^*)$ with a measure of error on a surrogate problem are called *calibration inequalities*. They are useful when the measure of error between g_n and g^* is easier to control than the one between f_n and f^*.

Example 8 (Binary classification)**.** *Binary classification corresponds to $\mathcal{Y} = \mathcal{Z} = \{-1, 1\}$ and $\ell(z, y) = \mathbf{1}_{z\neq y}$ (or equivalently $\ell(z, y) = 2\mathbf{1}_{z\neq y} - 1$). The classical surrogate consists of taking $\mathcal{H} = \mathbb{R}$, with $\varphi = id$ and $\psi = -id$. In this setting, we have $g^*(x) = \mathbb{E}_\rho[Y|X = x]$, and the decoding $f_n(x) := \operatorname{sign} g_n(x)$, for any $g_n(x) \in \mathcal{H}$. In this case $\mathcal{R}(f_n) - \mathcal{R}(f^*) = \mathbb{E}_X\left[\mathbf{1}_{f_n(X)\neq f^*(X)} \left|g^*(X)\right|\right] \le 2\left\|g_n - g^*\right\|_{L^1} \le 2\left\|g_n - g^*\right\|_{L^2}$. Note that in regression the excess risk reads as the square of the L^2 norm, explaining a loss of a power one half in convergence rates, when going from regression to classification (*e.g. *Chen and Shah, 2018).*

Differences between an empirical estimate and its population version are generally handled through concentration inequalities. In this work, we will leverage concentration on $\|g_n(x) - g(x)\|$ that is uniform for $x \in \operatorname{supp} \rho_{\mathcal{X}}$, motivating the introduction of Assumption 2.

Assumption 2 (Exponential concentration inequality)**.** *Suppose that for $n \in \mathbb{N}$, there exists two reals L_n and M_n, such that the tails of $\|g_n(x) - g(x)\|$ can be controlled for any $t > 0$ as*

$$\sup_{x \in \operatorname{supp} \rho_{\mathcal{X}}} \mathbb{P}_{\mathcal{D}_n} \left(\|g_n(x) - g(x)\| > t \right) \leq \exp\left(- \frac{L_n t^2}{1 + M_n t} \right). \tag{5.4}$$

Note that to satisfy Assumption 2, it is sufficient, yet *not necessary*, to have a uniform control on $g_n - g^*$, *i*.e., a control on the tail of $\|g_n - g^*\|_{L^\infty}$, since $\sup_x \mathbb{P}\left(A_x > t\right) \leq \mathbb{P}\left(\cup_x \{A_x > t\}\right) = \mathbb{P}\left(\sup_x A_x > t\right)$, with (A_x) a family of random variables indexed by $x \in \mathcal{X}$.

Usually, in bounds like (5.4), M_n is a constant of the problem, while L_n depends on the number of samples, therefore, we would like to give rates depending on L_n. Typically, in Bernstein inequalities (see Theorem 9 in Appendix), $L_n = n\sigma^{-2}$ with σ^2 a variance parameter and $M_n = c\sigma^{-2}$ with c a constant of the problem that does not depend on n.

5.2.2 Refined calibration

While it is sufficient to control the excess risk through an L^1-norm control on g from (5.3), it is not always necessary. In other words, the calibration bound in Lemma 6 is not always tight. Indeed, we do not predict optimally, that is, $\{f_n(x) \neq f^*(x)\}$ only if $g_n(x)$ and $g^*(x)$ do not lead to the same decoding $f_n(x)$ and $f^*(x)$. When $\mathcal{Z}$ is finite, this is characterized by $g_n(x)$ and $g^*(x)$ not falling in the same region R_z of $\mathcal{H}$, where

$$R_z = \left\{ \xi \in \mathcal{H} \middle| z \in \operatorname*{arg\,min}_{z' \in \mathcal{Z}} \langle \psi(z'), \xi \rangle \right\}.$$

To ensure that $g_n(x)$ and $g^*(x)$ fall in the same region, one can ensure that $g_n(x)$ is closer to $g^*(x)$ than $g^*(x)$ is on the frontier of those regions. Those frontiers are defined by points leading to at least two minimizers in (5.2):

$$F = \left\{ \xi \in \mathcal{H} \middle| \left| \operatorname*{arg\,min}_{z \in \mathcal{Z}} \langle \psi(z), \xi \rangle \right| > 1 \right\}.$$

The introduction of F is motivated by the following geometric results.

Lemma 7 (Refined surrogate control)**.** *When $\mathcal{Z}$ is finite, for any $x \in \operatorname{supp} \rho_{\mathcal{X}}$,*

$$\|g_n(x) - g^*(x)\| < d(g^*(x), F) \qquad \Rightarrow \qquad f_n(x) = f^*(x),$$

with d the extension of the norm distance to sets as $d(g^*(x), F) = \inf_{\xi \in F} \|g^*(x) - \xi\|$. This result allows refining the calibration control from Lemma 6 as

$$\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq 2 c_\psi \, \mathbb{E}_X \left[\mathbf{1}_{\|g_n(X) - g^*(X)\| \geq d(g^*(X), F)} \, \|g_n(X) - g^*(X)\| \right]. \tag{5.5}$$

Example 9 (Binary classification)**.** *In binary classification (*cf. *Example 8), $F = \{0\}$, and, for any $x \in \operatorname{supp} \rho_{\mathcal{X}}$, $d(g^*(x), F) = |g^*(x)|$. Lemma 7 is based on the fact that $f^*(x) \neq f_n(x)$ implies that* $\operatorname{sign} g^*(x) \neq \operatorname{sign} g_n(x)$ *which itself implies that $|g^*(x) - g_n(x)| = |g^*(x)| + |g_n(x)| \geq |g^*(x)|$.*

To leverage (5.5), we need to control $d(g^*(x), F)$ below and $\|g_n(x) - g^*(x)\|$ above. While upper bounds on $\|g_n(x) - g^*(x)\|$ are assumed to have been derived through concentration inequalities, lower bounds on $d(g^*(x), F)$ will be assumed as a given parameter of the problem, see (5.6) and (5.7).

Remark 8 (Scope of our work)**.** *While we derived the refined calibration inequality* (5.5) *for the surrogate conditional mean g^* and the associated pointwise metric $\|\cdot\|_{\mathcal{H}}$, similar inequality could be obtained for other types of surrogate methods. This suggests that our work could be extended to any smooth surrogate such as the ones considered by Nowak-Vila et al. (2020), as well as Fenchel-Young losses (Blondel et al., 2020).*

5.2.3 Geometric understanding

In this subsection, we detail how to understand geometrically Lemma 7. While the introduction of φ and ψ could seem arbitrary, it can be thought in a more intrinsic manner by considering the embedding $\varphi(y) = \delta_y$ belonging to the Banach space $\mathcal{H}$ of signed measured, $g^*(x) = \rho|_x$, with the bracket operator, for $\mu \in \mathcal{H}$ and $z \in \mathcal{Z}$, $\langle z, \mu \rangle = \int_{\mathcal{Y}} \ell(z, y)\mu(\mathrm{d}y)$, and the distance between signed measures being $d(\mu_1, \mu_2) = \sup_{z \in \mathcal{Z}} \langle z, \mu_1 - \mu_2 \rangle$. Note that Lemma 7 is a pointwise result, holding for any $x \in \mathcal{X}$, that is integrated over $\mathcal{X}$ afterwards. Therefore, it is enough to consider $\mathcal{X} = \{x\}$ and remove the dependency in $\mathcal{X}$ to understand it. The simplex $\Delta_{\mathcal{Y}}$ naturally splits into the decision region R_z for $z \in \mathcal{Z}$ as illustrated on Figure 5.1. The main idea of Lemma 7 is that one does not have to precisely estimate $g^*(x) = \rho|_x$ but only has to make sure that $g_n(x)$ falls in the same region on Figure 5.1.

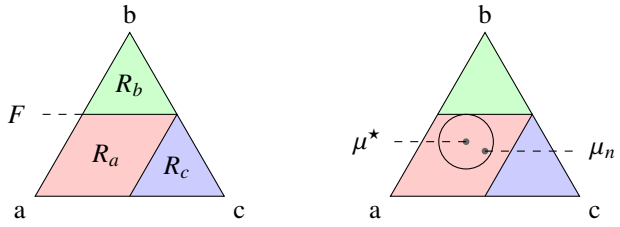

Figure 5.1: Illustration of Lemma 7. Simplex $\Delta_{\mathcal{Y}}$, for $\mathcal{Y} = \mathcal{Z} = \{a, b, c\}$ and ℓ a symmetric loss defined as $\ell(a, b) = \ell(a, c) = 1$ and $\ell(b, c) = 2$, while $\ell(z, z) = 0$. This leads to the decision regions R_z represented in colors. Given $x \in \mathcal{X}$, if $g^*(x)$ corresponds to a distribution $\mu^* := \rho|_x$ falling in R_a, and if $g_n(x)$ represented by $\hat{\mu}$ falls closer to μ^* than the distance between μ^* and the decision frontier F (represented by a circle on the right figure), then $\hat{\mu}$ is also in R_a, and therefore $f^*(x) = f_n(x) = a$.

5.3 Rate acceleration under margin condition

In this section, we introduce a condition that g^* is not too often close to the decision frontier F. It generalizes the so-called "Tsybakov margin condition" known for classification. Under this condition, we prove rates that generalize the results of Audibert and Tsybakov (2007) from binary classification to generic structured prediction problems, which opens the way to "superfast" rates in structured prediction.

5.3.1 No density separation

To get fast convergence rates, one has to make assumptions on the problem. A classical assumption is that g^* is smooth enough in order to get concentration bounds similar to Assumption 2 when considering a specific class of estimates g_n. In our decoding setting (Lemma 6), learning is made easy when it is easy to estimate in which region R_z the optimal g^* will fall in. This is in particular the case, when there is a margin $t_0 > 0$, for which, for no point $x \in \operatorname{supp} \rho_{\mathcal{X}}$, $g^*(x)$ falls at distance t_0 of the decision frontier F, motivating the following definition.

Assumption 3 (No-density separation)**.** *A surrogate solution g^* will be said to satisfy the* no-density separation*, if there exists a $t_0 > 0$, such that*

$$\mathbb{P}_X(d(g^*(X), F) < t_0) = 0. \tag{5.6}$$

This condition is alternatively called the hard margin condition*, or sometimes "Massart's noise condition" for binary classification (Massart and Nédélec, 2006).*

Remark 9 (Separation in $\mathcal{Y}$ and separation in $\mathcal{X}$)**.** *It is important to realize that* (5.6) *is a condition of separation in $\Delta_{\mathcal{Y}}$ that should hold for all $x \in \mathcal{X}$, but it does not state any separation between classes in $\mathcal{X}$ for $f^* : \mathcal{X} \to \mathcal{Z}$. To visualize it, consider the classification problem where $\mathcal{X} = [-1, 1]$, $\mathcal{Y} = \mathcal{Z} = \{-1, 1\}$ and $\ell(z, y) = \mathbf{1}_{z \neq y}$.*

- *A situation where $\rho_{\mathcal{X}}$ is uniform on $\mathcal{X}$ and $\mathbb{E}[Y|X = x] = 2 \cdot \mathbf{1}_{x \in p\mathbb{N} + \{a \,|\, |a| < p/4\}} - 1$, for $p = 1/50$, satisfies separation in $\Delta_{\mathcal{Y}}$* (5.6)*, but classes are not separated in $\mathcal{X}$.*

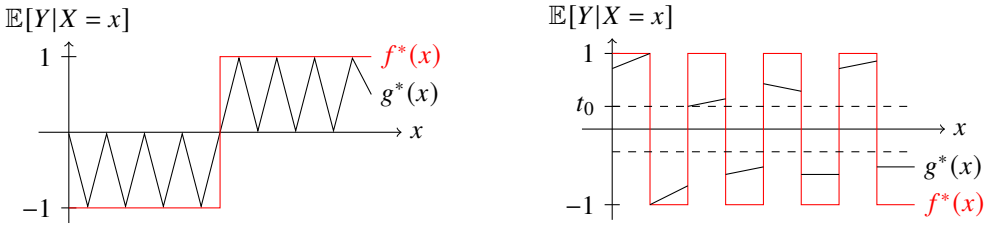

Figure 5.2: Illustration of Remark 9. We represent two instances of binary classification (see Examples 8 and 9). On the left example, when $\rho_{\mathcal{X}}$ is such that there is no mass where the sign of g^* changes, classes are separated in $\mathcal{X}$, yet the no-density separation is not verified. On the right, classes are not separated in $\mathcal{X}$, but the problem satisfies the no-density separation as there is no x such that $d(g^*(x), F) = |g^*(x)| < t_0$. Note that when $\rho_{\mathcal{X}}$ is uniform, the left problem satisfies a milder separation condition, introduced thereafter and called the 1-low-density separation.

- *A situation where ρ is uniform on $[-1,-.5] \cup [.5,1]$, with $\mathbb{E}[Y|X=x] = \text{sign}(x)(1-|x|)^p$, for $p > 0$, satisfies a separation of classes in $\mathcal{X}$ but does not satisfy* (5.6).

Note that continuity of g^* and the no-density separation in (5.6) *imply separation of classes in $\mathcal{X}$. Note also that to get concentration inequality such as* (5.4)*, one usually supposes that g^* is smooth. We refer the curious reader to Section 2.4 in Steinwart and Scovel (2007) for separation in $\mathcal{X}$.*

The introduction of Assumption 3 is motivated by the following result.

Theorem 1 (Rates under no-density separation)**.** *When ℓ is bounded by ℓ_∞* (i.e.*, $\ell(z,y) \leq \ell_\infty$ for any $(z,y) \in \mathcal{Z} \times \mathcal{Y}$) and satisfies Assumption 1, and $\mathcal{Z}$ is finite, under the no-density separation Assumption 3, and the concentration Assumption 2, the excess risk is controlled*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty \exp\left(-\frac{L_n t_0^2}{1 + M_n t_0}\right).$$

Proof. Because we make a mistake only when $d(g^*(x), F) \geq \|g_n(x) - g^*(x)\|$, we make no mistake when $\|g_n(x) - g^*(x)\| < t_0$; otherwise we can consider the worse error we are going to pay, that is ℓ_∞, leading to

$$\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty \, \mathbb{P}_X(\|g_n(x) - g^*(x)\| > t_0).$$

Taking the expectation with respect to $\mathcal{D}_n$ and using the fact that $\mathbb{E}_A \, \mathbb{P}_B(Z) = \mathbb{E}_A \, \mathbb{E}_B[\mathbf{1}_Z] = \mathbb{E}_B \, \mathbb{E}_A[\mathbf{1}_Z] = \mathbb{E}_B \, \mathbb{P}_A(Z)$, and plug-in the concentration inequality (5.4), we get the result. □

Example 10 (Image classification)**.** *In image classification, one can arguably assume that the class of an image is a deterministic function of this image. With the 0-1 loss, it implies that the image classification problem verifies the no-density separation. The same holds for any discrete problem where the label is a deterministic function of the input. Based on Theorem 1 and* (5.4) *in which M is generally a constant when L is proportional to the number of data, it is reasonable to ask for exponential convergences rates on such problems.*

5.3.2 Low density separation

While we presented the no-density separation first for readability, it is a strong assumption. Recall our example, Remark 9, with $\mathbb{E}[Y|X=x] = \text{sign}(x)(1-|x|)^p$, only around $x = 1$ and $x = -1$ is $d(g^*(x), F)$ not bounded away from zero. While the neighborhood of those points should be studied carefully, the error on all other points $x \in [-1+t, 1-t]$ can be controlled with exponential rates. The low-density separation, also known as the Tsybakov margin condition in binary classification, will allow a refined control to get fast rates in such a setting.

Assumption 4 (Low-density separation)**.** *A surrogate solution g^* is said to satisfy the* low-density separation*, if there exists $c_\alpha > 0$, and $\alpha > 0$, such that for any $t > 0$*

$$\mathbb{P}_X(d(g^*(X), F) < t) \leq c_\alpha t^\alpha. \tag{5.7}$$

This condition is alternatively called the margin condition*.*

The low-density separation spans all situations from the hard margin condition, that can be seen as $\alpha = +\infty$, to situations without any margin assumption corresponding to $\alpha = 0$. The coefficient α is an intrinsic measure of the easiness of finding f^* in the problem (ℓ, ρ). For example, the setting described in the last paragraph corresponds to the case $\alpha = 1/p$. We discuss the equivalence of Assumption 4 to definitions appearing in the literature in Remark 10.

Theorem 2 (Optimal rates under low density separation)**.** *Under refined calibration in* (5.5)*, concentration (Assumption 2), and low-density separation (Assumption 4), the risk is controlled as*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq 2c_\psi c_\alpha c \left(M_n^{\alpha+1} L_n^{-(\alpha+1)} + L_n^{-\frac{\alpha+1}{2}} \right),$$

for c a constant that only depends on α, that can be expressed through the Gamma function evaluated in quantity depending on α, meaning that when α is big, c behaves like α^α. Note that it is not possible to derive a better bound only given (5.4)*,* (5.5) *and* (5.7)*.*

Sketch for Theorem 2, details in Appendix 5.A.5. Based on the refined calibration inequality in (5.5), and using that $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\, \mathrm{d}t$, it is possible to show that the expectation of the excess risk behave like

$$\int_0^\infty \mathbb{P}_X(d(g^*(x), F) < t) \sup_x \mathbb{P}_{\mathcal{D}_n}(\|g_n(x) - g^*(x)\| > t)\, \mathrm{d}t.$$

Based on Assumptions 2 and 4, the integrand behaves like $t^\alpha \exp(-L_n t^2/(1 + M_n t))$. A change of variable and the study of the Gamma function leads to the result. We provide all the details in Appendix 5.A.5. Note that while we stated Theorem 2 under an exponential inequality of Bernstein type (Assumption 2), similar theorems can be derived for any type of exponential concentration inequality, as stated in Lemma 18 in Appendix 5.A.6. □

Theorem 2 is to put in perspective with the work of Nowak-Vila et al. (2019) which considers the same setup as ours, yet only succeeds to derive acceleration by a power $2(\alpha + 1)/(\alpha + 2)$, while we got an acceleration by a power $(\alpha + 1)/2$ as already mentioned in the related work section. This gain will appear more clearly in Theorem 6.

Remark 10 (Independence to the decomposition of ℓ)**.** *While we have stated results based on the quantity $d(g^*(x), F)$, generalization of the Tsybakov margin condition has also been expressed through the quantity $\inf_{z \neq z^*} \mathbb{E}_{Y \sim \rho|_x} \ell(z, Y) - \mathbb{E}_{Y \sim \rho|_x} \ell(z^*, Y)$ instead of $d(g^*(x), F)$ (Nowak-Vila et al., 2019). We show in Appendix 5.A.3 that the two definitions of the margin condition are equivalent.*

Remark 11 (Scope of our work)**.** *Our work relies on pointwise exponential concentration inequalities (Assumption 2) which are specially designed to work well with the Tsybakov margin condition. It is natural for localized averaging methods such as nearest neighbors, or for surrogate methods leading to L^∞ concentration. For surrogate methods leading to concentration of other quantities, it is possible to use similar tricks under different "margin" conditions (*e.g. *Steinwart and Scovel (2007) for a margin condition designed for the Hinge loss).*

Note that L^2 concentration on g_n toward g^* (such as the one derived by Marteau-Ferey et al. (2019) for logistic regression) could also be turned into fast convergence of f_n toward f^*, since, in essence, for points $x \in \mathcal{X}$ where $\rho(\mathrm{d}x)$ is high, the quantity $g^*(x) - g_n(x)$ will have a non-negligible contribution to $\|g^* - g_n\|_{L^2}$ – allowing to cast concentration in L^2 to concentration pointwise in x – and for points $x \in \mathcal{X}$ where $\rho(\mathrm{d}x)$ is negligible, it is acceptable to pay the worst error, since it will have a small contribution to the excess of risk.

Finally, note that it is also possible to let the right hand-side term in (5.4) *depends on x, and to modify Theorem 1 with $L = \mathbb{E}[L(X)]$.*

5.3.3 The importance of constants

In this subsection, we discuss the importance of constants when providing learning rates. Assumption 3 corresponds to asking for $g^*(x)$ never to enter a neighborhood of F defined through t_0. Similarly, when $\mathcal{X}$ is parametrized such that $\rho_{\mathcal{X}}$ is uniform, the parameter α in Assumption 4 corresponds to the speed at which $g^*(x)$ "get through" the decision frontier F. In order to have a higher α and optimize the dependency in n in the bound of Theorem 2, it is natural to think of infinitesimal perturbations of g^* to make it cross the

boundary orthogonally (or even jump over it and satisfy the no-density separation). To give a precise example, in binary classification, let us artificially add smoothness to the function $g^*(x) = x^q$ when approaching zero. Consider $g^* : [0,1] \to [-1,1], x \to c^{q-p}x^p \mathbf{1}_{x<c} + x^q \mathbf{1}_{x\geq c}$, and x uniform, and $p < q$. In this setting, α can be taken anywhere in $[0, p^{-1})$. Naively, we could ask for the biggest possible α in order to have the best dependency in n in the learning rates given by Theorem 2. While this approach will higher α, it will also higher c_α, compensating the gain one could expect from such a strategy. Indeed, for $\alpha \in [0, p^{-1}]$, at best, we can take $c_\alpha = \mathbf{1}_{\alpha<q^{-1}} + c^{1-q\alpha}\mathbf{1}_{\alpha\geq q^{-1}}$. This shows the importance of optimizing both α and $c_\alpha c$ to minimize the lower bound appearing in Theorem 2 when given a fixed number of sample n.

In a word, while we only give results that are optimized in n, when n is fixed, better bounds could be given by optimizing parameters and constants simultaneously. For example, when $\mathcal{X} = \mathbb{R}^d$ and g^* belongs to the Sobolev space H^m for all $m \in [0, m_*]$, and satisfies Assumption 4 for all $\alpha \in [0, \alpha_*]$, we expect the best bound, that could be derived from our proof technique, to be of form

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \min_{m\leq m_*, \alpha<\alpha_*} \alpha^\alpha c_\alpha c_\psi \|g^*\|_{H^m} n^{-\frac{m(\alpha+1)}{2d}}.$$

Yet, for simplicity, we will express those bounds as $bn^{-\frac{m_*(\alpha_*+1)}{2d}}$, for b a big constant.

5.4 Application to nearest neighbors

In this section, we consider the Bayes approximate risk estimator proposed by Stone (1977), with weights given by nearest neighbors (Cover and Hart, 1967). We prove, under regularity assumptions, concentration inequalities similar to (5.4), which allow us to derive exponential and polynomial rates. Given samples $(X_i, Y_i) \sim \rho^{\otimes n}$, $k \in \mathbb{N}$ and a metric d on $\mathcal{X}$, the estimator is

$$g_n(x) = \sum_{i=1}^n \alpha_i(x)\varphi(Y_i), \quad \text{with } \alpha_i(x) = \begin{cases} k^{-1} & \text{if } \sum_{j=1}^n \mathbf{1}_{d(x,X_j)\leq d(x,X_i)} < k \\ 0 & \text{if } \sum_{j=1}^n \mathbf{1}_{d(x,X_j)<d(x,X_i)} \geq k \\ (pk)^{-1} & \text{else, with } p = \sum_{j=1}^n \mathbf{1}_{d(x,X_j)=d(x,X_i)}. \end{cases} \tag{5.8}$$

To study the convergence of g_n, we introduce the noise free estimator $g_n^* = \sum_{i=1}^n \alpha_i(x) g^*(X_i)$. This allows separating the error due to the randomness of the labels $Y_i \sim \rho|_{X_i}$, and the error due to the difference between $g^*(x)$ and the averaging of g^* on the neighbors of x defining g_n. To control the fist error, we need a bounded moment condition on $\varphi(Y)$. We reuse an assumption from Bernstein (1937), that is classic in machine learning (*e.g.*, Caponnetto and De Vito, 2006; Lin et al., 2020).

Assumption 5 (Sub-exponential moment of $\rho|_x$)**.** *Suppose that there exists $\sigma^2, M > 0$ such that for any $x \in \operatorname{supp} \rho_{\mathcal{X}}$, for any $m \geq 2$, we have*

$$\mathbb{E}_{Y\sim\rho|_x}\left[\|\varphi(Y) - g^*(x)\|^m\right] \leq \frac{1}{2} m! \sigma^2 M^{m-2}.$$

Example 11 (Moment bound on $\varphi(Y)$)**.** *Assumption 5 is a classical assumption that is notably satisfied when $\varphi(Y)$ is bounded by M, with σ^2 its variance, or when $(\varphi(Y) \mid X)$ is Gaussian with covariance bounded by a constant independent of X (see a proof of this standard result by Fischer and Steinwart, 2020).*

To control the second error, we notice, for $x \in \operatorname{supp} \rho_{\mathcal{X}}$, that the quantity $\left\|g^*(x) - g_n^*(x)\right\|$ behaves like $\sup_{x'\in\mathcal{B}(x,r)} \|g^*(x) - g^*(x')\|$, with r such that $\rho_{\mathcal{X}}(\mathcal{B}(x,r)) \approx k/n$, such an r modeling the distance between x and its k-th neighbor. This motivates the following assumption.

Assumption 6 (Modified Lipschitz condition (Chaudhuri and Dasgupta, 2014))**.** *g^* is said to verify the β-Modified Lipschitz condition if there exists $c_\beta > 0$ such that for any $x, x' \in \operatorname{supp} \rho_{\mathcal{X}}$*

$$\|g^*(x) - g^*(x')\| \leq c_\beta \rho_{\mathcal{X}}(\mathcal{B}(x, d(x,x')))^\beta,$$

where d is the distance on $\mathcal{X}$, and $\mathcal{B}(x,t) \subset \mathcal{X}$ the ball of center x and radius t.

Typically, the β that appears in Assumption 6 is linked with the dimension of a subset of $\mathcal{X}$ containing most of the mass of $\rho_{\mathcal{X}}$ (see below). This will slow the rates according to this dimension parameter, a property referred to as the curse of dimensionality.

Example 12 (Classical assumptions)**.** *When $\mathcal{X} = \mathbb{R}^d$, if g is β'-Hölder continuous, and $\rho_{\mathcal{X}}$ is regular in the sense that, there exists a constant c and $t^* > 0$ such that for $x \in \operatorname{supp} \rho_{\mathcal{X}}$ and any $t \in [0, t^*]$, $\rho_{\mathcal{X}}(\mathcal{B}(x,t)) \geq c\lambda(\mathcal{B}(x,t))$, with λ the Lebesgue measure on $\mathcal{X}$, then g satisfies the modified Lipschitz condition with $\beta = \beta'/d$. The condition on $\rho_{\mathcal{X}}$ is usually split in a condition of minimal mass of $\rho_{\mathcal{X}}$, and a condition of regular boundaries of* $\operatorname{supp} \rho_{\mathcal{X}}$ *(*e.g.*, Audibert and Tsybakov, 2007). We provide more details in Appendix 5.B.1.*

We now state convergence results, respectively proven in Appendices 5.B.2, 5.B.3 and 5.B.4, in which the constant values b_1 to b_6 appear explicitly. Note that results provided by Lemma 12 are already known in the literature (Györfi et al., 2002), while Theorems 3 and 4 were only known in binary classification, but we generalize them to any discrete structured prediction problem. It should be noted that rates in Theorem 4 match the minimax rates derived by Audibert and Tsybakov (2007) in the case of binary classification.

Lemma 12 (Nearest neighbors concentration)**.** *Under Assumptions 5 and 6, there exist constants $b_1, b_2, b_3 > 0$, such that for any $x \in \operatorname{supp} \rho_{\mathcal{X}}$ and any $t > 0$,*

$$\mathbb{P}_{\mathcal{D}_n}\left(\left\|g_n(x) - g_n^*(x)\right\| > t\right) \leq 2\exp\left(-\frac{b_1 k t^2}{1 + b_2 t}\right).$$

And for $t > (k/2n)^{\beta}$, when $\rho_{\mathcal{X}}$ is continuous[1]

$$\mathbb{P}_{\mathcal{D}_n}\left(\left\|g_n^*(x) - g^*(x)\right\| > t\right) \leq \exp\left(-b_3 n t^{\frac{1}{\beta}}\right).$$

Theorem 3 (Nearest neighbors fast rates under no-density assumption)**.** *When ℓ is bounded by ℓ_∞, satisfies Assumption 1, and $\mathcal{Z}$ is finite, under the no-density separation, Assumption 3, and Assumptions 5 and 6, there exist two constants $b_4, b_5 > 0$ that do not depend on n or k such that for any $n \in \mathbb{N}^*$ and any k such that $(k/2n)^{\beta} < t_0$, we have*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq 2\ell_\infty \exp(-b_4 k) + \ell_\infty \exp(-b_5 n). \tag{5.9}$$

Theorem 4 (Nearest neighbors fast rates under low-density assumption)**.** *When ℓ satisfies Assumption 1, and $\mathcal{Z}$ is finite, under the low-density separation, Assumption 4, and Assumptions 5 and 6, considering the scheme $k_n = \left\lfloor k_0 n^{\frac{2\beta}{2\beta+1}} \right\rfloor$, for any $k_0 > 0$, there exists a constant $b_6 > 0$ that does not depend on n such that for any $n \in \mathbb{N}^*$,*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq b_6 n^{-\frac{\beta(\alpha+1)}{2\beta+1}}. \tag{5.10}$$

Remark 13 (Scope of our work)**.** *The same type of argument works for other local averaging methods, such as Nadaraya-Watson (Nadaraya, 1964; Watson, 1962), local polynomials (Cleveland, 1979; Audibert and Tsybakov, 2007) or decision trees (Breiman et al., 1984).*

5.5 Application to reproducing kernel ridge regression

In this section, we consider the kernel ridge regression estimate g_n of g^* first proposed by Ciliberto et al. (2016), and we prove, under regularity assumptions, uniform concentration inequalities similar to (5.4), which allow us to derive superfast rates at the end of the section. Given a symmetric, positive semi-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, the kernel ridge regression estimation g_n of g^* is defined similarly to (5.8) yet with weights $\alpha(x) \in \mathbb{R}^n$ defined as

$$\alpha(x) = (\hat{K} + \lambda)^{-1} \hat{K}_x, \quad \hat{K} = \left(\frac{1}{n} k(X_i, X_j)\right)_{i,j \leq n} \in \mathbb{R}^{n \times n}, \ \hat{K}_x = \left(\frac{1}{n} k(x, X_i)\right)_{i \leq n} \in \mathbb{R}^n.$$

To state regularity assumptions, we introduce a minimal setup linked to the reproducing kernel k. To keep the exposition clear, we relegate technicalities in Appendix 5.C. We define the operator K operating on functions $f \in L^2(\mathcal{X}, \mathcal{H}, \rho_{\mathcal{X}})$ and $K_{\mathcal{X}}$ operating on $f \in L^2(\mathcal{X}, \mathbb{R}, \rho_{\mathcal{X}})$, both defined as

$$(Kf)(x') = \int_{\mathcal{X}} k(x', x) f(x) \, \mathrm{d}\rho_{\mathcal{X}}(x).$$

[1] Note that this topological assumption ease derivations but is not fundamental for such non-asymptotic results.

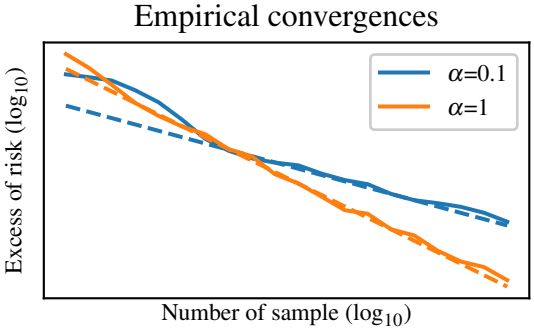

Figure 5.3: Empirical convergence rates. We consider binary classification, with $\mathcal{X} = [-1, 1]$, $g^*(x) = \text{sign}(x) * |x|^{\frac{1}{\alpha}}$, for $\alpha \in \{.1, 1\}$ and $\rho_{\mathcal{X}}$ uniform. We plot in solid the logarithm of the excess risk averaged over 100 trials against the logarithm of the number of samples for $n \in [10, 10^6]$, and plot in dashed the expected slope of those curves due to Theorem 4 (*i.e.*, we fit the constant C in the rate $Cn^{-\gamma}$ with γ obtained from the bound in (5.10)).

Inheriting from the symmetry and positive semi-definiteness of k, K is self-adjoint with the spectrum in $\mathbb{R}_+$. To study the convergence of g_n to g^*, it is useful to introduce the approximate orthogonal projection on $\text{im}\, K^{\frac{1}{2}}$, defined for $\lambda > 0$ as

$$g_\lambda = K(K + \lambda)^{-1} g^*.$$

We introduce three assumptions linked with the regularity of the problem, referred to as the capacity condition, interpolation inequality and source condition. Those are classical assumptions to prove uniform rates of the kernel ridge regression estimates. They could be found, in particular, by Fischer and Steinwart (2020) under the respective names of (EVD), (EMB) and (SRC), but also by Pillaud-Vivien et al. (2018a); Lin et al. (2020). Our assumptions differ in that they are expressed for vector-valued functions, which usually generate compactness issues (Caponnetto and De Vito, 2006). However, when $\mathcal{Z}$ is finite, $\mathcal{H}$ is finite dimensional, and K can be shown to be a compact operator, which allows considering fractional power without definition issues.

Assumption 7 (Capacity condition)**.** *Suppose* $\text{Tr}\left(K_{\mathcal{X}}^{\sigma}\right) < +\infty$ *for* $\sigma \in [0, 1]$.

Assumption 8 (Interpolation inequality)**.** *Assume the existence of* $p \in [0, \frac{1}{2}]$, $c_p > 0$ *such that*

$$\forall\, g \in (\ker K)^{\perp}, \qquad \|K^p g\|_{L^\infty} \leq c_p \|g\|_{L^2}.$$

Assumption 9 (Source condition)**.** *Suppose* $g^* \in \text{im}\, K^q$ *for* $q \in (p, 1]$.

When $q = 1/2$, the source condition is often expressed as g^* belonging to the reproducing kernel Hilbert space associated to the kernel k. Note that when k is bounded, Assumptions 7 and 8 hold with $\sigma = 1$ and $p = 1/2$. In those assumptions, for p and σ the smaller, and for q the bigger, the faster the convergence rates will be.

Example 13 (Classical assumptions)**.** *For Assumption 8 to hold, minimal mass and regular support of* ρ*, similarly to Example 12, are often assumed, as well as regularity of functions in* $\text{im}\, K^p$*, in coherence with Remark 11. For Assumption 9 to hold, it is classical to assume regularity of* g^**, matching the regularity of function spaces derived from the kernel* k*. The value of* σ *in Assumption 7 often comes as a bonus of regularity assumptions on* ρ *and specificity of the RKHS implied by* k*. See Example 2 by Pillaud-Vivien et al. (2018a) and Section 4 by Fischer and Steinwart (2020) as well as references therein for concrete examples.*

We now state convergence results respectively proven in Appendices 5.C.5 and 5.C.6, 5.C.7, and 5.C.8. Lemma 14 is a generalization to vector-valued functions of kernel ridge regression uniform convergence rates known for real-valued function (see Fischer and Steinwart, 2020). Note that a similar result to Theorem 5 was provided for binary classification by Koltchinskii and Beznosova (2005), but we generalize exponential rates with kernel ridge regression to any discrete structured prediction problem. Theorem 6 is new, even in the context of binary classification. It states that, while, up to now, only rates in $n^{-1/4}$ were known for f_n (Ciliberto et al., 2020), one can indeed hope for arbitrarily fast rates, depending on the hardness of the problem, read in the value of $\alpha \in [0, \infty)$.

Lemma 14 (Reproducing kernel concentration)**.** *Under Assumptions 7, 8 and 9, for any $\lambda > 0$,*

$$\|g_\lambda - g^*\|_{L^\infty} \leq b_1 \lambda^{q-p}.$$

With $b_1 = c_p \|K^{-q} g^*\|_{L^2}$. Moreover, when the kernel k is bounded and under Assumption 5, there exists three constants $b_2, b_3, b_4, b_5 > 0$ that does not depend nor on λ nor on n such that

$$\mathbb{P}\left(\|g_n - g_\lambda\|_\infty > t\right) \leq b_2 \lambda^{-\sigma} \exp\left(-b_3 n \lambda^{2p}\right) + 4 \exp\left(-\frac{n\lambda^{2p+\sigma} t^2}{b_4 + b_5 t}\right).$$

As long as $b_3 n \geq \lambda^{-p}$, and $\lambda \leq \min\left(\|K\|_{\text{op}}, 1\right)$.

Theorem 5 (Kernel ridge regression fast rates under no-density assumption)**.** *When the loss ℓ is bounded, satisfies Assumption 1 and $\mathcal{Z}$ is finite, under the t_0-no-density separation condition, and Assumptions 5, when k is bounded, if $\lambda_n = \lambda$, for any $\lambda > 0$ such that $\|g^* - g_\lambda\|_{L^\infty} < t_0$, then there exist two constants $b_6, b_7 > 0$ such that, for any $n \in \mathbb{N}^*$,*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq b_6 \exp(-b_7 n), \tag{5.11}$$

with f_n given by the kernel ridge regression surrogate estimate.

Theorem 6 (Kernel ridge regression fast rates under low-density assumption)**.** *When ℓ satisfies Assumption 1, is bounded and $\mathcal{Z}$ is finite, under the α-low-density separation condition, and Assumptions 5, 7, 8 and 9, if $\lambda_n = \lambda_0 n^{-\frac{1}{2q+\sigma}}$, for any $\lambda_0 > 0$, there exists $b_8 > 0$, such that for any $n \in \mathbb{N}^*$,*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq b_8 n^{-\frac{(q-p)(1+\alpha)}{2q+\sigma}}. \tag{5.12}$$

5.6 Conclusion

In this paper, we have shown how, for discrete problems, to leverage exponential concentration inequalities derived on continuous surrogate problems, in order to derive faster rates than rates directly obtained through calibration inequalities. Those rates are arbitrarily fast, depending on a parameter characterizing the hardness of the discrete problem. We have shown how this method directly applies to local averaging methods and to kernel ridge regression, which allows deriving "superfast" rates for any discrete structured prediction problem.

This opens the way to several follow-up, such as

- Application follow-up, consisting of tackling concrete problem instances, such as predicting properties of DNA-sequence (Jaakkola et al., 2000), *e.g.*, gene mutations responsible for diseases, with well-designed kernels on DNA in order to higher the exponent appearing in Theorem 6.
- Computational follow-up, pushing our analysis further to understand how to design better algorithms on discrete problems. For example, by adding a regularization pushing g_n away from the decision frontier F, and adding a term in $\mathbf{1}_{\|g_n(x) - g^*(x)\| > d(g_n(x), F)}$ in (5.5) for the analysis.
- Theoretical follow-up, to widen our analysis to other types of smooth surrogates, and to parametric methods, such as deep learning models, assuming that functions are parametrized by a parameter θ, that some analysis gives concentration on $\theta_n - \theta^*$ similar to (5.4) and that calibration inequalities relate the error on θ with the error between $f_n = f_{\theta_n}$ and $f^* = f_{\theta^*}$.

Appendix

5.A Fast rates

In the following, we consider $\mathcal{X}$ and $\mathcal{Y}$ to be Polish spaces, *i.e.*, separable completely metrizable topological spaces, in order to define the distribution ρ. We also consider $\mathcal{Z}$ endowed with a topology that makes it compact, and that makes $z \to \mathbb{E}_{Y\sim\mu}\,\ell(z,Y)$ continuous for any $\mu \in \Delta_{\mathcal{Y}}$, in order to have minimizer well-defined. For a Polish space $\mathcal{A}$, we denote by $\Delta_{\mathcal{A}}$ the simplex formed by the set of Borel probability measures on this space. For $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$, we denote by $\rho|_x$ the conditional distribution of Y given x, and by $\rho_{\mathcal{X}}$ the marginal distribution over $\mathcal{X}$. We suppose $\mathcal{H}$ separable Hilbert and that the mapping φ is measurable in order to define the pushforward measure $\varphi_*\rho|_x$. We assume that, for $\rho_{\mathcal{X}}$-almost every x, $(\varphi(Y)|X=x)$ has a second moment, in order to consider the conditional mean $g^*(x)$ as the solution of the well-defined problem consisting of minimizing $\|\xi - \varphi(Y)\|^2$ for $\xi \in \mathcal{H}$. We consider ψ to be continuous, in order to have the decoding problem well posed.

5.A.1 Proof of Lemma 6

With the notation of Lemma 6, for $x \in \operatorname{supp}\rho_{\mathcal{X}}$

$$\begin{aligned}
\mathbb{E}_{Y\sim\rho|_x}\left[\ell(f_n(x),Y) - \ell(f^*(x),Y)\right] &= \langle\psi(f_n(x)) - \psi(f^*(x)), g^*(x)\rangle_{\mathcal{H}} \\
&= \langle\psi(f_n(x)), g_n(x)\rangle + \langle\psi(f_n(x)), g^*(x) - g_n(x)\rangle - \langle\psi(f^*(x)), g^*(x)\rangle \\
&\leq \langle\psi(f^*(x)), g_n(x)\rangle + \langle\psi(f_n(x)), g^*(x) - g_n(x)\rangle - \langle\psi(f^*(x)), g^*(x)\rangle \\
&= \langle\psi(f_n(x)) - \psi(f^*(x)), g^*(x) - g_n(x)\rangle \\
&\leq \|\psi(f_n(x)) - \psi(f^*(x))\|_{\mathcal{H}}\,\|g^*(x) - g_n(x)\|_{\mathcal{H}} \\
&\leq 2c_\psi\,\|g^*(x) - g_n(x)\|_{\mathcal{H}}\,,
\end{aligned}$$

where the inequality $\langle\psi(f_n(x)), g_n(x)\rangle \leq \langle\psi(f^*(x)), g_n(x)\rangle$ is due to the fact that $f_n(x)$ minimizes the functional $z \to \langle\psi(z), g_n(x)\rangle$. Integrating over $\mathcal{X}$ leads to the results in Lemma 6.

5.A.2 Proof of Lemma 7

The first part of the lemma is a geometrical result stating that to go from two elements ξ_1 and ξ_2 in $\Delta_{\varphi(\mathcal{Y})}$, leading to two different decoding, one has to pass by a point $\xi_{1/2} \in F$, where there is at least two possible decodings. Let us make it clearer. Consider $x \in \operatorname{supp}\rho_{\mathcal{X}}$ and suppose that $f_n(x) \neq f^*(x)$, define the path

$$\begin{array}{rccc}
\zeta: & [0,1] & \to & \Delta_{\varphi(\mathcal{Y})} \\
& \lambda & \to & \lambda g_n(x) + (1-\lambda)g^*(x).
\end{array}$$

Consider $d : \Delta_{\varphi(\mathcal{Y})} \to \mathcal{Z}$ the decoding function used to retrieve f^* and f_n, from g^* and g_n, satisfying $d(\xi) \in \arg\min_{z\in\mathcal{Z}}\langle\psi(z),\xi\rangle$. Consider the path $d\circ\zeta : [0,1] \to \mathcal{Z}$, it goes from $\zeta(0) = f^*(x)$ to $\zeta(1) = f_n(x)$. Consider λ_∞ the supremum of $(d\circ\zeta)^{-1}(f^*(x))$. We will show that $\zeta(\lambda_\infty) \in F$, this will lead to

$$\|g_n(x) - g^*(x)\| = \|g_n(x) - \zeta(\lambda_\infty)\| + \|\zeta(\lambda_\infty) - g^*(x)\| \geq \|\zeta(\lambda_\infty) - g^*(x)\| \geq d(g^*(x), F),$$

and to Lemma 7 by contraposition.

To show that $\zeta(\lambda_\infty) \in F$, we will show that $f^*(x) \in \arg\min_z \langle \psi(z), \zeta(\lambda_\infty)\rangle \not\subset \{f^*(x)\}$. By definition of the supremum, there exists a sequence $(\lambda_p)_{p\in\mathbb{N}}$ converging to λ_∞ such that

$$f^*(x) \in \arg\min_z \left\langle \psi(z), \lambda_p g_n(x) + (1-\lambda_p) g^*(x) \right\rangle,$$

meaning that for all $z \neq f^*(x)$

$$\left\langle \psi(f^*(x)), \lambda_p g_n(x) + (1-\lambda_p) g^*(x) \right\rangle \leq \left\langle \psi(z), \lambda_p g_n(x) + (1-\lambda_p) g^*(x) \right\rangle.$$

By continuity of the scalar product, it holds for $p = \infty$, which means $f^*(x) \in \arg\min_z \langle \psi(z), \zeta(\lambda_\infty)\rangle$. Now, suppose that $\arg\min_z \langle \psi(z), \zeta(\lambda_\infty)\rangle = \{f^*(x)\}$, this means that for all $z \neq f^*(x)$,

$$\langle \psi(f^*(x)), \lambda_\infty g_n(x) + (1-\lambda_\infty) g^*(x)\rangle < \langle \psi(z), \lambda_\infty g_n(x) + (1-\lambda_\infty) g^*(x)\rangle.$$

By continuity of this function according to λ, this means that this still holds for $\lambda_\infty + \varepsilon_z$ for $\varepsilon_z > 0$. Taking $\varepsilon = \inf_{z\in\mathcal{Z}} \varepsilon_z$, it means that $\lambda_\infty + \varepsilon \in (d \circ \zeta)^{-1}(f^*(x))$. When $\mathcal{Z}$ is finite, $\varepsilon > 0$, which contradicts the definition of λ_∞. Therefore, $\zeta(\lambda_\infty) \in F$.

The second part of Lemma 7 follows from derivations in Appendix 5.A.1.

Remark 15 (Extension to discrete cases)**.** *Note that the same argument can be generalized to discrete problems – which could be defined as $\mathcal{Z}$ endowed with a topology that makes $z \to \mathbb{E}_{Y\sim\mu}[\ell(z,Y)]$ continuous with respect to z, and $\mathcal{Z} \setminus \{z\}$ locally compact for any $z \in \mathcal{Z}$ – that are not degenerate, in the sense that $\rho_{\mathcal{X}}$ almost all $x \in \mathcal{X}$, there exists $t > 0$ such that the cardinality of the set defined as $\left\{z \,\middle|\, \mathbb{E}_{Y\sim\rho|_x}[\ell(z,Y)] - \inf_{z'\in\mathcal{Y}} \mathbb{E}_{Y\sim\rho|_x}[\ell(z',Y)] < t\right\}$ if finite. This holds for classification with infinite countable classes, but it does not for regression on the set of rational numbers.*

Remark 16 (Extension to general cases)**.** *To remove the condition $\mathcal{Z}$ finite, one can change the definition of $d(g^*(x), F)$ to $\inf_{\xi\in\mathcal{H};\{f^*(x)\}\neq\arg\min\langle\psi(z),\xi\rangle} \|\xi - g^*(x)\|$, in order to make Lemma 7 hold for any $\mathcal{Z}$.*

5.A.3 Equivalence between generalizations of the Tsybakov margin condition

While we state the margin condition with $d(g^*(x), F)$, it could also be stated with $d(g^*(x), F \cap \mathrm{Conv}(\varphi(\mathcal{Y})))$ or with, which is the quantity considered by (Nowak-Vila et al., 2019),

$$\gamma(x) = \inf_{z\neq z^*} \mathbb{E}_{Y\sim\rho|_x} \ell(z,Y) - \mathbb{E}_{Y\sim\rho|_x} \ell(z^*,Y) = \inf_{z\neq z^*} \langle \psi(z) - \psi(z^*), g^*(x)\rangle.$$

Indeed, when $\mathcal{Z}$ is finite and ℓ is proper in the sense that $\ell(\cdot, y) = \ell(\cdot, z)$ implies $z = y$, and that there is no z that minimizes a linear combination of $(\ell(\cdot, y))_{y\in\mathcal{Y}}$ without minimizing a convex combination of the same family, we have the existence of two constants such that

$$c\gamma(x) \leq d(g^*(x), F \cap \mathrm{Conv}(\varphi(\mathcal{Y}))) \leq d(g^*(x), F) \leq c'\gamma(x).$$

Mildness of our condition

Let z' be the argmin defining γ, geometric properties of the scalar product imply the existence of a $\xi \in (\psi(z') - \psi(z^*))^\perp$ such that

$$\langle \psi(z') - \psi(z^*), g^*(x)\rangle = \|\psi(z') - \psi(z^*)\| \, \|g^*(x) - \xi\|.$$

Therefore,

$$\langle \psi(z') - \psi(z^*), g^*(x)\rangle \geq \min_{y,y'} \|\psi(y) - \psi(y')\| \, \|g^*(x) - \xi\|.$$

Note that, by definition of ξ, $\langle \xi, \psi(z')\rangle = \langle \xi, \psi(z^*)\rangle$. If $\xi \in R_{z^*}$ then $\xi \in F$, otherwise $\xi \notin R_{z^*}$ and then, there exists a point between ξ and $g^*(x)$ that belongs to the decision frontier (see Appendix 5.A.2 for a proof - for which we need some regularity assumption such as $\mathcal{Z}$ finite). In every case,

$$\|g^*(x) - \xi\| \geq d(g^*(x), F).$$

This implies the existence of c'.

Strength of our condition

For any g_n such that $f_n(x) = z$, we have

$$\begin{aligned}\langle \psi(z) - \psi(z^*), g^*(x) \rangle &= \langle \psi(z), g^*(x) - g_n(x) \rangle + \langle \psi(z), g_n(x) \rangle - \langle \psi(z^*), g^*(x) \rangle \\ &\leq \langle \psi(z), g^*(x) - g_n(x) \rangle + \langle \psi(z^*), g_n(x) \rangle - \langle \psi(z^*), g^*(x) \rangle \\ &\leq 2c_\psi \left\| g^*(x) - g_n(x) \right\|.\end{aligned}$$

If we take the infimum on both sides we have

$$d(g^*(x), F) = \inf_{g_n(x) \notin R_{f^*(x)}} \|g_n(x) - g^*(x)\| \geq \frac{1}{2c_\psi} \inf_{z \neq z^*} \langle \psi(z) - \psi(z^*), g^*(x) \rangle,$$

where the left equality is provided, when $\mathcal{Z}$ is finite, by a similar reasoning to the one in Appendix 5.A.2. This implies the existence of c. Note also that if the loss is proper in the sense that if z minimizes $\langle \psi(z), \xi \rangle$ for a $\xi \in \mathcal{H}$, there exists a $\xi \in \operatorname{Conv} \varphi(\mathcal{Y})$ such that z minimizes $\langle \psi(z), \xi \rangle$, we can consider $g_n(x) \in \operatorname{Conv}(\varphi(\mathcal{Y}))$, and therefore restrict F to $F \cap \operatorname{Conv} \varphi(\mathcal{Y})$. Finally, we have shown that, when $\mathcal{Z}$ finite and ℓ proper

$$c\gamma(x) \leq d(g^*(x), F \cap \operatorname{Conv}(\varphi(\mathcal{Y}))) \leq d(g^*(x), F) \leq c'\gamma(x).$$

This explains why we would consider $\gamma(x)$, $d(g^*(x), F \cap \operatorname{Conv}(\varphi(\mathcal{Y})))$ or $d(g^*(x), F)$ to define the margin condition, it will only change the value of constants in Assumptions 3 and 4.

5.A.4 Refinement of Theorem 1

It is possible to refine Theorem 1 to remove the condition that the loss ℓ is bounded. In the following, we omit the dependency of L_n and M_n to n.

Lemma 17 (Refinement of Theorem 1)**.** *Under refined calibration* (5.5)*, concentration, Assumption 2, and no-density separation, Assumption 3, the risk is controlled as*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq 4c_\psi L^{-1/2} \exp\left(-\frac{t_0^2 L}{2}\right)^{1/2} + 4c_\psi M L^{-1} \exp\left(-\frac{t_0 L}{2M}\right).$$

Note that it is not possible to derive a better bound only given (5.4)*,* (5.5) *and* (5.6)*. Yet when ℓ is bounded by ℓ_∞, we have*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty \exp\left(-\frac{L t_0^2}{1 + M t_0}\right).$$

Proof. Using the calibration inequality along with the no-density separation one, we get

$$\begin{aligned}\mathcal{R}(f_n) - \mathcal{R}(f^*) &\leq 2c_\psi \, \mathbb{E}_X \left[\mathbf{1}_{\|g_n(X) - g^*(X)\| \geq t_0} \|g_n(X) - g^*(X)\| \right] \\ &= 2c_\psi \int_{t_0}^{\infty} \mathbb{P}_X \left(\|g_n(X) - g^*(X)\| \geq t \right) \mathrm{d}t.\end{aligned}$$

Taking the expectation over $\mathcal{D}_n$ and using concentration inequality we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) &\leq 2c_\psi \int_{t_0}^{\infty} \mathbb{P}_{X, \mathcal{D}_n} \left(\|g_n(X) - g^*(X)\| \geq t \right) \mathrm{d}t \\ &\leq 2c_\psi \int_{t_0}^{\infty} \exp\left(-\frac{L t^2}{1 + M t}\right) \mathrm{d}t.\end{aligned}$$

We only need to study the integral $\int_{t_0}^{\infty} \exp\left(-\frac{Lt^2}{1+Mt}\right) \mathrm{d}t$. We first clean the dependency on t inside the exponential using that

$$\frac{1}{2}\left(\exp(-Lt^2) + \exp\left(-\frac{Lt}{M}\right)\right) \leq \exp\left(-\frac{Lt^2}{1 + Mt}\right) \leq \exp\left(-\frac{Lt^2}{2}\right) + \exp\left(-\frac{Lt}{2M}\right).$$

We are left with the study of $\int_{t_0}^{\infty} \exp(-At^p)\,\mathrm{d}t$, for $p \in \{1, 2\}$ and $A > 0$. The case $p = 1$, directly leads to $A^{-1}\exp(-At_0)$, explaining the part in L/M. The case $p = 2$ is similar to the Gaussian integral, and can be handled with the following tricks

$$\begin{aligned}\int_{t_0}^{\infty} \exp(-At^2)\,\mathrm{d}t &= \frac{1}{2}\int_{(-\infty,-t_0]\cup[t_0,\infty)} \exp(-At^2)\,\mathrm{d}t \\ &= \frac{1}{2}\left(\int_{((-\infty,-t_0]\cup[t_0,\infty))^2} \exp(-A\,\|x\|^2))\,\mathrm{d}x\right)^{1/2}.\end{aligned}$$

This last integral corresponds to integrate the function $x \to \exp(-A\,\|x\|^2)$ for $x \in \mathbb{R}^2$ on the domain $((-\infty, -t_0] \cup [t_0, \infty))^2$. This function being positive and the domain being included in the domain $\{\|x\| \geq t_0\}$ and containing the domain $\left\{\|x\| \geq \sqrt{2}t_0\right\}$, we get

$$\int_{\left\{\|x\|\geq\sqrt{2}t_0\right\}}^{\infty} \exp(-A\,\|x\|^2)\,\mathrm{d}x \leq \left(2\int_{t_0}^{\infty} \exp(-At^2)\,\mathrm{d}t\right)^2 \leq \int_{\{\|x\|\geq t_0\}}^{\infty} \exp(-A\,\|x\|^2)\,\mathrm{d}x.$$

Using polar coordinate we get

$$\int_{\{\|x\|\geq t_0\}}^{\infty} \exp(-A\,\|x\|^2)\,\mathrm{d}x = 2\pi\int_{t_0}^{\infty} r\exp(-Ar^2)\,\mathrm{d}r = \pi A^{-1}\exp(-At_0^2).$$

Therefore,

$$2^{-1}\sqrt{\pi}A^{-1/2}\exp(-A2t_0^2)^{1/2} \leq \int_{t_0}^{\infty} \exp(-At^2)\,\mathrm{d}t \leq 2^{-1}\sqrt{\pi}A^{-1/2}\exp(-At_0^2)^{1/2}.$$

This explains the rates in L. □

5.A.5 Proof of Theorem 2

Using the calibration and Bernstein inequalities we get, omitting the dependency of L_n and M_n to n,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}\mathcal{R}(f_n) - \mathcal{R}(f^*) &\leq 2c_\psi\,\mathbb{E}_{\mathcal{D}_n,X}\left[\mathbf{1}_{d(g_n(X),g^*(X))\geq d(g^*(X),F)}\,\|g_n(X) - g^*(X)\|\right] \\ &= 2c_\psi\int_0^{\infty} \mathbb{P}_{\mathcal{D}_n,X}\left(\mathbf{1}_{d(g_n(X),g^*(X))\geq d(g^*(X),F)}\,\|g_n(X) - g^*(X)\| \geq t\right)\mathrm{d}t \\ &= 2c_\psi\int_0^{\infty} \mathbb{E}_X\,\mathbb{P}_{\mathcal{D}_n}\left(\|g_n(X) - g^*(X)\| \geq \max\{t, d(g^*(X), F)\}\right)\mathrm{d}t \\ &\leq 2c_\psi\int_0^{\infty} \mathbb{E}_X\exp\left(-\frac{L\max\{t, d(g^*(X),F)\}^2}{1 + M\max\{t, d(g^*(X),F)\}^2}\right)\mathrm{d}t \\ &= 2c_\psi\int_0^{\infty} \mathbb{E}_X\left[\mathbf{1}_{d(g^*(X),F)<t}\exp\left(-\frac{Lt^2}{1+Mt}\right)\right]\mathrm{d}t \\ &\quad + 2c_\psi\int_0^{\infty} \mathbb{E}_X\left[\mathbf{1}_{d(g^*(X),F)\geq t}\exp\left(-L\frac{d(g^*(X),F)^2}{1 + Md(g^*(X),F)}\right)\right]\mathrm{d}t \\ &= 2c_\psi\int_0^{\infty} \mathbb{P}_X\left(d(g^*(X),F) < t\right)\exp\left(-\frac{Lt^2}{1+Mt}\right)\mathrm{d}t \\ &\quad + 2c_\psi\,\mathbb{E}_X\left[d(g^*(X),F)\exp\left(-\frac{Ld(g^*(X),F)^2}{1 + Md(g^*(X),F)}\right)\right].\end{aligned}$$

Let us begin by working on the first term. We have, using the low-density separation hypothesis

$$\int_0^{\infty} \mathbb{P}_X\left(d(g^*(X),F) < t\right)\exp\left(-\frac{Lt^2}{1+Mt}\right)\mathrm{d}t \leq c_\alpha\int_0^{\infty} t^\alpha\exp(-\frac{Lt^2}{1+Mt})\,\mathrm{d}t.$$

Recall the expression of the Gamma integral

$$\int_0^\infty t^\alpha \exp(-Lt)\,\mathrm{d}t = \frac{\Gamma(\alpha+1)}{L^{\alpha+1}} \qquad \text{and} \qquad \int_0^\infty t^\alpha \exp(-Lt^2)\,\mathrm{d}t = \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{2L^{\frac{\alpha+1}{2}}}.$$

Let us briefly talk about optimality. Up to now, we have only used three inequalities: calibration, exponential concentration and low-density separation. Therefore, when those inequalities hold as equalities, we get a lower bound of order on the excess of risk as

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}\mathcal{R}(f_n) - \mathcal{R}(f^*) &\geq 2c_\psi c_\alpha \int_0^\infty t^\alpha \exp(-\frac{Lt^2}{1+Mt})\,\mathrm{d}t \\
&\geq 2c_\psi c_\alpha \int_0^\infty \frac{1}{2} t^\alpha \left(\exp(-Lt^2) + \exp\left(-\frac{Lt}{M}\right)\right)\mathrm{d}t \\
&= 2c_\psi c_\alpha \left(\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{4} L^{-\frac{\alpha+1}{2}} + \frac{\Gamma(\alpha+1)}{2} M^{\alpha+1} L^{-(\alpha+1)}\right).
\end{aligned}$$

For the upper bound, using that $\exp(-a/1+b) \leq \exp(-a/2) + \exp(-a/2b)$, we get

$$\begin{aligned}
\int_0^\infty t^\alpha \exp(-\frac{Lt^2}{1+Mt})\,\mathrm{d}t &\leq \int_0^\infty t^\alpha \exp(-\frac{Lt^2}{2})\,\mathrm{d}t + \int_0^\infty t^\alpha \exp(-\frac{Lt}{2M})\,\mathrm{d}t \\
&= 2^{\frac{\alpha-1}{2}}\Gamma\left(\frac{\alpha+1}{2}\right) L^{-\frac{\alpha+1}{2}} + 2^{\alpha+1}\Gamma(\alpha+1) M^{\alpha+1} L^{-(\alpha+1)}.
\end{aligned}$$

Let study the second term in the excess of risk inequality. To enhance readability, write $\eta(X) = d(g^*(X), F)$. We will first dissociate the two parts in the exponential with

$$\mathbb{E}_X\left[\eta(X)\exp\left(-\frac{L\eta(X)^2}{1+M\eta(X)}\right)\right] \leq \mathbb{E}_X\left[\eta(X)\left(\exp\left(-\frac{L\eta(X)^2}{2}\right) + \exp\left(-\frac{L\eta(X)}{2M}\right)\right).\right]$$

We are left with studying $\mathbb{E}[\eta(X)\exp(-A\eta(X)^p)]$, for $A > 0$ and $p \in \{1, 2\}$. The function $t \to t\exp(-At^p)$ achieves its maximum in $t_0 = (pA)^{-1/p}$, increasing before and decreasing after. Notice that the quantity

$$\begin{aligned}
\mathbb{P}(\eta(X) < t_0)\,\mathbb{E}_X\left[\eta(X)\exp(-A\eta(X)^p) \,|\, \eta(X) < t_0\right] &\leq c_\alpha t_0^{\alpha+1}\exp(-At_0^p) \\
&= c_\alpha p^{-\frac{\alpha+1}{p}}\exp(-p^{-1/p})A^{-\frac{\alpha+1}{p}},
\end{aligned}$$

is exactly of the same order as the control we had on the first term in the excess of risk decomposition. This suggests considering the following decomposition

$$\begin{aligned}
\mathbb{E}_X\left[\eta(X)\exp(-A\eta(X)^p)\right] &= \mathbb{P}(\eta(X) < t_0)\,\mathbb{E}_X\left[\eta(X)\exp(-A\eta(X)^p) \,|\, \eta(X) < t_0\right] \\
&+ \sum_{i=0}^\infty \mathbb{P}(2^i t_0 \leq \eta(X) < 2^{i+1}t_0)\,\mathbb{E}_X\left[\eta(X)\exp(-A\eta(X)^p) \,\middle|\, 2^i t_0 \leq \eta(X) < 2^{i+1}t_0\right] \\
&\leq c_\alpha t_o^{\alpha+1}\exp(-At_0^p) + \sum_{i=0}^\infty c_\alpha 2^\alpha (2^i t_0)^{\alpha+1}\exp(-At_0^p(2^i)^p) \\
&= c_\alpha t_o^{\alpha+1}\left(\exp(-p^{-1/p}) + \sum_{i=0}^\infty 2^\alpha 2^{i(\alpha+1)}\exp(-p^{-1/p}2^{ip})\right).
\end{aligned}$$

The convergence of the last series, ensures the existence of a constant c such that

$$\mathbb{E}_X\left[\eta(X)\exp\left(-\frac{L\eta(X)^2}{1+M\eta(X)}\right)\right] \leq c\left(\left(\frac{L}{2M}\right)^{-(\alpha+1)} + \left(\frac{L}{2}\right)^{-\frac{\alpha+1}{2}}\right).$$

Adding everything together, we get the existence of two constants c', c'', such that

$$\mathbb{E}_{\mathcal{D}_n}\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq 2c_\psi c_\alpha\left(c' M^{\alpha+1}L^{-(\alpha+1)} + c'' L^{-\frac{\alpha+1}{2}}\right).$$

This ends the proof by considering $c = \max(c', c'')$.

5.A.6 Refinement of Theorem 2

Some convergence analyses lead to exponential inequalities that are not of Bernstein type, indeed, our result still holds in those settings, as mentioned by the following lemma. In the following, we omit the dependency of L_n and M_n to n.

Lemma 18 (Refinement of Theorem 2)**.** *Under the assumptions of Theorem 2, if the concentration is not given by Assumption 2 but given, for some positive constants $(a_i, b_i, p_i)_{i\leq m}$, by, for all $x \in \operatorname{supp}\rho_{\mathcal{X}}$ and $t > 0$,*

$$\mathbb{P}_{\mathcal{D}_n}(\|g_n(x) - g^*(x)\| > t) \leq \sum_{i=1}^{n} a_i \exp(-b_i t^{p_i}).$$

Then the excess of risk is controlled by

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq c \sum_{i=1}^{n} a_i b_i^{-\frac{\alpha+1}{p_i}},$$

for a constant c that does not depend on $(a_i, b_i)_{i\leq m}$.

Proof. First of all, remark that the proof of Theorem 2 is linear in $\mathbb{P}_{\mathcal{D}_n}(\|g_n(x) - g^*(x)\| > t)$, therefore we only need to prove this lemma for (a, b, p), for which we proceed as in Theorem 2

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) &\leq 2c_\psi \, \mathbb{E}_{\mathcal{D}_n, X} \left[\mathbf{1}_{\|g_n(X)-g(X)\| \geq d(g(X),F)} \, \|g_n(X) - g(X)\| \right] \\
&= 2c_\psi \int_0^\infty \mathbb{P}_{\mathcal{D}_n, X} \left(\mathbf{1}_{\|g_n(X)-g(X)\| \geq d(g(X),F)} \, \|g_n(X) - g(X)\| \geq t \right) \mathrm{d}t \\
&= 2c_\psi \int_0^\infty \mathbb{E}_X \, \mathbb{P}_{\mathcal{D}_n} \left(\|g_n(X) - g(X)\| \geq \max\{t, d(g(X), F)\} \right) \mathrm{d}t \\
&\leq 2c_\psi a \int_0^\infty \mathbb{E}_X \exp\left(-b \max\{t, d(g(X), F)\}^p \right) \mathrm{d}t \\
&= 2c_\psi a \int_0^\infty \mathbb{E}_X \left[\mathbf{1}_{d(g(X),F)<t} \exp\left(-bt^p\right) \right] \mathrm{d}t \\
&\qquad + 2c_\psi a \int_0^\infty \mathbb{E}_X \left[\mathbf{1}_{d(g(X),F)\geq t} \exp\left(-bd(g(X), F)^p\right) \right] \mathrm{d}t \\
&= 2c_\psi a \int_0^\infty \mathbb{P}_X \left(d(g(X), F) < t \right) \exp\left(-bt^p\right) \mathrm{d}t \\
&\qquad + 2c_\psi a \, \mathbb{E}_X \left[d(g(X), F) \exp\left(-bd(g(X), F)^p\right) \right].
\end{aligned}$$

Let us begin by working on the first term. We have, using the low-density separation hypothesis

$$\begin{aligned}
\int_0^\infty \mathbb{P}_X \left(d(g(X), F) < t \right) \exp\left(-bt^2\right) \mathrm{d}t &\leq c_\alpha \int_0^\infty t^\beta \exp(-bt^p) \, \mathrm{d}t. \\
&= b^{-\frac{1+\beta}{p}} c_\alpha \int_0^\infty (b^{1/p}t)^\beta \exp(-(b^{1/p}t)^p) \, \mathrm{d}(b^{1/p}t). \\
&= b^{-\frac{1+\beta}{p}} c_\alpha \int_0^\infty t^\beta \exp(-t^p) \, \mathrm{d}t = c_\alpha \Gamma(\beta, p) b^{-\frac{1+\beta}{p}}.
\end{aligned}$$

Let study the second term in the excess of risk inequality. To enhance readability, write $\eta(X) = d(g(X), F)$. We are left with studying $\mathbb{E}[\eta(X) \exp(-b\eta(X)^p)]$. The function $t \to t \exp(-bt^p)$ achieves its maximum in $t_0 = (pb)^{-1/p}$, increasing before and decreasing after. Notice that the quantity

$$\begin{aligned}
\mathbb{P}(\eta(X) < t_0) \, \mathbb{E}_X \left[\eta(X) \exp(-b\eta(X)^p) \,|\, \eta(X) < t_0 \right] &\leq c_\alpha t_0^{\beta+1} \exp(-bt_0^p) \\
&= c_\alpha p^{-\frac{\beta+1}{p}} \exp(-p^{-1/p}) b^{-\frac{\beta+1}{p}},
\end{aligned}$$

is exactly of the same order as the control we had on the first term in the excess of risk decomposition. This

suggests considering the following decomposition

$$\begin{aligned}\mathbb{E}_X\left[\eta(X)\exp(-b\eta(X)^p)\right] &= \mathbb{P}(\eta(X) < t_0)\,\mathbb{E}_X\left[\eta(X)\exp(-b\eta(X)^p) \,|\, \eta(X) < t_0\right] \\ &\quad + \sum_{i=0}^{\infty} \mathbb{P}(2^i t_0 \le \eta(X) < 2^{i+1}t_0)\,\mathbb{E}_X\left[\eta(X)\exp(-b\eta(X)^p) \,\middle|\, 2^i t_0 \le \eta(X) < 2^{i+1}t_0\right] \\ &\le c_\alpha t_o^{\beta+1}\exp(-bt_0^p) + \sum_{i=0}^{\infty} c_\alpha 2^\beta (2^i t_0)^{\beta+1}\exp(-bt_0^p(2^i)^p) \\ &= c_\alpha t_o^{\beta+1}\left(\exp(-p^{-1/p}) + \sum_{i=0}^{\infty} 2^\beta 2^{i(\beta+1)}\exp(-p^{-1/p}2^{ip})\right).\end{aligned}$$

The convergence of the last series ensures the existence of a constant c' such that

$$\mathbb{E}_X\left[\eta(X)\exp(-b\eta(X)^p)\right] \le c' b^{-\frac{\beta+1}{p}}.$$

Adding everything together ends the proof of this lemma. Note that we have the same type of optimality as the one stated in Theorem 2. □

Because we use concentration inequalities for terms that are not necessarily centered, we usually get that (5.4) only holds for $t > \varepsilon_0$ where, typically $\varepsilon_0 = \left\|\mathbb{E}_{\mathcal{D}_n} g_n(x) - g^*(x)\right\|$, we can bypass this problem by adding in $\mathbf{1}_{t<\varepsilon_0}$ in the probability, motivating the study leading to the following lemma.

Lemma 19 (Handling bias in concentration inequality)**.** *Under the assumptions of Theorem 2, if the concentration is not given by Assumption 2 but given, for an $\varepsilon_0 > 0$, by, for all $x \in \operatorname{supp}\rho_{\mathcal{X}}$ and $t > 0$,*

$$\mathbb{P}_{\mathcal{D}_n}(\|g_n(x) - g^*(x)\| > t) \le \mathbf{1}_{t<\varepsilon_0}.$$

Then the excess of risk is controlled by

$$\mathbb{E}_{\mathcal{D}_n}\mathcal{R}(f_n) - \mathcal{R}(f^*) \le 2c_\psi c_\alpha \varepsilon_0^{\alpha+1}.$$

Proof. We retake the beginning of the proof of Theorem 2, and change its ending with

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}\mathcal{R}(f_n) - \mathcal{R}(f^*) &\le 2c_\psi\,\mathbb{E}_{\mathcal{D}_n,X}\left[\mathbf{1}_{\|g_n(X)-g(X)\|\ge d(g(X),F)}\,\|g_n(X) - g(X)\|\right] \\ &= 2c_\psi\int_0^\infty \mathbb{P}_{\mathcal{D}_n,X}\left(\mathbf{1}_{\|g_n(X)-g(X)\|\ge d(g(X),F)}\,\|g_n(X) - g(X)\| \ge t\right)\mathrm{d}t \\ &= 2c_\psi\int_0^\infty \mathbb{E}_X\,\mathbb{P}_{\mathcal{D}_n}\left(\|g_n(X) - g(X)\| \ge \max\{t, d(g(X),F)\}\right)\mathrm{d}t \\ &\le 2c_\psi\int_0^\infty \mathbb{E}_X\,\mathbf{1}_{t<\varepsilon_0}\mathbf{1}_{d(g(X),F)<\varepsilon_0}\,\mathrm{d}t = 2c_\psi\varepsilon_0\,\mathbb{P}_X\left(d(g(X),F) < \varepsilon_0\right)\mathrm{d}t.\end{aligned}$$

This leads to the result after applying the α-margin condition. □

5.B Nearest neighbors

5.B.1 Usual assumptions to derive nearest neighbors convergence rates

Assumption 6 can be seen as the backbone to control $\left\|g_n^*(x) - g^*(x)\right\|$ in a uniform manner. This assumption that relates the regularity of g^* with the density of $\rho_{\mathcal{X}}$ has been historically approached in the following manner. Assume that g^* is β'-Hölder, that is, for any $x, x' \in \operatorname{supp}\rho_{\mathcal{X}}$

$$\|g^*(x) - g^*(x')\| \le a_1 d(x,x')^{\beta'}.$$

Suppose that $\mathcal{X} = \mathbb{R}^d$, that $\rho_{\mathcal{X}}$ is continuous against λ, the Lebesgue measure, with minimal mass in the sense that there exists a $p_{\min} > 0$ such that $\frac{\mathrm{d}\rho_{\mathcal{X}}}{\mathrm{d}\lambda}(\mathcal{X})$ does not intersect $(0, p_{\min})$, and that $\operatorname{supp}\rho_{\mathcal{X}}$ has regular boundaries in the sense that there exist $a_2, t_0 > 0$ such that for any $x \in \operatorname{supp}\rho_{\mathcal{X}}$ and $t \in (0, t_0)$

$$\lambda\left(\mathcal{B}(x,t) \cap \operatorname{supp}\rho_{\mathcal{X}}\right) \ge a_2\lambda\left(\mathcal{B}(x,t)\right).$$

For example an orthant satisfies this property with $a_2 = 2^{-d}$ and $t_0 = \infty$, and $\mathcal{B}(0,1)$ satisfies this property with $a_2 = \lambda\left(\mathcal{B}(0,1) \cap \mathcal{B}(1,1)\right) / \lambda\left(\mathcal{B}(0,1)\right)$ and $t_0 = 1$. In such a setting, we get

$$\|g^*(x) - g^*(x')\| \leq a_1 d(x,x')^{\beta'} = a_1 \left(\frac{\lambda(\mathcal{B}(x, d(x,x')))}{\lambda(\mathcal{B}(0,1))}\right)^{\frac{\beta'}{d}}.$$

When $d(x,x') < t_0$, we have

$$\lambda(\mathcal{B}(x, d(x,x'))) \leq a_2^{-1} \lambda\left(\mathcal{B}(x, d(x,x')) \cap \operatorname{supp} \rho_{\mathcal{X}}\right) \leq a_2^{-1} p_{\min}^{-1} \rho_{\mathcal{X}}(\mathcal{B}(x, d(x,x')))^{\beta}.$$

This means that for any $x \in \operatorname{supp}\rho_{\mathcal{X}}$ and $x' \in \mathcal{B}(x, t_0)$ we have, with $\beta = \frac{\beta'}{d}$ and the constant $a_3 = a_1 a_2^{-\beta} p_{\min}^{-\beta} \lambda(\mathcal{B}(0,1))^{-\beta}$

$$\|g^*(x) - g^*(x')\| \leq a_3 \rho_{\mathcal{X}}(\mathcal{B}(x, d(x,x')))^{\beta}.$$

While, we actually do not need the bound to hold for $d(x,x') > t_0$ in the following proof, to check the veracity of our remark on Assumption 6, one can verify that under our assumptions on $\rho_{\mathcal{X}}$, $\operatorname{supp}\rho_{\mathcal{X}}$ is bounded, and therefore g^* is too. And if g^* is bounded by c_φ, by considering $a_3' = \max\left(2c_\varphi a_2^{-\beta} p_{\min}^{-\beta} t_0^{-\beta'}, a_3\right)$, this bound holds for any $x, x' \in \operatorname{supp}\rho_{\mathcal{X}}$.

5.B.2 Proof of Lemma 12

Control of the variance term. For $x \in \rho_{\mathcal{X}}$, the variance term can be written

$$\left\|g_n(x) - g_n^*(x)\right\| = \left\|\frac{1}{k}\sum_{i=1}^{k} \varphi(Y_{(i)}) - \mathbb{E}\left[Y_{(i)} \,\middle|\, X_{(i)}\right]\right\|.$$

Where the index (i) is such that $X_{(i)}$ is the i-th nearest neighbor of x in $(X_i)_{i\leq n}$. Since, given $(X_i)_{i\leq n}$, the $(Y_i)_{i\neq n}$ are independent, distributed according to $\otimes_{i\leq n}\rho|_{X_i}$, we can use a concentration inequality to control it. We recall Bernstein concentration inequality in such spaces, derived by Yurinskii (1970), we will use the formulation of Corollary 1 from Pinelis and Sakhanenko (1986).

Theorem 7 (Concentration in Hilbert space (Pinelis and Sakhanenko, 1986))**.** *Let denote by $\mathcal{A}$ a Hilbert space and by (ξ_i) a sequence of independent random vectors on $\mathcal{A}$ such that $\mathbb{E}[\xi_i] = 0$, and that there exists $M, \sigma^2 > 0$ such that for any $m \geq 2$*

$$\sum_{i=1}^{n} \mathbb{E}\left[\|\xi_i\|^m\right] \leq \frac{1}{2} m! \sigma^2 M^{m-2}.$$

Then for any $t > 0$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \xi_i\right\| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + 2tM}\right).$$

This explains Assumption 5, allowing, because there is only k ξ_i active in $\sum_{i=1}^n \alpha_i(x)\xi_i$, to get

$$\mathbb{P}_{\mathcal{D}_n}\left(\left\|g_n(x) - g_n^*(x)\right\| > t\right) \leq 2\exp\left(-\frac{kt^2}{2\sigma^2 + 2Mt}\right).$$

Control of the bias term. Under the Modified Lipschitz condition, Assumption 6,

$$\begin{aligned}\left\|g_n^*(x) - g_n(x)\right\| &= \left\|\sum_{i=1}^{n} \alpha_i(x)\left(g_n(x) - g^*(X_i)\right)\right\| \leq \sum_{i=1}^{n} \alpha_i(x)\left\|g_n(x) - g^*(X_i)\right\| \\ &\leq c_\beta \sum_{i=1}^{n} \alpha_i(x) \rho_{\mathcal{X}}\left(\mathcal{B}(x, d(x,X_i))\right)^{\beta} \leq c_\beta \rho_{\mathcal{X}}\left(\mathcal{B}(x, d(x,X_k(x)))\right)^{\beta}.\end{aligned}$$

When $\rho_{\mathcal{X}}$ is continuous, it follows from the probability integral transform (also known as universality of the uniform) that $\rho_{\mathcal{X}}\left(\mathcal{B}(x, d(x, X_k(x)))\right)$ behaves like the k-th order statistics of a sample $(U_i)_{i\leq n}$ of n uniform distributions on $[0,1]$. Therefore, for any $s \in [0,1]$

$$\mathbb{P}_{\mathcal{D}_n}\left(\rho_{\mathcal{X}}\left(\mathcal{B}(x, d(x, X_k(x)))\right) > s\right) = \mathbb{P}\left(\sum_{i=1}^{n} \mathbf{1}_{U_i<s} \leq k\right).$$

Recall the multiplicative Chernoff bound, stating that for $(Z_i)_{i\leq n}$ n independent random variables in $\{0,1\}$, if $Z = \sum_{i=1}^{n} Z_i$, and $\mu = \mathbb{E}[Z]$, for any $\delta > 0$

$$\mathbb{P}\left(Z \leq (1-\delta)\mu\right) \leq \exp\left(-\frac{\delta^2\mu}{2}\right).$$

Since, for $s \in [0,1]$, $\mathbb{E}[\mathbf{1}_{U_i<s}] = \mathbb{P}(U_i < s) = s$, we get, when $k \leq ns/2$

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathbf{1}_{U_i<s} \leq k\right) \leq \exp\left(-\frac{(ns-k)^2}{2ns}\right) \leq \exp\left(-\frac{ns}{8}\right).$$

With $s = c_\beta^{-1} t^{\frac{1}{\beta}}$, we get

$$\mathbb{P}_{\mathcal{D}_n}\left(\left\|g_n^*(x) - g_n(x)\right\| > t\right) \leq \exp\left(-\frac{nt^{\frac{1}{\beta}}}{8c_\beta}\right).$$

Remark that when g^* is β' Hölder, we get the same result with $\rho_{\mathcal{X}}(\mathcal{B}(x,t))$ instead of $t^{\frac{1}{\beta}}$ by considering $\mathbf{1}_{X_i\in\mathcal{B}(x,t)}$ instead of $\mathbf{1}_{U_i\leq t}$. Note that there is a way to bound a Binomial distribution with a Gaussian for t smaller than the mean of the binomial distribution, which would lead to a bound that holds for any $t > 0$ (Slud, 1977).

5.B.3 Proof of Theorem 3

Using the proof of Theorem 1, we get

$$\mathbb{E}_{\mathcal{D}_n}\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty \mathbb{P}_{\mathcal{D}_n}\left(\left\|g(x) - g_n^*(x)\right\| > t_0\right).$$

Because $\|g_n(x) - g^*(x)\| > t_0$ implies that either $\left\|g_n(x) - g_n^*(x)\right\| > t_0/2$ or $\left\|g_n^*(x) - g^*(x)\right\| > t_0/2$, we get using Lemma 12

$$\mathbb{P}_{\mathcal{D}_n}\left(\left\|g(x) - g_n^*(x)\right\| > t_0\right) \leq 2\exp\left(-\frac{b_1 k t_0^2}{4 + 2b_2 t_0}\right) + \exp\left(-2^{-\frac{1}{\beta}} b_3 n t_0^{\frac{1}{\beta}}\right) + \mathbf{1}_{t_0<(k/2n)^\beta}.$$

This explains the result of Theorem 3.

5.B.4 Proof of Theorem 4

First of all for $t > 0$, and $x \in \operatorname{supp}\rho_{\mathcal{X}}$, because $\|g_n(x) - g^*(x)\| > t$ implies that either $\left\|g_n(x) - g_n^*(x)\right\| > t/2$ or $\left\|g_n^*(x) - g^*(x)\right\| > t/2$, we have the inclusion of events:

$$\{\mathcal{D}_n \mid \|g_n(x) - g^*(x)\| > t\} \subset \left\{\mathcal{D}_n \,\middle|\, \left\|g_n(x) - g_n^*(x)\right\| > t/2\right\} \cup \{\mathcal{D}_n \mid \|g_n(x) - g^*(x)\| > t/2\},$$

which translates in terms of probability as

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_n}\left(\|g_n(x) - g^*(x)\| > 2t\right) &\leq \mathbb{P}_{\mathcal{D}_n}\left(\left\|g_n(x) - g_n^*(x)\right\| > t\right) + \mathbb{P}_{\mathcal{D}_n}\left(\left\|g_n^*(x) - g^*(x)\right\| > t\right). \\ &\leq 2\exp\left(-\frac{b_1 k t^2}{1 + b_2 t}\right) + \exp\left(-b_3 n t^{\frac{1}{\beta}}\right) + \mathbf{1}_{t<\left(\frac{k}{2n}\right)^\beta}.\end{aligned}$$

Using the refinements of Theorem 2 exposed in Appendix 5.A.6, we get that there exists a constant $c > 0$ that does not depend on k or n such that

$$\mathbb{E}_{\mathcal{D}_n}\mathcal{R}f_n - \mathcal{R}(f^*) \leq c\left(k^{-\frac{\alpha+1}{2}} + n^{-\beta(\alpha+1)} + (nk^{-1})^{\beta(\alpha+1)}\right).$$

We optimize this last quantity with respect to k by taking $k = n^\gamma$, and choosing γ such that $\gamma = 2(1-\gamma)\beta$ leading to $\gamma = 2\beta/(2\beta+1)$ and to rates in n to the power minus $\beta(\alpha+1)/(2\beta+1)$.

5.B.5 Numerical experiments

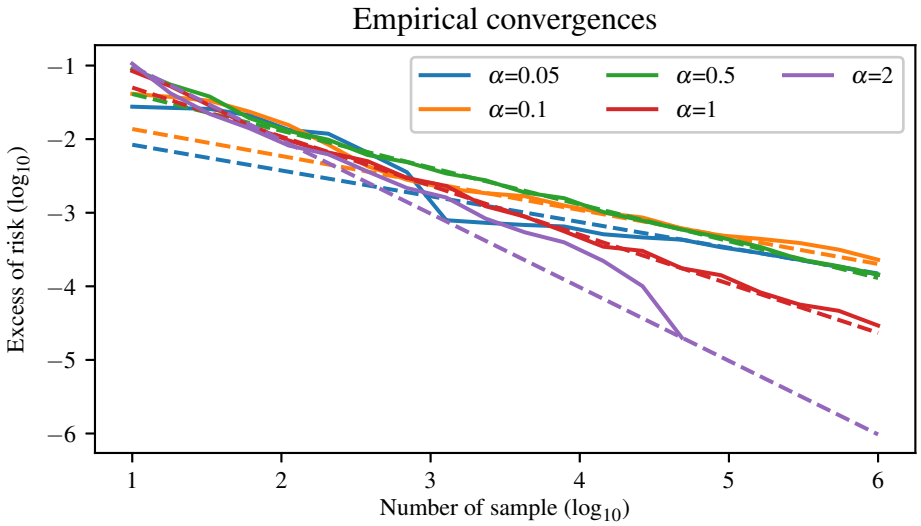

Figure 5.4: Supplement to Figure 5.3. We specify the error is evaluated on 100 points forming a regular partition of $\mathcal{X} = [-1, 1]$, and the expectation $\mathbb{E}_{\mathcal{D}_n}$ is approximated by considering 100 datasets. The violet curve is cropped at $n \approx 10^5$, because the error was null afterwards with our evaluation parameters (only 100 points to evaluate the error), forbidding us to consider the logarithm of the excess of risk.

Interestingly, on numerical simulations such as the one presented on Figure 5.3, we observed two regimes. A first regime where bound is meaningless because of constants being too big, and where the error decreases independently of the exponent expected through Theorem 4, and a final regime where rates correspond to the bond given by the theorem. Note that when $\alpha >> 1$, with our computation parameter, we do not get to really illustrate convergence rates, as this final regime get place for bigger n than what we have considered ($n_{\max} = 10^6$), this being partly due to the constant c_β in Assumption 6 being big for the g^* we considered. Furthermore, note that, for example, if a problem satisfied Assumption 3 with a tiny t_0, we expect that exponential convergence rates are only going to be observed for $n > N$, with N huge, and for which the excess of risk is already minuscule.

5.C Kernel proofs

In this section, we study L^∞ convergence rates of the kernel ridge regression estimate. We use the L^2-proof scheme of Caponnetto and De Vito (2006) with the remark of Ciliberto et al. (2016) to factorize the action of K on $L^2(\mathcal{X}, \mathcal{H}, \rho_\mathcal{X})$ through its action on $L^2(\mathcal{X}, \mathbb{R}, \rho_\mathcal{X})$. We retake the work of Pillaud-Vivien et al. (2018a) to relax the source condition, and use Fischer and Steinwart (2020) to cast in L^∞ thanks to interpolation inequality. While those results, leading to Lemma 14, are not new, we present them entirely to provide the reader with self-contained materials.

5.C.1 Construction of reproducing kernel Hilbert space (RKHS)

In the following, we suppose that k is bounded by κ^2.

Vector-valued RKHS. To study the estimator g_n, it is useful to introduce the reproducing kernel Hilbert space $\mathcal{G}$ associated with k and $\mathcal{H}$ (Aronszajn, 1950). To define $\mathcal{G}$, define the atoms $k_x : \mathcal{H} \to \mathcal{G}$ and the scalar product, for $x, x' \in \mathcal{X}$ and $\xi, \xi' \in \mathcal{H}$ as

$$\langle k_x \xi, k_{x'} \xi' \rangle_\mathcal{G} = \langle \xi, \Gamma(x, x') \xi' \rangle = k(x, x') \langle \xi, \xi' \rangle_\mathcal{H} .$$

Where Γ is the vector valued kernel inherited from k as $\Gamma(x,x') = k(x,x')I_{\mathcal{H}}$ (Schwartz, 1964). $\mathcal{G}$ is defined as the closure, under the metric induced by this scalar product, of the span of the atoms $k_x\xi$ for $x \in \mathcal{X}$ and $\xi \in \mathcal{H}$. Note that k_x is linear, and continuous of norm $\|k_x\|_{\text{op}} = \sqrt{k(x,x)}$. When $k(\cdot,x)$ is square integrable for all $x \in \operatorname{supp}\rho_{\mathcal{X}}$, $\mathcal{G}$ is homomorphic to a functional space in $L^2(\mathcal{X},\mathcal{H},\rho_{\mathcal{X}})$ through the linear mapping S that associates the atom $k_x\xi$ in $\mathcal{G}$ to the function $k(\cdot,x)\xi$ in L^2, defined formally as

$$\begin{array}{rccc} S: & \mathcal{G} & \to & L^2 \\ & \gamma & \to & x \to k_x^\star\gamma. \end{array}$$

While intrinsically similar, it is useful to distinguish between $\mathcal{G}$ and $\operatorname{im} S \subset L^2$. Note that S is continuous, since on atom $k_x\xi$, $\|Sk_x\xi\|_{L^2} \leq \|k_x(\cdot)\|_{L^2}\|\xi\|_{\mathcal{H}} \leq k(x,x)\|\xi\|_{\mathcal{H}} = \|k_x\xi\|_{\mathcal{G}}$. The fact that S is a bounded operator justifies the introduction of the following operators.

Central operators. In the following, we will make an extensive use of $S^\star : L^2(\mathcal{X},\mathcal{H},\rho_{\mathcal{X}}) \to \mathcal{G}$ the adjoint of S, defined as $S^\star g = \mathbb{E}_{\rho_{\mathcal{X}}}[k_X g(X)]$; the covariance operator $\Sigma : \mathcal{G} \to \mathcal{G}$, defined as $\Sigma := S^\star S = \mathbb{E}_{\rho_{\mathcal{X}}}[k_X k_X^\star]$; and its action on L^2, $K : L^2(\mathcal{X},\mathcal{H},\rho_{\mathcal{X}}) \to L^2(\mathcal{X},\mathcal{H},\rho_{\mathcal{X}})$, defined as $Kg := SS^\star g = \mathbb{E}_X[k(\cdot,X)g(X)]$. Finally, we have defined the four central operators

$$\begin{array}{ll} S\gamma = k_{(\cdot)}^\star\gamma, & S^\star g = \mathbb{E}_{\rho_{\mathcal{X}}}[k_X g(X)] \\ \Sigma := S^\star S = \mathbb{E}_{\rho_{\mathcal{X}}}[k_X k_X^\star], & Kg := SS^\star g = \mathbb{E}_X[k(\cdot,X)g(X)]. \end{array} \tag{5.13}$$

It should be noted that this construction is usually avoided since, based on the fact that the Frobenius norm of K behave like $\dim(\mathcal{H})$, meaning that when $\mathcal{H}$ is infinite dimensional, K is not a compact operator. However, since we consider $\mathcal{Z}$ finite, we can always consider $\mathcal{H} = \mathbb{R}^{|\mathcal{Z}|}$ with $\varphi(y) = (\ell(z,y))_{z\in\mathcal{Z}}$ and $\psi(z) = (\mathbf{1}_{z=z'})_{z'\in\mathcal{Z}}$, and moreover, we will see that a way can be worked out, even when $\mathcal{H}$ is infinite dimensional, which was already shown by Ciliberto et al. (2016).

Relation between real-valued versus vector-valued RKHS.

Usually convergence in RKHS is studied for real-valued functions. We need convergence results for vector-valued functions. As mentioned above, we only need the results for Euclidean space, however, we will do it for functions that are maps going into potentially infinite dimensional Hilbert space. Indeed, this does not lead to major complications. We provide here one way to get around this issue. An alternative formal way to proceed can be found (Ciliberto et al., 2016).

Real-valued RKHS. We build the real-valued RKHS $\mathcal{G}_{\mathcal{X}}$ as the closure of the span of the atoms $\bar{k}_x$ for $x \in \mathcal{X}$, under the metric induced by the scalar product $\langle \bar{k}_x, \bar{k}_{x'} \rangle = k(x,x')$. Similarly, we build $\bar{S}$, $\bar{S}^\star$, $\bar{\Sigma}$ and $\bar{K}$. We shall see that the action of Σ on $\mathcal{G}$ can be factorized through its actions $\bar{\Sigma}$ on $\mathcal{G}_{\mathcal{X}}$.

Algebraic equivalences. Based on the fact that $\|\bar{k}_x\|_{\mathcal{G}_{\mathcal{X}}} = \|k_x\|_{\text{op}} = \sqrt{k(x,x)}$, it is possible to build an isometry that match $\bar{k}_x$ in $\mathcal{G}_{\mathcal{X}}$ to k_x in the space of continuous linear operator from $\mathcal{H}$ to $\mathcal{G}$. With $(\bar{e}_i)_{i\in\mathbb{N}}$ an orthogonal basis of $\mathcal{G}_{\mathcal{X}}$, and $(f_j)_{j\in\mathbb{N}}$ an orthogonal basis of $\mathcal{H}$, we get an orthogonal basis $(e_i f_j)_{i,j\in\mathbb{N}}$ of $\mathcal{G}$. This is exactly the construction $\mathcal{G} = \mathcal{G}_{\mathcal{X}} \otimes \mathcal{H}$ of (Ciliberto et al., 2016).

Note that for μ_1, μ_2 two measures on $\mathcal{X}$, we can check that

$$\begin{aligned} \left\|\mathbb{E}_{X\sim\mu_1}[k_X k_X^\star]\,\mathbb{E}_{X_0\sim\mu_2}[k_{X_0}]\right\|_{\text{op}}^2 &= \left\|\mathbb{E}_{X\sim\mu_1}[\bar{k}_X \bar{k}_X^\star]\,\mathbb{E}_{X_0\sim\mu_2}[\bar{k}_{X_0}]\right\|_{\mathcal{G}_{\mathcal{X}}}^2 \\ &= \mathbb{E}_{X,X'\sim\mu_1;X,X'\sim\mu_2}[k(X_0,X)k(X,X')k(X',X_0')]. \end{aligned}$$

This explains that we will allow ourselves to write derivations of the type

$$\left\|(\Sigma+\lambda)^{-\frac{1}{2}}k_x g_n(x)\right\|_{\mathcal{G}} \leq \left\|(\bar{\Sigma}+\lambda)^{-\frac{1}{2}}\bar{k}_x\right\|_{\mathcal{G}_{\mathcal{X}}} \|g_n(x)\|_{\mathcal{H}}\,.$$

Note also that for $g := \sum_{ij} c_{ij} e_i f_j \in \mathcal{G}$, with $\sum_{ij} c_{ij}^2 = 1$, $\bar{c}_i := (c_{ij})_{j\in\mathbb{N}} \in \ell^2$, $\bar{A}$ a self-adjoint operator on

$\mathcal{G}_\mathcal{X}$ and A its version on $\mathcal{G}$, we have

$$\|Ag\|_\mathcal{G}^2 = \sum_{ijk} c_{ij}c_{kj} \left\langle \bar{A}\bar{e}_i, \bar{A}\bar{e}_k \right\rangle_\mathcal{G} = \sum_{ij} \left\langle \bar{c}_i, \bar{c}_j \right\rangle_{\ell^2} \left\langle \bar{A}\bar{e}_i, \bar{A}\bar{e}_k \right\rangle_{\mathcal{G}_\mathcal{X}}$$
$$\leq \sum_{ij} \|\bar{c}_i\|_{\ell^2} \left\|\bar{c}_j\right\|_{\ell^2} \left\langle \bar{A}\bar{e}_i, \bar{A}\bar{e}_k \right\rangle_{\mathcal{G}_\mathcal{X}} = \left\| \bar{A} \sum_i \|\bar{c}_i\|_{\ell^2} \bar{e}_i \right\|_{\mathcal{G}_\mathcal{X}}^2 \leq \left\|\bar{A}\right\|_{\text{op}}^2,$$

which explains why we will consider derivations of the type

$$\left\|(\Sigma+\lambda)^{\frac{1}{2}}(\hat{\Sigma}+\lambda)^{-1}(\Sigma+\lambda)^{\frac{1}{2}}\right\|_{\text{op}} \leq \left\|(\bar{\Sigma}+\lambda)^{\frac{1}{2}}(\hat{\bar{\Sigma}}+\lambda)^{-1}(\bar{\Sigma}+\lambda)^{\frac{1}{2}}\right\|_{\text{op}}.$$

Finally, notice that because of the same consideration, if $(\bar{u}_i)_{i\in\mathbb{N}} \in \mathcal{G}_\mathcal{X}^\mathbb{N}$ diagonalize $\bar{A}$, $(u_i f_j)_{i,j\leq\mathbb{N}} \in \mathcal{G}^{\mathbb{N}\times\mathbb{N}} \simeq \mathcal{G}^\mathbb{N}$ diagonalize A in $\mathcal{G}$. This justifies the consideration of fractional operators A^p for $p \in \mathbb{R}_+$, such as in Assumptions 8 and 9. Based on this equivalence, we will forget the bar notations, we incite the careful and attentive reader to recover them.

5.C.2 Estimate g_n as an empirical approximate projection on RKHS

To obtain bounds like (5.4), it is sufficient to control the convergence of g_n to g^* in L^∞. Assumption 8 allow us to cast in L^2 the study of the convergence in L^∞. The convergence of g_n toward g^* can be split in two terms, a term expressing the convergence of g_λ toward g^* that is based on geometrical properties and a term expressing the convergence of g_n toward g_λ, that is based on concentration inequalities in $\mathcal{G}$, such as the ones given by Pinelis and Sakhanenko (1986); Minsker (2017). For this last term, we need to characterize g_n and g_λ with the following lemma.

Lemma 20 (Approximation of integral operators)**.** *g_n can be understood as the empirical approximation of g_λ since*

$$g_n = S(\mathbb{E}_{\hat{\rho}}[k_X k_X^\star] + \lambda)^{-1} \mathbb{E}_{\hat{\rho}}[k_X \varphi(Y)], \qquad g_\lambda = S(\mathbb{E}_\rho[k_X k_X^\star] + \lambda)^{-1} \mathbb{E}_\rho[k_X \varphi(Y)],$$

with $\hat{\rho} = \frac{1}{n}\sum_{i=1}^n \delta_{X_i} \otimes \delta_{Y_i}$,

Proof. Indeed, the expression of g_n and its convergences toward g^* will be understood thanks to the operator S and its derivatives. When $\operatorname{im} S$ is closed in L^2, on can be defined the orthogonal projection of g^* to $\operatorname{im} S$, with the L^2 metric as $\pi_{\operatorname{im} S}(g^*) = S(S^\star S)^\dagger S^\star g^*$. When $\operatorname{im} S$ is not closed, or equivalently when Σ has positive eigenvalues converging to zero, one can define approximate orthogonal projection, through eigenvalue thresholding or Tikhonov regularization. This last choice leads to the estimate

$$g_\lambda = S(\Sigma+\lambda)^{-1}S^\star g^* = S(S^\star S+\lambda)^{-1}S^\star g^* = SS^\star(SS^\star+\lambda)^{-1}g^* = K(K+\lambda)^{-1}g^*.$$

Note that, because of the Bayes optimum characterization of g^*, $S^\star g^* = \mathbb{E}_\rho[k_X\varphi(Y)]$. This explains the characterization of g_λ.

Interestingly, the approximation of ρ by $\hat{\rho}$ can be thought with the approximation of $L^2(\mathcal{X},\mathcal{H},\rho_\mathcal{X})$ by $\ell^2(\mathcal{H}^n) \simeq L^2(\mathcal{X},\mathcal{H},\hat{\rho}_\mathcal{X})$ where for $\Xi = (\xi_i), Z = (\zeta_i) \in \mathcal{H}^n$,

$$\langle \Xi, Z\rangle_{\ell^2} = \frac{1}{n}\sum_{i=1}^n \langle \xi_i, \zeta_i\rangle_\mathcal{H},$$

and with the empirical probability measure $\hat{\rho} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i} \otimes \delta_{y_i}$. We redefine the natural homomorphism of $\mathcal{G}$ into ℓ^2 with

$$\begin{array}{rccc} \hat{S}: & \mathcal{G} & \to & \ell^2 \\ & \gamma & \to & \left(k_{x_i}^\star \gamma\right)_{i\leq n}. \end{array}$$

We check that its adjoint is, for $\Xi \in \mathcal{H}^n$ and $\gamma \in \mathcal{G}$

$$\left\langle \hat{S}^\star \Xi, \gamma \right\rangle_\mathcal{G} = \left\langle \Xi, \hat{S}\gamma \right\rangle_{\ell^2} = \frac{1}{n}\sum_{i=1}^n \left\langle \xi_i, k_{x_i}^\star \gamma \right\rangle_\mathcal{H} = \left\langle \frac{1}{n}\sum_{i=1}^n k_{x_i}\xi_i, \gamma \right\rangle_\mathcal{G}.$$

Similarly, we define $\hat{K} : \mathcal{H}^n \to \mathcal{H}^n$ and $\hat{\Sigma} : \mathcal{G} \to \mathcal{G}$, with

$$\hat{K}\Xi = \hat{S}\hat{S}^\star\Xi = \left(\frac{1}{n}\sum_{i=1}^n k(x_j, x_i)\xi_i\right)_{j\leq n}, \qquad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n k_{x_i} \otimes k_{x_i} = \mathbb{E}_{\hat{\rho}_\mathcal{X}}[k_X k_X^\star].$$

Finally, we define $\hat{\Phi} = (\varphi(y)_i)_{i\leq n} \in \mathcal{H}^n$, so that

$$\widehat{S^\star g^*} := \mathbb{E}_{\hat{\rho}}[\varphi(Y)\cdot k_X] = \hat{S}^\star\hat{\Phi}.$$

Finally, we can express g_n as

$$g_n = S(\hat{\Sigma}+\lambda)^{-1}\hat{S}^\star\hat{\Phi} = S(\hat{S}^\star\hat{S}+\lambda)^{-1}\hat{S}^\star\hat{\Phi} = S\hat{S}^\star(\hat{S}\hat{S}^\star+\lambda)^{-1}\hat{\Phi} = S\hat{S}^\star(\hat{K}+\lambda)^{-1}\hat{\Phi}.$$

This explains the equivalence between g_n defined at the beginning of Section 5.5 and the g_n expressed in the lemma, that will be used for derivations of theorems. □

5.C.3 Linear algebra and equivalent assumptions to Assumptions 7, 8

To proceed with the study of the convergence of g_n toward g_λ in L^2, it is helpful to pass by $\mathcal{G}$. To do so, we need to express Assumptions 7 and 8 in $\mathcal{G}$, which we can do using the following linear algebra property.

Lemma 21 (Linear algebra on compact operators)**.** *There exist $(u_i)_{i\in\mathbb{N}}$ an orthogonal basis of $\mathcal{G}_\mathcal{X}$, $(v_i)_{i\in\mathbb{N}}$ an orthogonal basis of $L^2(\mathcal{X}, \mathbb{R}, \rho_\mathcal{X})$, and $(\lambda_i)_{i\in\mathbb{N}}$ a decreasing sequence of positive real number such that*

$$S = \sum_{i\in\mathbb{N}} \lambda_i^{1/2} u_i v_i^\star, \qquad S^\star = \sum_{i\in\mathbb{N}} \lambda_i^{1/2} v_i u_i^\star, \qquad \Sigma = \sum_{i\in\mathbb{N}} \lambda_i u_i u_i^\star, \qquad K = \sum_{i\in\mathbb{N}} \lambda_i v_i v_i^\star, \tag{5.14}$$

where the convergence of series has to be understood with the operator norms. Moreover, we have that, if the kernel k is bounded by κ^2,

$$\sum_{i\in\mathbb{N}} \lambda_i \leq \kappa^2 < +\infty.$$

Therefore, K and Σ are trace-class, and S and $S^\star$ are Hilbert-Schmidt.

Proof. Notice that $\Sigma = \mathbb{E}_X[k_X \otimes k_X]$ and that $\|k_x \otimes k_x\|_{\mathrm{op}(\mathcal{G}_\mathcal{X})} = \|k_x\|_{\mathcal{G}_\mathcal{X}} = k(x,x) \leq \kappa^2$. Therefore, Σ is a nuclear operator, so it is trace class and so it is compact.

The first point results from diagonalization of kernel operators, known as Mercer's Theorem (Mercer, 1909; Steinwart and Scovel, 2012). Σ is a compact operator, therefore, the Spectral Theorem gives the existence of a sequence $(\lambda_i) \in \mathbb{R}^\mathbb{N}$ and an orthonormal basis $(u_i) \in \mathcal{G}_\mathcal{X}^\mathbb{N}$ of $\mathcal{G}_\mathcal{X}$ such that

$$\Sigma = \sum_{i\in\mathbb{N}} \lambda_i u_i u_i^\star,$$

where the convergence has to be understood with the operator norm. Because Σ is of the form $S^\star S$, one can consider (λ_i) a decreasing sequence of positive eigenvalues. Then, by defining, for all $i \in \mathbb{N}$ with $\lambda_i > 0$,

$$v_i = \lambda_i^{-1/2} S u_i$$

we check that (v_i) are orthonormal, and we complete them to form an orthonormal basis of $(L^2(\mathcal{X}, \mathbb{R}, \rho_\mathcal{X}))$. Finally, we check that

$$S = \sum_{i\in\mathbb{N}} \lambda_i^{1/2} v_i u_i^\star,$$

and that the other equalities hold too.

To check the second assertion, we use that $k_x k_x^\star$ is rank one when operating on $\mathcal{G}_\mathcal{X}$ and therefore

$$\begin{aligned}\mathrm{Tr}\,\Sigma = \mathrm{Tr}\left(\mathbb{E}_X[k_X k_X^\star]\right) = \mathbb{E}_X\left[\mathrm{Tr}\left(k_X k_X^\star\right)\right] = \mathbb{E}_X\left[\left\|k_X k_X^\star\right\|_{\mathrm{op}(\mathcal{G}_\mathcal{X})}\right] \\ = \mathbb{E}_X\left[\|k_X\|_{\mathcal{G}_\mathcal{X}}\right] = \mathbb{E}_X[k(x,x)] \leq \kappa^2.\end{aligned}$$

This shows that S and $S^\star$ are Hilbert-Schmidt operators and that K is also trace class. □

This allows us to cast in $\mathcal{G}_\mathcal{X}$ the assumptions expressed in L^2.

Lemma 22 (Equivalence of capacity condition)**.** *For $\sigma \in (0, 1]$, it is equivalent to suppose that*

- $\mathrm{Tr}_{L^2(\mathcal{X},\mathcal{H},\rho_\mathcal{X})}(K^\sigma) < +\infty$.
- $\mathrm{Tr}_{\mathcal{G}_\mathcal{X}}(\Sigma^\sigma) < +\infty$.
- $\sum_{i\in\mathbb{N}} \lambda_i^\sigma < +\infty$.

In Assumption 7, the smaller σ, the faster the λ_i decreases, the easier it will be to approximate Σ based on approximation of ρ. This appears explicitly in Theorem 8. Indeed, for $\sigma = 0$, the condition should be defined as Σ of finite rank. Note that when k is bounded, we know that Σ is trace class, and therefore, Assumption 7 holds with $\sigma = 1$.

Lemma 23 (Interpolation inequality in RKHS)**.** *Assumption 8 implies that*

$$\forall\, \gamma \in \mathcal{G}, \qquad \|S\gamma\|_{L^\infty} \le c_p \left\|\Sigma^{\frac{1}{2}-p}\gamma\right\|_\mathcal{G}. \tag{5.15}$$

Proof. We begin by showing the property for $\gamma \in \mathcal{G}_\mathcal{X}$. When $\gamma = \sum_{i\in\mathbb{N}} c_i v_i$ with $\sum_{i\in\mathbb{N}} c_i^2 < +\infty$, denote $g = \sum_{i\in\mathbb{N}} \lambda_i^{\frac{1}{2}-p} c_i u_i$, we have $g \in L^2$, therefore, using Assumption 8,

$$\|S\gamma\|_{L^\infty} = \|K^p g\|_{L^\infty} \le c_p \|g\|_{L^2} = c_p \left\|\Sigma^{\frac{1}{2}-p}\gamma\right\|_{\mathcal{G}_\mathcal{X}}.$$

This ends the proof for $\mathcal{G}_\mathcal{X}$. Note also that when the result of the Lemma holds, then Assumption 8 holds for any $g \in \mathrm{im}_{L^2(\mathcal{X},\mathbb{R},\rho_\mathcal{X})} K^{\frac{1}{2}-p}$.

Let us switch to $\mathcal{G}$ now. Let $\gamma \in \mathcal{G}$, and denote $g = S\gamma$. Suppose that g achieve it maximum in x_∞, define the direction $\xi = {}^{g(x_\infty)}/_{\|g(x_\infty)\|_\mathcal{H}}$, and define $g_\xi : x \to \langle g(x), \xi\rangle_\mathcal{H} \in L^2(\mathcal{X}, \mathbb{R}, \rho_\mathcal{X})$, and $\gamma_\xi = \sum_{j\in\mathbb{N}} \langle g_\xi, v_i\rangle_{L^2} u_i \in \mathcal{G}_\mathcal{X}$. We have

$$\|S\gamma\|_{L^\infty} = \left\|S\gamma_\xi\right\|_{L^\infty} \le c_p \left\|\Sigma^{\frac{1}{2}-p}\gamma_\xi\right\|_{\mathcal{G}_\mathcal{X}} \le c_p \left\|\Sigma^{\frac{1}{2}-p}\gamma\right\|_\mathcal{G}.$$

When g does not achieve its maximum, one can do a similar reasoning by considering a basis $(f_i)_{i\in\mathbb{N}}$ of $\mathcal{H}$ and decomposition γ on the basis $(u_i f_j)_{i,j\in\mathbb{N}}$, before summing the directions. □

In Assumption 8, the bigger $1/2 - p$ the more we are able to control our problem in $\mathcal{G}$, the better. Note that this reformulation of the interpolation inequality allows generalizing it for p smaller than zero. Note that when k is bounded, $\|(S\gamma)(x)\|_\mathcal{H} = \left\|k_X^\star \gamma\right\|_\mathcal{H} \le \|k_X\|_{\mathrm{op}} \|\gamma\|_\mathcal{G} = \sqrt{k(x,x)}\, \|\gamma\|_\mathcal{G}$, hence Assumption 8 holds with $p = 1/2$.

5.C.4 Linear algebra with atoms k_x and useful inequalities

From the study of the convergence of g_n to g_λ will emerge two quantities linked to eigenvalues of Σ and the position of k_x regarding eigenspaces, that are

$$\mathcal{N}(\lambda) = \mathrm{Tr}\left((\Sigma + \lambda)^{-1}\Sigma\right), \qquad \mathcal{N}_\infty(\lambda) = \sup_{x\in\mathrm{supp}\,\rho_\mathcal{X}} \left\|(\Sigma + \lambda)^{-\frac{1}{2}} k_x\right\|_{\mathrm{op}}. \tag{5.16}$$

While those quantities could be bounded with brute force consideration, Assumptions 7 and 8 will help to control them more subtly.

Proposition 24 (Characterization of capacity condition)**.** *The property $\sum_{i\in\mathbb{N}} \lambda_i^\sigma < +\infty$, can be rephrased in terms of eigenvalues of Σ as the existence of $a_1 > 0$ such that, for all $i > 0$,*

$$\lambda_i \le a_1 (i+1)^{-\frac{1}{\sigma}}. \tag{5.17}$$

Proof. Denote by u_i and S_n the respective quantities λ_i^σ and $\sum_{i=1}^n u_i$. Because S_n converge, it is a Cauchy sequence, so there exists N such that for any $p > q > N$, $S_p - S_q = \sum_{i=q+1}^p u_i \le 1$. In particular, considering $p = 2q$, and because (λ_i) is decreasing, we have $q u_{2q} \le \sum_{i=q+1}^{2q} u_i \le 1$. Therefore, we have that for all $i > 2N$, $u_i \le 3(i+1)^{-1}$, considering $(a_1)^\sigma = 3 + \max_{i\le 2N}\{(i+1)u_i\}$, we get the desired result. □

Proposition 25 (Characterization of $\mathcal{N}$)**.** *When* $\operatorname{Tr}(K^\sigma) < +\infty$*, with* $a_2 = \int_0^\infty \frac{a_1}{a_1 + t^{\frac{1}{\sigma}}}\,\mathrm{d}t$*,*

$$\forall\, \lambda > 0, \qquad \mathcal{N}(\lambda, r) \leq a_2 \lambda^{-\sigma}. \tag{5.18}$$

Proof. Expressed with eigenvalues, we have

$$\mathcal{N}(\lambda) = \operatorname{Tr}\left((\Sigma + \lambda)^{-1}\Sigma\right) = \sum_{i\in\mathbb{N}} \frac{\lambda_i}{\lambda_i + \lambda}.$$

Using that $\lambda_i \leq a_1(i+1)^{-\frac{1}{\sigma}}$, that $x \to \frac{x}{x+a}$ is increasing with respect to x for any $a > 0$ and the series-integral comparison, we get for $\sigma \in (0, 1]$

$$\begin{aligned}\mathcal{N}(\lambda) &\leq \sum_{i\in\mathbb{N}} \frac{a_1(i+1)^{-\frac{1}{\sigma}}}{a_1(i+1)^{-\sigma} + \lambda} \leq \int_0^\infty \frac{a_1 t^{-\frac{1}{\sigma}}}{a_1 t^{-\frac{1}{\sigma}} + \lambda}\,\mathrm{d}t = \int_0^\infty \frac{a_1}{a_1 + \lambda t^{\frac{1}{\sigma}}}\,\mathrm{d}t \\ &= \lambda^{-\sigma} \int_0^\infty \frac{a_1}{a_1 + (\lambda^\sigma t)^{\frac{1}{\sigma}}}\,\mathrm{d}(\lambda^\sigma t) = a_2 \lambda^{-\sigma},\end{aligned}$$

where we check the convergence of the integral. □

Indeed, Assumption 8 has a profound linear algebra meaning, it is a condition on $\rho_{\mathcal{X}}$-almost all the vector $k_x \in \mathcal{G}_{\mathcal{X}}$ not to be excessively supported on the eigenvector corresponding to small eigenvalue of Σ.

Proposition 26 (Characterization of interpolation condition)**.** *The interpolation Assumption 8 implies that, for all* $i \in \mathbb{N}$

$$\sup_{x\in\rho_{\mathcal{X}}} \left|\langle k_x, u_i\rangle_{\mathcal{G}_{\mathcal{X}}}\right| \leq c_p \lambda_i^{\frac{1}{2}-p}. \tag{5.19}$$

Proof. Consider the decomposition of $k_x \in \mathcal{G}_{\mathcal{X}}$ according to the eigenvectors of Σ, with $a_i(x) = \langle k_x, u_i\rangle$. The interpolation condition Assumption 8, expressed in $\mathcal{G}_{\mathcal{X}}$ with Lemma 23, leads to for any $\gamma_{\mathcal{X}} \in \mathcal{G}_{\mathcal{X}}$, and $S\gamma_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}$,

$$|(S\gamma_{\mathcal{X}})(x)| = \left|\langle k_x, \gamma_{\mathcal{X}}\rangle_{\mathcal{G}_{\mathcal{X}}}\right| \leq \|S\gamma_{\mathcal{X}}\|_{L^\infty} \leq c_p \left\|\Sigma^{\frac{1}{2}-p}\gamma_{\mathcal{X}}\right\|_{\mathcal{G}_{\mathcal{X}}}$$

This implies that

$$\langle k_x, \gamma_{\mathcal{X}}\rangle^2_{\mathcal{G}_{\mathcal{X}}} = \left(\sum_{i\in\mathbb{N}} \langle k_x, u_i\rangle \langle \gamma_{\mathcal{X}}, u_i\rangle\right)^2 \leq c_p^2 \left\|\Sigma^{\frac{1}{2}-p}\gamma_{\mathcal{X}}\right\|^2_{\mathcal{G}_{\mathcal{X}}} = c_p^2 \sum_{i\in\mathbb{N}} \lambda_i^{1-2p} \langle \gamma_{\mathcal{X}}, u_i\rangle^2 .$$

Taking $\gamma_{\mathcal{X}} = u_i$, we get that

$$|\langle k_x, u_i\rangle| \leq c_p \lambda_i^{\frac{1}{2}-p}.$$

This result relates the interpolation condition to the fact that k_x is not excessively supported on the eigenvectors corresponding to vanishing eigenvalues of Σ. □

Proposition 27 (Characterization of $\mathcal{N}_\infty(\lambda, r)$)**.** *Under the interpolation condition, Assumption 8, we have with* $a_3 = c_p(2p)^{-p}(1-2p)^{\frac{1}{2}-p}$*, or* $a_3 = c_p$ *when* $p = 1/2$*,*

$$\mathcal{N}_\infty(\lambda) \leq a_3 \lambda^{-p}. \tag{5.20}$$

Proof. First of all, notice that

$$\begin{aligned}\left\|(\Sigma+\lambda)^{-\frac{1}{2}}k_x\right\|_{\mathcal{G}_{\mathcal{X}}} &= \sup_{\|\gamma_{\mathcal{X}}\|_{\mathcal{X}}=1} \left\langle \gamma_{\mathcal{X}}, (\Sigma+\lambda)^{-\frac{1}{2}}k_x\right\rangle_{\mathcal{G}_{\mathcal{X}}} = \sup_{c;\sum_{i\in\mathbb{N}} c_i^2 = 1} \sum_{i\in\mathbb{N}} \frac{c_i \langle k_x, u_i\rangle}{(\lambda+\lambda_i)^{\frac{1}{2}}} \\ &\leq c_p \sup_{c;\sum_{i\in\mathbb{N}} c_i^2=1} \sum_{i\in\mathbb{N}} \frac{c_i \lambda_i^{\frac{1}{2}-p}}{(\lambda+\lambda_i)^{\frac{1}{2}}} \leq \sup_{t\in\mathbb{R}_+} c_p \frac{t^{\frac{1}{2}-p}}{(\lambda+t)^{\frac{1}{2}}}.\end{aligned}$$

When $p \in (0, 1/2)$, this last function is zero in zero and in infinity, therefore its maximum t_0 verifies, taking the derivative of its logarithm,

$$\frac{1/2 - p}{t_0} = \frac{1}{2(t_0 + \lambda)} \quad \Rightarrow \quad t_0 = \frac{(1-2p)\lambda}{2p} \quad \Rightarrow \quad \sup_{t \in \mathbb{R}_+} \frac{t^{\frac{1}{2}-p}}{(\lambda + t)^{\frac{1}{2}}} = (2p)^{-p}(1-2p)^{\frac{1}{2}-p}\lambda^{-p}.$$

The cases $p \in \{0, 1\}$ are easy to treat. □

In the previous analysis, one fact does not appear, it is that Σ and k_x are linked to one another, since $\Sigma = \mathbb{E}_X[k_X k_X^\star]$. The following remark builds on it to relate $\mathcal{N}$ and $\mathcal{N}_\infty$.

Remark 28 (Relation between interpolation and capacity condition)**.** *The capacity and interpolation condition are related by the fact that it is unreasonable not to consider that $p \leq \sigma/2$.*

Proof. Because $k_x k_x^\star$ is of rank one in $\mathcal{G}_\mathcal{X}$, we have

$$\begin{aligned}\mathcal{N}(\lambda) &= \operatorname{Tr}\left((\Sigma + \lambda)^{-1}\Sigma\right) = \mathbb{E}_X\left[\operatorname{Tr}\left((\Sigma + \lambda)^{-1} k_X k_X^\star\right)\right] = \mathbb{E}_X\left[\operatorname{Tr}\left(k_X^\star(\Sigma + \lambda)^{-1} k_X\right)\right] \\ &= \mathbb{E}_X\left[\left\|k_X^\star(\Sigma + \lambda)^{-1} k_X\right\|_{\mathrm{op}}\right] = \mathbb{E}_X\left[\left\|(\Sigma + \lambda)^{-\frac{1}{2}} k_X\right\|_{\mathcal{G}_\mathcal{X}}^2\right].\end{aligned}$$

So indeed, $\mathcal{N}(\lambda)$ is the expectation of the square $\left\|(\Sigma + \lambda)^{-\frac{1}{2}} k_X\right\|_{\mathcal{G}_\mathcal{X}}$, when $\mathcal{N}_\infty(\lambda)$ is the supremum of this last quantity. Therefore,

$$\mathcal{N}(\lambda) \leq \mathcal{N}_\infty(\lambda)^2$$

Supposing that the dependencies in λ proved above are tight, we should have $\sigma \geq 2p$, which is the statement of this remark. We refer the reader to Lemma 6.2 of Fischer and Steinwart (2020) for more consideration to relates σ and p (reading p and $\alpha/2$ with their notations) □

5.C.5 Geometrical control of the residual $\|g_\lambda - g^*\|_{L^\infty}$

The proof of the first assertion in Lemma 14 follows from, using Assumption 9, with $g_0 \in K^{-q} g^*$,

$$\begin{aligned}g_\lambda - g^* &= (K(K + \lambda)^{-1} - I)g^* = -\lambda(K + \lambda)^{-1} g^* = -\lambda(K + \lambda)^{-1} K^q g_0 \\ &= -\lambda K^p (K + \lambda)^{-1} K^{q-p} g_0.\end{aligned}$$

Then using Assumption 8,

$$\begin{aligned}\|g^* - g_\lambda\|_\infty &\leq c_p \lambda \left\|K^{q-p}(K + \lambda)^{-1}\right\|_{\mathrm{op}} \|g_0\|_{L^2} \\ &\leq c_p \lambda \left\|K(K + \lambda)^{-1}\right\|_{\mathrm{op}}^{q-p} \left\|(K + \lambda)^{-1}\right\|_{\mathrm{op}}^{1+p-q} \|g_0\|_{L^2} \\ &\leq c_p \lambda 1^{q-p} \lambda^{-(1+p-q)} \|g_0\|_{L^2} = b_1 \lambda^{q-p},\end{aligned}$$

where we have used that $\left\|K(K + \lambda)^{-1}\right\|_{\mathrm{op}} = \|K\|_{\mathrm{op}} / (\|K\|_{\mathrm{op}} + \lambda) \leq 1$ and that $\left\|(K + \lambda)^{-1}\right\| \leq \lambda^{-1}$.

5.C.6 Convergence of $\|g_n - g_\lambda\|$ through concentration inequality

For the proof of the second assertion in Lemma 14, we will put ourselves in $\mathcal{G}$. For this, we define in $\mathcal{G}$

$$\gamma = \mathbb{E}_\rho[k_X \varphi(Y)], \qquad \gamma_\lambda = (\Sigma + \lambda)^{-1}\gamma, \qquad \hat{\gamma} = \mathbb{E}_{\hat{\rho}}[k_X \varphi(Y)], \tag{5.21}$$

so that $g_\lambda = S\gamma_\lambda$, and $g_n = S(\hat{\Sigma} + \lambda)^{-1}\hat{\gamma}$.

Decomposition into a matrix and a vector term

We begin by expressing $g_n - g_\lambda$ in $\mathcal{G}$ with

$$\begin{aligned} g_n - g_\lambda &= S\left((\hat{\Sigma}+\lambda)^{-1}\hat{\gamma} - (\Sigma+\lambda)^{-1}\gamma\right) \\ &= S\left((\hat{\Sigma}+\lambda)^{-1}(\hat{\gamma}-\gamma) + ((\hat{\Sigma}+\lambda)^{-1} - (\Sigma+\lambda)^{-1})\gamma)\right) \\ &= S\left((\hat{\Sigma}+\lambda)^{-1}(\hat{\gamma}-\gamma) + (\hat{\Sigma}+\lambda)^{-1}(\Sigma-\hat{\Sigma})(\Sigma+\lambda)^{-1}\gamma)\right) \\ &= S\left((\hat{\Sigma}+\lambda)^{-1}((\hat{\gamma}-\hat{\Sigma}\gamma_\lambda) - (\gamma - \Sigma\gamma_\lambda))\right), \end{aligned}$$

where we have used that $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$. Therefore, using the expression, Lemma 23, of Assumption 8 in $\mathcal{G}$, we get

$$\begin{aligned} \|g_n - g_\lambda\|_{L^\infty} \leq c_p \left\|\Sigma^{\frac{1}{2}-p}(\hat{\Sigma}+\lambda)^{-1}(\Sigma+\lambda)^{\frac{1}{2}+p}\right\|_{\text{op}} \times \cdots \\ \left\|(\Sigma+\lambda)^{-(\frac{1}{2}+p)}((\hat{\gamma}-\hat{\Sigma}\gamma_\lambda) - (\gamma-\Sigma\gamma_\lambda))\right\|_{\mathcal{G}}. \end{aligned}$$

On the one hand, we have concentration of matrix term toward $\Sigma^{\frac{1}{2}-p}(\Sigma+\lambda)^{-\left(\frac{1}{2}-p\right)} \preceq I$. On the other hand, we have concentration of the vector $\hat{\gamma} - \hat{\Sigma}\gamma_\lambda$ toward $\gamma - \Sigma\gamma_\lambda$. Indeed, the concentration of the matrix term is hard to prove (it is only a conjecture), therefore we will go for another decomposition, that will result in similar rates when $p \geq 0$, that is

$$\begin{aligned} \|g_n - g_\lambda\|_{L^\infty} &\leq c_p \left\|\Sigma^{\frac{1}{2}-p}(\Sigma+\lambda)^{-\frac{1}{2}}\right\|_{\text{op}} \mathcal{A}(\lambda)\mathcal{B}(\lambda) \\ \mathcal{A}(\lambda) &= \left\|(\Sigma+\lambda)^{\frac{1}{2}}(\hat{\Sigma}+\lambda)^{-1}(\Sigma+\lambda)^{\frac{1}{2}}\right\|_{\text{op}}, \\ \mathcal{B}(\lambda) &= \left\|(\Sigma+\lambda)^{-\frac{1}{2}}((\hat{\gamma}-\hat{\Sigma}\gamma_\lambda) - (\gamma-\Sigma\gamma_\lambda))\right\|_{\mathcal{G}}. \end{aligned} \tag{5.22}$$

Recall the definition of the following important quantity that are going to pop up from the analysis

$$\mathcal{N}(\lambda) = \text{Tr}\left((\Sigma+\lambda)^{-1}\Sigma\right), \qquad \mathcal{N}_\infty(\lambda) = \sup_{x \in \text{supp}\,\rho_{\mathcal{X}}} \left\|(\Sigma+\lambda)^{-\frac{1}{2}}k_x\right\|_{\text{op}}. \tag{5.16}$$

Extra matrix term

We control the extra matrix term with

$$\left\|\Sigma^{\frac{1}{2}-p}(\Sigma+\lambda)^{-\frac{1}{2}}\right\|_{\text{op}} = \left\|\Sigma^{\frac{1}{2}-p}(\Sigma+\lambda)^{-\left(\frac{1}{2}-p\right)}\right\|_{\text{op}} \|(\Sigma+\lambda)^{-p}\|_{\text{op}} \leq \lambda^{-p}.$$

Using that $\left\|(\Sigma+\lambda)^{-1}\right\|_{\text{op}} \leq \lambda^{-1}$ and that $\left\|(\Sigma+\lambda)^{-1}\Sigma\right\|_{\text{op}} \leq {}^{\|\Sigma\|_{\text{op}}}/_{(\|\Sigma\|_{\text{op}}+\lambda)} \leq 1$.

Matrix concentration

Let us make explicit the concentration in the matrix term with

$$\begin{aligned} (\Sigma+\lambda)^{\frac{1}{2}}(\hat{\Sigma}+\lambda)^{-1}(\Sigma+\lambda)^{\frac{1}{2}} &= I + (\Sigma+\lambda)^{\frac{1}{2}}\left((\hat{\Sigma}+\lambda)^{-1} - (\Sigma+\lambda)^{-1}\right)(\Sigma+\lambda)^{\frac{1}{2}} \\ &= I + (\Sigma+\lambda)^{\frac{1}{2}}(\hat{\Sigma}+\lambda)^{-1}\left(\Sigma - \hat{\Sigma}\right)(\Sigma+\lambda)^{-1}(\Sigma+\lambda)^{\frac{1}{2}}. \end{aligned}$$

From here, notice the following implications (that are actually equivalence)

$$\begin{aligned} \Sigma - \hat{\Sigma} \preceq t(\Sigma+\lambda) \quad &\Rightarrow \quad \hat{\Sigma}+\lambda \succeq (1-t)(\Sigma+\lambda) \\ &\Rightarrow \quad (\hat{\Sigma}+\lambda)^{-1} \preceq (1-t)^{-1}(\Sigma+\lambda)^{-1}. \\ &\Rightarrow \quad (\hat{\Sigma}+\lambda)^{-1} - (\Sigma+\lambda)^{-1} \preceq t(1-t)^{-1}(\Sigma+\lambda)^{-1}. \\ &\Rightarrow \quad (\Sigma+\lambda)^{\frac{1}{2}}\left((\hat{\Sigma}+\lambda)^{-1} - (\Sigma+\lambda)^{-1}\right)(\Sigma+\lambda)^{\frac{1}{2}} \preceq t(1-t)^{-1} \\ &\Rightarrow \quad (\Sigma+\lambda)^{\frac{1}{2}}(\hat{\Sigma}+\lambda)^{-1}(\Sigma+\lambda)^{\frac{1}{2}} \preceq (1-t)^{-1}. \end{aligned}$$

The probability of the event $\Sigma - \hat{\Sigma} \preceq t(\Sigma + \lambda)$, can be studied through the probability of the event $(\Sigma+\lambda)^{-\frac{1}{2}}(\Sigma-\hat{\Sigma})(\Sigma+\lambda)^{-\frac{1}{2}} \preceq t$, which can be studied through concentration of self adjoint operators. Finally, we have shown that

$$\left\|(\Sigma + \lambda)^{-\frac{1}{2}}(\Sigma - \hat{\Sigma})(\Sigma + \lambda)^{-\frac{1}{2}}\right\|_{\text{op}} \leq t \quad \Rightarrow \quad \mathcal{A}(\lambda) \leq \frac{1}{1-t}. \tag{5.23}$$

The best result that we are aware of, for covariance matrix inequality, is the extension to self-adjoint Hilbert-Schmidt operators provided by Minsker (2017) in Section 3.2 of its concentration inequality on random matrices Theorem 3.1. It can be formulated as the following.

Theorem 8 (Concentration of self-adjoint operators (Minsker, 2017))**.** *Let denote by $(\xi_i)_{i\leq n}$ a sequence of independent self-adjoint operator acting on a separable Hilbert space $\mathcal{A}$, such that* $\ker(\mathbb{E}[\xi_i]) = \mathcal{A}$*, that are bounded by a constant $M \in \mathbb{R}$, in the sense $\|\xi_i\|_{\text{op}} \leq M$, with finite variance $\sigma^2 = \left\|\mathbb{E}\sum_{i=1}^n \xi_i^2\right\|_{\text{op}}$. For any $t > 0$ such that $6t^2 \geq (\sigma^2 + {}^{Mt}/_3)$,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \xi_i\right\|_{\text{op}} > t\right) \leq 14\, r\left(\sum_{i=1}^n \mathbb{E}\,\xi_i^2\right) \exp\left(-\frac{t^2}{2\sigma^2 + 2tM/3}\right),$$

with $r(\xi) = {}^{\text{Tr}\,\xi}/_{\|\xi\|_{\text{op}}}$.

Let us define ξ that goes from $\mathcal{X}$ to the space of self-adjoint operator action on $\mathcal{G}_\mathcal{X}$ as

$$\xi(x) = (\Sigma + \lambda)^{-\frac{1}{2}} k_x k_x^\star (\Sigma + \lambda)^{-\frac{1}{2}}. \tag{5.24}$$

We have that $(\Sigma + \lambda)^{-\frac{1}{2}}(\Sigma - \hat{\Sigma})(\Sigma + \lambda)^{-\frac{1}{2}} = \mathbb{E}_\rho[\xi(X)] - \frac{1}{n}\sum_{i=1}^n \xi(x_i)$. To apply operator concentration, we need to bound ξ and its variance.

Bound on ξ. To bound ξ we proceed with, because $k_x k_x^\star$ is of rank one,

$$\begin{aligned}\|\xi(x)\|_{\text{op}} &= \left\|(\Sigma + \lambda)^{-\frac{1}{2}} k_x k_x^\star (\Sigma + \lambda)^{-\frac{1}{2}}\right\|_{\text{op}} = \text{Tr}\left((\Sigma + \lambda)^{-\frac{1}{2}} k_x k_x^\star (\Sigma + \lambda)^{-\frac{1}{2}}\right) \\ &= \text{Tr}\left(k_x^\star (\Sigma + \lambda)^{-1} k_x\right) = \left\|(\Sigma + \lambda)^{-\frac{1}{2}} k_x\right\|_{\mathcal{G}_\mathcal{X}}^2 \leq \mathcal{N}_\infty(\lambda)^2.\end{aligned}$$

Variance of ξ. For the variance of ξ we proceed by noticing that

$$\begin{aligned}\mathbb{E}\,\xi(X) &= \mathbb{E}_X (\Sigma + \lambda)^{-\frac{1}{2}} k_X k_X^\star (\Sigma + \lambda)^{-\frac{1}{2}} = (\Sigma + \lambda)^{-\frac{1}{2}}\, \mathbb{E}_X\left[k_X k_X^\star\right](\Sigma + \lambda)^{-\frac{1}{2}} \\ &= (\Sigma + \lambda)^{-\frac{1}{2}} \Sigma (\Sigma + \lambda)^{-\frac{1}{2}} = (\Sigma + \lambda)^{-1}\Sigma.\end{aligned}$$

Hence,

$$\mathbb{E}\,\xi(X)^2 \leq \sup_{x\in\mathcal{X}} \|\xi(x)\|_{\text{op}}\, \mathbb{E}[\xi(X)] \leq \mathcal{N}_\infty(\lambda)^2 (\Sigma + \lambda)^{-1}\Sigma.$$

And as a consequence

$$\left\|\mathbb{E}\,\xi(x)^2\right\| \leq \mathcal{N}_\infty(\lambda)^2,$$

where we have used that $\left\|(\Sigma + \lambda)^{-1}\Sigma\right\|_{\text{op}} = {}^{\|\Sigma\|_{\text{op}}}/_{(\|\Sigma\|_{\text{op}} + \lambda)} \leq 1$.

Concentration bound on ξ. Using the self-adjoint concentration theorem, we get for any $t > 0$, such that $6nt^2 \geq \mathcal{N}_\infty(\lambda)^2(1 + {}^t/_3)$,

$$\mathbb{P}_{\mathcal{D}_n}\left(\left\|\mathbb{E}_{\hat{\rho}}[\xi] - \mathbb{E}_\rho[\xi]\right\|_{\text{op}} > t\right) \leq 14\, \frac{\|\Sigma\|_{\text{op}} + \lambda}{\|\Sigma\|_{\text{op}}} \mathcal{N}(\lambda) \exp\left(-\frac{nt^2}{2\mathcal{N}_\infty(\lambda)^2(1 + t/3)}\right).$$

Therefore, using the contraposition of the prior implication, we get

$$\mathbb{P}_{\mathcal{D}_n}\left(\mathcal{A}(\lambda) > \frac{1}{1-t}\right) \leq 14\, \frac{\|\Sigma\|_{\text{op}} + \lambda}{\|\Sigma\|_{\text{op}}} \mathcal{N}(\lambda) \exp\left(-\frac{nt^2}{2\mathcal{N}_\infty(\lambda)^2(1 + t/3)}\right). \tag{5.25}$$

Decomposition of vector term in a variance and a bias term

Let switch to the vector term, consider $\xi : \mathcal{X} \times \mathcal{Y} \to \mathcal{G}$, defined as

$$\xi = (\Sigma + \lambda)^{-\frac{1}{2}} k_x(\varphi(y) - k_x^\star \gamma_\lambda).$$

It allows expressing in simple form the vector term as

$$\mathcal{B}(\lambda) = \left\| \frac{1}{n} \sum_{i=1}^{n} \xi(X_i, Y_i) - \mathbb{E}_{(X,Y)\sim\rho}[\xi(X,Y)] \right\|.$$

We can study this term through concentration inequality in $\mathcal{G}$. To proceed we will dissociate the variability due to Y to the one due to X, recalling that $g_\lambda(x) = k_x^\star \gamma_\lambda$ and going for the following decomposition

$$\begin{aligned} \xi(x,y) &= \xi_v(x,y) + \xi_b(x) \\ \xi_v(x,y) &= (\Sigma + \lambda)^{-\frac{1}{2}} k_x(\varphi(y) - g^*(x)), \\ \xi_b(x) &= (\Sigma + \lambda)^{-\frac{1}{2}} k_x(g^*(x) - g_\lambda(x)), \end{aligned} \tag{5.26}$$

which corresponds to the decomposition

$$\begin{aligned} \mathcal{B}(\lambda) &\leq \mathcal{B}_v(\lambda) + \mathcal{B}_b(\lambda) \\ \mathcal{B}_v(\lambda) &= \left\| \mathbb{E}_{\hat{\rho}}[\xi_v(X,Y)] - \mathbb{E}_\rho[\xi_v(X,Y)] \right\| \\ \mathcal{B}_b(\lambda) &= \left\| \mathbb{E}_{\hat{\rho}}[\xi_b(X,Y)] - \mathbb{E}_\rho[\xi_b(X,Y)] \right\|. \end{aligned} \tag{5.27}$$

The first term is due to the error because of having observed $\varphi(y)$ rather than $g^*(x)$, often called "variance", and the second term is due to the aiming for g_λ instead of g^* often called "bias".

Control of the variance

To control the variance term, we will use the Bernstein inequality stated Theorem 7.

Bound on the moment of ξ_v. First of all notice that

$$\|\xi_v(x,y)\|_{\mathcal{G}} \leq \left\| (\Sigma + \lambda)^{-\frac{1}{2}} k_x \right\|_{\text{op}} \|\varphi(y) - g^*(x)\|_{\mathcal{H}}.$$

Therefore, under Assumption 5, for $m \geq 2$:

$$\begin{aligned} \mathbb{E}_{(X,Y)\sim\rho}\left[\|\xi_v(X,Y)\|^m\right] &\leq \mathbb{E}_{X\sim\rho_{\mathcal{X}}}\left[\left\| (\Sigma + \lambda)^{-\frac{1}{2}} k_x \right\|_{\text{op}}^m \mathbb{E}_{Y\sim\rho|_X}\left[\|\varphi(y) - g^*(x)\|_{\mathcal{H}}^m \right] \right] \\ &\leq \frac{1}{2} m! \sigma^2 M^{m-2} \, \mathbb{E}_{X\sim\rho_{\mathcal{X}}}\left[\left\| (\Sigma + \lambda)^{-\frac{1}{2}} k_x \right\|_{\text{op}}^m \right]. \end{aligned}$$

We bound the last term with

$$\begin{aligned} \mathbb{E}_{X\sim\rho_{\mathcal{X}}}\left[\left\| (\Sigma + \lambda)^{-\frac{1}{2}} k_x \right\|_{\text{op}}^m \right] &\leq \sup_{x \in \operatorname{supp}\rho_{\mathcal{X}}} \left\| (\Sigma + \lambda)^{-\frac{1}{2}} k_x \right\|_{\text{op}}^{m-2} \mathbb{E}_{X\sim\rho_{\mathcal{X}}}\left[\left\| (\Sigma + \lambda)^{-\frac{1}{2}} k_x \right\|_{\text{op}}^2 \right] \\ &= \mathcal{N}_\infty(\lambda)^{(m-2)} \mathcal{N}(\lambda). \end{aligned}$$

Concentration on ξ_v. Applying Theorem 7, we get, for any $t > 0$, that

$$\mathbb{P}\left(\mathcal{B}_v(\lambda) > t\right) \leq 2 \exp\left(-\frac{nt^2}{2\sigma^2 \mathcal{N}(\lambda) + 2M\mathcal{N}_\infty(\lambda) t} \right). \tag{5.28}$$

Control of the bias

To control the bias, we recall a simpler version of Bernstein concentration inequality, that is a corollary of Theorem 7.

Theorem 9 (Concentration in Hilbert space (Pinelis and Sakhanenko, 1986)). *Let denote by $\mathcal{A}$ a Hilbert space and by (ξ_i) a sequence of independent random vectors on $\mathcal{A}$ such that $\mathbb{E}[\xi_i] = 0$, that are bounded by a constant M, with finite variance $\sigma^2 = \mathbb{E}[\sum_{i=1}^n \|\xi_i\|^2]$. For any $t > 0$,*

$$\mathbb{P}(\left\|\sum_{i=1}^n \xi_i\right\| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + 2tM/3}\right).$$

Bound on ξ_b. We have

$$\|\xi_b(x)\|_{\mathcal{G}} \leq \sup_{x\in\operatorname{supp}\rho_{\mathcal{X}}} \left\|(\Sigma+\lambda)^{-\frac{1}{2}}k_x\right\|_{\text{op}} \|g_\lambda(x) - g^*(x)\|_{\mathcal{H}} \leq \mathcal{N}_\infty(\lambda)\,\|g_\lambda - g^*\|_\infty .$$

Therefore, with Appendix 5.C.5, we get

$$\|\xi_b(x)\|_{\mathcal{G}} \leq b_1\lambda^{q-p}\mathcal{N}_\infty(\lambda).$$

Variance of ξ_b. For the variance we proceed with

$$\|\xi_b(x)\|_{\mathcal{G}}^2 \leq \mathcal{N}_\infty(\lambda)^2\,\|g_\lambda(x) - g^*(x)\|_{\mathcal{H}}^2 .$$

Therefore,

$$\mathbb{E}[\|\xi_b(X)\|^2] \leq \mathcal{N}_\infty(\lambda)^2\,\|g_\lambda - g^*\|_{L^2}^2 .$$

Using the derivations made in Appendix 5.C.5, we have, using that $q \leq 1$,

$$\begin{aligned}\|g_\lambda - g^*\|_{L^2} &= \lambda\left\|(K+\lambda)^{-1}K^q g_0\right\|_{L^2} \leq \lambda\left\|(K+\lambda)^{-(1-q)}\right\|_{\text{op}}\|(K+\lambda)^{-q}K^q\|_{\text{op}}\,\|g_0\|_{L^2}\\ &\leq \lambda^q\,\|g_0\|_{L^2} .\end{aligned}$$

Concentration on ξ_b. Adding everything together, we get

$$\mathbb{P}\left(\mathcal{B}_b(\lambda) > t\right) \leq 2\exp\left(-\frac{nt^2}{2\left(\lambda^{2q}\mathcal{N}_\infty(\lambda)^2\,\|g_0\|_{L^2}^2 + b_1\lambda^{q-p}\mathcal{N}_\infty(\lambda)t/3\right)}\right). \tag{5.29}$$

Note that based on the bound on the variance, we would like $\mathcal{N}_\infty(\lambda)^2\lambda^{2q} \approx \lambda^{2(q-p)}$ to be smaller than $\mathcal{N}(\lambda) \approx \lambda^{-\sigma}$. It is the case since $q > p$.

Union bound

To control $\|g_n - g_\lambda\|_{L^\infty} \leq c_p\lambda^{-p}\mathcal{A}(\lambda)(\mathcal{B}_v(\lambda) + \mathcal{B}_b(\lambda))$, we need to perform a union bound on the control of $\mathcal{A}$ and the control of $\mathcal{B} := \mathcal{B}_v + \mathcal{B}_b$, we use that for any $t > 0$ and $0 < s < 1$, $c_p\lambda^{-p}\mathcal{A}\mathcal{B} > t$ implies $\mathcal{A} > {}^1/_{(1-s)}$ or $\mathcal{B} > {}^{(1-s)t\lambda^p}/_{c_p}$. Similarly, $\mathcal{B}_v + \mathcal{B}_b > t$, implies that either $\mathcal{B}_v > {}^t/_2$, either $\mathcal{B}_b > {}^t/_2$. Therefore, we have, the following inclusion of events (with respect to $\mathcal{D}_n$)

$$\{\|g_n - g_\lambda\|_{L^\infty} > t\} \subset \left\{\mathcal{A} > \frac{1}{1-s}\right\} \cup \left\{\mathcal{B}_v > \frac{(1-s)t\lambda^p}{2c_p}\right\} \cup \left\{\mathcal{B}_b > \frac{(1-s)t\lambda^p}{2c_p}\right\}.$$

In terms of probability this leads to

$$\mathbb{P}_{\mathcal{D}_n}\left(\|g_n - g_\lambda\|_{L^\infty} > t\right) \leq \mathbb{P}_{\mathcal{D}_n}\left(\mathcal{A} > \frac{1}{1-s}\right) + \mathbb{P}_{\mathcal{D}_n}\left(\mathcal{B} > \frac{(1-s)t\lambda^p}{c_p}\right). \tag{5.30}$$

Looking closer it is the term in $\mathcal{B}$ that will be the more problematic, therefore we would like s to be small. If we take s to be a constant with respect to t, we will get something that behaves like $\mathbb{P}(B > t\lambda^p)$, which is the

best we can hope for (this also explain why we divide $\mathcal{B} > t$ in $\mathcal{B}_v > t/2$ or $\mathcal{B}_b > t/2$). We will consider $s = 1/2$. We express concentration based on the expression of $\mathcal{N}$ and $\mathcal{N}_\infty$, assuming $\lambda \leq \|\Sigma\|_{\text{op}}$, and $n > a_3^2 \lambda^{-2p}$

$$\mathbb{P}_{\mathcal{D}_n}(\mathcal{A} > 2) \leq 28 a_2 \lambda^{-\sigma} \exp(-\frac{n\lambda^{2p}}{10 a_3^2}).$$

Similarly, we get, when $\lambda \leq 1$, using that $\lambda^{-\sigma} \geq 1$

$$\mathbb{P}_{\mathcal{D}_n}(\mathcal{B}_v > t/4) \leq 2 \exp\left(-\frac{n\lambda^\sigma t^2}{32\sigma^2 a_2 + 8 M a_3 \lambda^{-p} t}\right).$$

For the bias term, we can proceed at a brutal bounding, based on the fact that for $\lambda \leq 1$, $\lambda^{q-p} \leq 1 \leq \lambda^{-\sigma}$, to get

$$\mathbb{P}_{\mathcal{D}_n}(\mathcal{B}_b > t/4) \leq 2 \exp\left(-\frac{n\lambda^\sigma t^2}{32 a_3^2 \|g_0\|_{L^2} + 8 b_1 a_3 \lambda^{-p} t/3}\right).$$

With $b_4 = \max(32\sigma^2 a_2, 32 a_3^2 \|g_0\|_{L^2})$ and $b_5 = \max(8 M a_3, 8 b_1 a_3/3)$, we get the following union bound

$$\mathbb{P}_{\mathcal{D}_n}\left(\mathcal{B} > \frac{t\lambda^p}{2}\right) \leq 4 \exp\left(-\frac{n\lambda^{2p+\sigma} t^2}{b_4 + b_5 t}\right).$$

We proceed with the union bound on $\|g_n - g_\lambda\|_{L^\infty}$ as

$$\mathbb{P}_{\mathcal{D}_n}(\|g_n - g_\lambda\|_{L^\infty} > t) \leq b_2 \lambda^{-\sigma} \exp(-b_3 n \lambda^{2p}) + 4 \exp\left(-\frac{n\lambda^{2p+\sigma} t^2}{b_4 + b_5 t}\right),$$

with $b_2 = 28 a_2$ and $b_3^{-1} = 10 a_3^2$, as long as $b_3 n > \lambda^{-2p}$, and $\lambda \leq \max(1, \|K\|_{\text{op}})$.

Refinement of Lemma 14

Remark that the uniform control in Lemma 14 is more than we need, we only need control for each x as described in Assumption 2. Indeed, if $p(x)$ is such that there exists a constant $\tilde{c_p}$ (that does not depend on x or λ), such that for any $i \in \mathbb{N}$

$$\langle k_x, u_i \rangle_{\mathcal{G}_\mathcal{X}} \leq \tilde{c_p} \lambda_i^{p(x)},$$

then considering that

$$g_n(x) - g_\lambda(x) = k_x^\star(\gamma_n - \gamma_\lambda) = k_x^\star (\Sigma + \lambda)^{-\frac{1}{2}} (\Sigma + \lambda)^{\frac{1}{2}} (\gamma_n - \gamma_\lambda),$$

we can improve the results of Lemma 14 by replacing p by $p(x)$. While we considered $p = \sup_{x \in \rho_\mathcal{X}} p(x)$ as a consequence of our proof scheme, one can expect to end up with the $\mathbb{E}_X[\lambda^{p(X)}]$ instead of λ^p when deriving the proof of Theorems 5 and 6 (for which one has to refine Theorem 2 in order to integrate dependency of L to x, similarly to what is done in Lemma 17), which will lead to better rates. Yet, because of the complexity of expressing a quantity of the type $\mathbb{E}_X[\varphi(p(X))]$, for some function φ, we decided not to present this improved version in the paper.

5.C.7 Proof of Theorem 5

Based on the proof of Theorem 1, we know that

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty \mathbb{P}_{\mathcal{D}_n}(\|g_n - g^*\|_\infty > t_0).$$

Now we use that

$$\mathbb{P}_{\mathcal{D}_n}(\|g_n - g^*\|_\infty > t_0) \leq \mathbb{P}_{\mathcal{D}_n}(\|g_n - g_\lambda\|_\infty > t_0 - \|g_\lambda - g^*\|_\infty).$$

The result follows from derivations in Appendix 5.C.6, where we used that when k is bounded, Assumptions 7 and 8 are verified with $\sigma = 1$ and $p = 1/2$. Note that we do not need the source assumption, since we can bound directly $\|g_\lambda - g^*\|_{L^2} \leq \|g_\lambda - g^*\|_{L^\infty} < t_0$ while retaking the proof in Appendix 5.C.6. Moreover, the results of this last proof holds under the condition $n\lambda b_3 > 1$, but, since $\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty$, we can augment the constant b_6 so that the result in Theorem 5 still holds for any $n \in \mathbb{N}^*$.

5.C.8 Proof of Theorem 6

We can rephrase Lemma 14, using a union bound

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_n}(\|g_n - g^*\| > t) &\leq \mathbb{P}_{\mathcal{D}_n}(\|g_n - g_\lambda\| > t/2) + \mathbb{P}_{\mathcal{D}_n}(\|g_\lambda - g^*\| > t/2) \\ &\leq b_2\lambda^{-\sigma}\exp\left(-b_3 n\lambda^{2p}\right) + 4\exp\left(-\frac{n\lambda^{2p+\sigma}t^2}{4b_4 + 2b_5 t}\right) + \mathbf{1}_{t\leq 2\lambda^{q-p}}.\end{aligned}$$

Using variant of Theorem 2 presented in Appendix 5.A.6, we get

$$\begin{aligned}\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty b_2\lambda^{-\sigma}\exp\left(-b_3 n\lambda^{2p}\right) &+ 2c_\psi c_\alpha 2^{\alpha+1}\lambda^{(q-p)(\alpha+1)} \\ &+ 2c_\psi c_\alpha c\left(b_4^{\frac{\alpha+1}{2}}(n\lambda^{2p+\sigma})^{-\frac{\alpha+1}{2}} + b_5^{\alpha+1}(n\lambda^{2p+\sigma})^{-(\alpha+1)}\right).\end{aligned}$$

As long as $\lambda \leq \max(\|K\|_{\text{op}}, 1)$ and $n \geq (b_3\lambda^{2p})^{-1}$. We optimize those rates with $\lambda = \lambda_0 n^{-\gamma}$, and γ satisfying

$$2\gamma(q-p) = 1 - \gamma(2p+\sigma) \qquad \Rightarrow \qquad \gamma = (2q+\sigma)^{-1}.$$

This leads to, for n after a certain $N \in \mathbb{N}^*$

$$\begin{aligned}\mathcal{R}(f_n) - \mathcal{R}(f^*) &\leq \ell_\infty b_2\lambda_0^{-\sigma} n^{\frac{\sigma}{2q+\sigma}}\exp\left(-b_3 n\lambda_0^{2p} n^{\frac{2(q-p)+\sigma}{2q+\sigma}}\right) \\ &\quad + 2c_\psi c_\alpha 2^{\alpha+1}\lambda_0^{(q-p)(\alpha+1)} n^{-\frac{(q-p)(\alpha+1)}{2q+\sigma}} \\ &\quad + 2c_\psi c_\alpha c\left(b_4^{\frac{\alpha+1}{2}}\lambda_0^{\frac{(2p+\sigma)\alpha+1}{2}} n^{-\frac{(q-p)(\alpha+1)}{2q+\sigma}} + b_5^{\alpha+1}\lambda_0^{(2p+\sigma)\alpha+1} n^{-\frac{2(q-p)(\alpha+1)}{2q+\sigma}}\right) \\ &\leq b_8 n^{-\frac{2(q-p)(\alpha+1)}{2q+\sigma}}.\end{aligned}$$

Since ℓ is bounded, $\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty$, and we can always higher b_8, in order to have the inequality for any $n \in \mathbb{N}^*$.

Chapter 6

Exponential Convergence Rates for SVM

The following is a reproduction of Cabannes and Vigogna (2022).

Classification is often the first problem described in introductory machine learning classes. Generalization guarantees of classification have historically been offered by Vapnik-Chervonenkis theory. Yet those guarantees are based on intractable algorithms, which has led to the theory of surrogate methods in classification. Guarantees offered by surrogate methods are based on calibration inequalities, which have been shown to be highly suboptimal under some margin conditions, failing short to capture exponential convergence phenomena. Those "super fast" rates are becoming well understood for smooth surrogates, but the picture remains blurry for non-smooth losses such as the hinge loss, associated with the renowned support vector machines. In this paper, we present a simple mechanism to obtain fast convergence rates, and we investigate its usage for SVM. In particular, we show that SVM can exhibit exponential convergence rates even without assuming the hard Tsybakov margin condition.

6.1 Introduction

To solve a problem with computer calculations, classical computer science consists in handcrafting a set of rules. In contrast, machine learning is based on the collection of a vast amount of solved instances of this problem, and on the automatic tuning of an algorithm that maps inputs defining the problem to the desired outputs. Denote by x in a space $\mathcal{X}$ the inputs, by $y \in \mathcal{Y}$ the outputs, and by $f : \mathcal{X} \to \mathcal{Y}$ the input/output mappings. To learn a mapping f^*, it is customary to introduce an explicit metric of error, and search for the function that minimizes it. Define this metric through a loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that quantifies how bad a prediction $f(x)$ is when we observe y. If we assume the existence of a distribution over I/O pairs, $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$, that generates the instances of the problem we mean to solve, we aim to minimize the average loss value

$$\mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\rho}\left[\ell(f(X), Y)\right]. \tag{6.1}$$

In practice, this "risk" $\mathcal{R}$ can be evaluated approximately with samples $\mathcal{D}_n = (X_i, Y_i)_{i\leq n}$, collected by the machine learning scientist and assumed to have been drawn independently accordingly to ρ.

We shall focus on the binary classification problem where $\mathcal{Y} = \{-1, 1\}$, and ℓ is the zero-one loss $\ell(y, z) = \mathbf{1}_{y\neq z}$. In this setting, the risk $\mathcal{R}(f)$ captures the probability of mistakes of a classifier f, and its minimizer is characterized by

$$f^* = \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f) = \operatorname{sign}\eta, \qquad \text{where} \qquad \eta(x) = \mathbb{E}\left[Y \mid X = x\right]. \tag{6.2}$$

Ideally, leveraging the dataset $\mathcal{D}_n$, we would like to find a mapping $f_{\mathcal{D}_n} : \mathcal{X} \to \mathcal{Y}$ that is close to be optimal, in the sense that the excess of risk $\mathcal{E}(f_{\mathcal{D}_n}) = \mathcal{R}(f_{\mathcal{D}_n}) - \mathcal{R}(f^*)$ is as small as it could be. Since this quantity is actually random, inheriting from the randomness of the samples, we will focus on controlling its average.[1] In particular, we will show that, when our model is well-specified, as the number of samples grows, this average decays actually much faster than what usual statistical learning theory suggests. We give a brief historical review of related literature before précising our contributions.

[1] Note that one could also be interested in controlling its tail, but this is out-of-score of the current literature on "super fast" rates.

6.1.1 Statistical learning theory

The classical approach to minimize (6.1) without the knowledge of ρ but with the sole access to samples $\mathcal{D}_n \sim \rho^{\otimes n}$ is to restrict the search over functions in a class $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$, and look for an empirical risk minimizer

$$f^*_{\mathcal{D}_n} \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \mathcal{R}_{\mathcal{D}_n}, \qquad \text{where} \qquad \mathcal{R}_{\mathcal{D}_n}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i). \tag{6.3}$$

If we denote by $f^*_{\mathcal{F}}$ the minimizer of $\mathcal{R}$ in $\mathcal{F}$, using the fact that $\mathcal{R}_{\mathcal{D}_n}(f^*_{\mathcal{F}}) \geq \mathcal{R}_{\mathcal{D}_n}(f^*_{\mathcal{D}_n})$, the excess of risk can be bounded as

$$\mathcal{R}(f^*_{\mathcal{D}_n}) - \mathcal{R}(f^*) \leq \underbrace{2 \sup_{f \in \mathcal{F}} \left| \mathcal{R}(f) - \mathcal{R}_{\mathcal{D}_n}(f) \right|}_{\text{estimation error}} + \underbrace{\mathcal{R}(f^*_{\mathcal{F}}) - \mathcal{R}(f^*)}_{\text{approximation error}}. \tag{6.4}$$

This bound can be seen as highly suboptimal because it bounds the deviation of a random function with the worst deviation in the function class. However, for any class $\mathcal{F}$, there exists an "adversarial" distribution ρ for which convergence rates (of the excess of risk toward zero as a function of the number of samples $n \in \mathbb{N}$) derived through this bound can not be improved beside lowering some multiplicative constants (Vapnik, 1995). On the one hand, the estimation error can be controlled with general tools to bound the supremum of a random process (*e.g.*, Dudley, 1967), and will decrease with a decrease in the size of the class $\mathcal{F}$. On the other hand, the approximation error depends on assumptions of the problem, and the bigger the size of the class $\mathcal{F}$, the less restrictive it will be to assume that f^* is not too different from $f^*_{\mathcal{F}}$. Hence, there is a clear trade-off between controlling both errors, which should be balanced in order to optimize a bound on the full excess of risk.

6.1.2 Surrogate methods

In practice, due to the combinatorial nature of discrete-valued functions, finding the empirical risk minimizer (6.3) is often an intractable problem (*e.g.*, Höffgen and Simon, 1992; Arora et al., 1997). Therefore, people have approached the original problem with other perspectives. A straightforward approach is given by *plug-in classifiers*, *i.e.*, classifiers of the form $\operatorname{sign} \hat{\eta}$, for $\hat{\eta}$ some estimator of η. For example, such an estimator can be constructed as $\hat{\eta}(x) = \sum_{i=1}^{n} \alpha_i(x) Y_i$, for $\alpha_i(x)$ some weights that specify how much the observation Y_i made at the point X_i should diffuse to the point x (see Friedman, 1994, for an example). Another popular approach to solve classification problems is provided by *support vector machines* (SVM), which were introduced from geometric considerations to maximize the margin between the classes $\{x \in \mathcal{X} \mid f^*(x) = y\}$ for $y \in \{-1, 1\}$ (Cortes and Vapnik, 1995).

These two approaches can be conjointly understood as introducing a surrogate loss $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ and looking for a continuous-valued function $g : \mathcal{X} \to \mathbb{R}$ that solves the surrogate problem

$$f = \operatorname{sign} g, \qquad g^* \in \operatorname*{arg\,min}_{g : \mathcal{X} \to \mathbb{R}} \mathcal{R}_S(g), \qquad \text{where} \qquad \mathcal{R}_S(g) = \mathbb{E}_{(X,Y)\sim\rho} \left[L(g(X), Y) \right], \tag{6.5}$$

where the notation S stands for "surrogate". To an estimate $g : \mathcal{X} \to \mathbb{R}$ of g^* we associate an estimate $f : \mathcal{X} \to \mathcal{Y}$ of f^* through the decoding step $f = \operatorname{sign} g$. In particular, using the variational characterization of the mean, η can be estimated through $L(z, y) = |z - y|^2$. Regarding SVM, they are related to the *hinge loss* (see, *e.g.*, Steinwart and Christmann, 2008)

$$L(z, y) = \max(0, 1 - zy), \tag{6.6}$$

Surrogate methods benefit from their relative easiness to optimize and the quality of their practical results. Arguably, they define the current state of the art in classification, softmax regression being particularly popular to train neural networks on classification tasks.

Surrogate methods were studied in depth by Bartlett et al. (2006), who proposed a generic framework to relate the excess of the original risk to the excess of surrogate risk through an inequality of the type

$$\mathcal{R}(f) - \mathcal{R}(f^*) \leq \psi \left(\mathcal{R}_S(g) - \mathcal{R}_S(g^*) \right), \tag{6.7}$$

where $f = \operatorname{sign} g$ and ψ is a concave function, uniquely defined from L and verifying $\psi(0) = 0$. The use of a concave function is motivated by Jensen inequality, allowing to integrate an inequality derived pointwise (conditionally on an input x).

6.1.3 Exponential convergence rates

On the one hand, calibration inequalities (6.7) are appealing, as they allow casting directly rates derived on the surrogate problem to rates on the original problem. On the other hand, because ψ has to be concave, rates in $O(n^{-r})$ on the surrogate problem can not be cast as better rates on the original problem, corresponding to the optimal inequalities where $\psi(x) = cx$ for some $c > 0$, Yet, one can find cases where the sign of η can be estimated much faster than η itself, even when this sign is estimated with surrogate methods. In particular, Mammen and Tsybakov (1999) (see also Massart and Nédélec, 2006) introduced the following condition.

Assumption 10 (Hard margin condition)**.** *The binary classification problem defined through the distribution ρ is said to verify the (Tsybakov) hard margin condition if the conditional mean η is bounded away from zero, i.e.,*

$$\exists \eta_0 > 0; \qquad |\eta(X)| > \eta_0 \qquad a.s., \tag{6.8}$$

where the notation a.s. stands for almost surely. Equivalently, $|\eta|^{-1} \in L^\infty(\rho_\mathcal{X})$.

Indeed, as shown in Appendix 6.A, under Assumption 10, leveraging sign equality, we get for any estimate $g_{\mathcal{D}_n} : \mathcal{X} \to \mathbb{R}$ computed from the dataset $\mathcal{D}_n$,

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\operatorname{sign} g_{\mathcal{D}_n})] - \mathcal{R}(f^*) \leq \mathbb{P}_{\mathcal{D}_n}\left(\left\|g_{\mathcal{D}_n} - \eta\right\|_{L^\infty} > \eta_0\right).$$

As a consequence, an exponential concentration inequality on the L^∞ distance between $g_{\mathcal{D}_n}$ and η directly translates to exponential convergence rates on the average excess of risk. In particular, estimation methods for η based on Hölder classes of functions, such as local polynomials, are known to be well-behaved with respect to the L^∞ norm (see, *e.g.*, the construction of covering number by Kolmogorov and Tikhomirov, 1959). This was leveraged by Audibert and Tsybakov (2007) in a seminal paper that shows how better rates can be achieved on the classification problem under Assumption 10 and a variety of weaker conditions (described later in Assumption 11).

Surprisingly, such an approach has remained somehow less popular than approaches based on calibration inequalities, and we are missing a framework to fully apprehend fast rates phenomena. Some results were achieved by Koltchinskii and Beznosova (2005). Recently, Cabannes et al. (2021c) showed that this result generalizes to any discrete output learning problem, and that approaches that naturally lead to concentration in L^2 could be turned into fast rates based on interpolation inequalities that relate the L^2 norm with the L^∞ one (notably reusing the work of Fischer and Steinwart (2020) on interpolation spaces). Exploiting the work of Marteau-Ferey et al. (2019), this can be generalized to any self-concordant loss (using self-concordance to reduce the problem to a least-squares problem); and, through the work of Lin et al. (2020), to any spectral filtering technique (beyond Tikhonov regularization), such as stochastic gradient descent, which was actually shown earlier for binary classification by Pillaud-Vivien et al. (2018b) and Nitanda and Suzuki (2019). In the same stream of research, Vigogna et al. (2022) proposed a general framework to study exponential rates for smooth losses in multiclass classification beyond least-squares.

6.1.4 Contribution

The proofs of exponential convergence in the works quoted above are all based on the basic mechanism outlined in Audibert and Tsybakov (2007). Unfortunately, such a mechanism does not easily extend to the hinge loss. Does this mean that support vector machines do not exhibit super fast rates, and thus they are inferior to other surrogate methods? The practice seems to answer negatively. In this paper, we give a firm theoretical answer to this question. In particular, we show not only that support vector machines do achieve exponential rates, but also that they can do so even without assuming the hard margin condition. Our main contribution is to introduce a general framework to prove exponential convergence rates, and show how this framework can be applied to the hinge loss while only considering classical assumptions.

Outline. Our general strategy is illustrated on Figure 6.1 and consists in first finding a relation

$$\mathcal{R}_S(g_\theta) - \mathcal{R}_S(g_{\theta^*}) \geq \|\theta - \theta^*\|,$$

for some natural parameter θ in a Banach space Θ parametrizing a class of functions $g_\theta \in \mathcal{F}$, and then show that $\operatorname{sign} g_\theta = \operatorname{sign} g_{\theta^*}$ when $\|\theta - \theta^*\|$ is small enough, that is,

$$\exists \varepsilon > 0; \qquad \|\theta - \theta^*\| \leq \varepsilon \quad \Rightarrow \quad \operatorname{sign} g_\theta = \operatorname{sign} g_{\theta^*}.$$

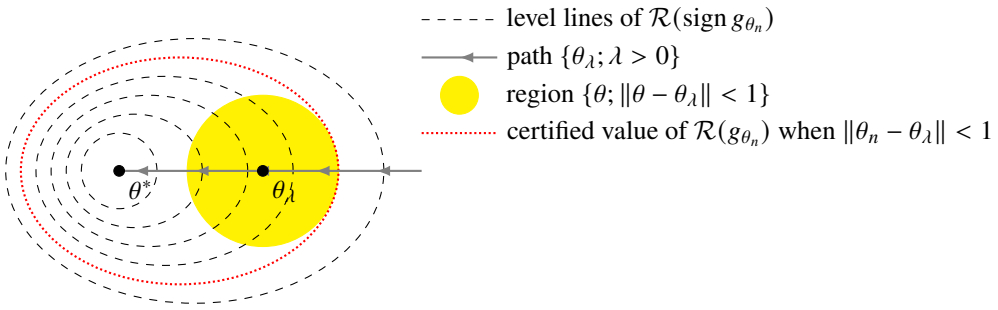

Figure 6.1: Our convergence analysis consists in relating natural concentration given by surrogate methods to the original excess of risk without passing by the surrogate excess of risk. As the drawing shows, concentration in parameter space Θ can be cast as deviation on the original excess of risk. Yet, such a casting relation depends on the geometry of this picture, which itself depends on what surrogate is used, what is the function to learn, how a regularized estimator approached it, and how our empirical estimate concentrates around the regularized estimator. Note that this figure illustrates an abstract mechanism that generalizes the simpler mechanism we use to derive exponential convergence rates.

Assuming that $\operatorname{sign} g_{\theta^*} = f^*$, we deduce that, as shown in Appendix 6.A,

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\operatorname{sign} g_{\theta_n})] - \mathcal{R}(f^*) \leq \mathbb{P}_{\mathcal{D}_n}\left(\mathcal{R}_S(g_{\theta_n}) - \mathcal{R}_S(g_{\theta^*}) \geq \varepsilon\right),$$

where g_{θ_n} is an estimate of g_θ based on the samples $\mathcal{D}_n$. Finally, we conclude with an exponential concentration inequality that controls the deviation of the excess of risk based on classical statistical learning theory.

6.2 Exponential convergence of SVM

This section is devoted to the proof of exponential convergence rates for the hinge loss. We shall fix the notation $\mathcal{R}_S$ as the surrogate risk associated with (6.6). All the proofs are collected in Appendix 6.A.

6.2.1 Refined calibration for the hinge loss

We start by introducing the classical weak margin condition (Mammen and Tsybakov, 1999).

Assumption 11 (Weak margin condition)**.** *The binary classification problem defined through the distribution ρ is said to verify the (Tsybakov) p-margin condition, with $p \in (0, \infty)$, if there exists a constant $c > 0$ such that*

$$\mathbb{P}_{\rho_\mathcal{X}}(0 < |\eta(X)| < t) \leq ct^p, \tag{6.9}$$

where the notation $\rho_\mathcal{X}$ denotes the marginal of ρ over $\mathcal{X}$.

Assumption 11 is equivalent to asking for the inverse of the conditional mean $|\eta|^{-1}$ (with the convention $0^{-1} = 0$) to belong to the Lorentz space $L^{p,\infty}(\rho_\mathcal{X})$ (also known as weak-L^p space), which is the Banach space endowed with the norm (quasi-norm and quasi-Banach if $p < 1$)

$$\|f\|_{p,\infty} = \sup_{t>0} t\, \mathbb{P}_{\rho_\mathcal{X}}(f(X) > t)^{\frac{1}{p}}, \tag{6.10}$$

where the $\rho_\mathcal{X}$ denotes the marginal of ρ with respect to $\mathcal{X}$. This definition can be extended to the case $p = \infty$ by setting $L^{p,\infty}(\rho_\mathcal{X}) = L^\infty(\rho_\mathcal{X})$, which characterizes the hard margin condition in Assumption 10. We will also use $\|\cdot\|_p$, for $p \in [1, \infty]$, to denote the L^p-norm on $\mathcal{X}$ endowed with $\rho_\mathcal{X}$.

We now relate the excess of risk on the hinge loss to the deviation in these spaces.

Lemma 29 (Weak-L^q concentration due to the hinge loss)**.** *For any functions $g_1, g_2 : \mathcal{X} \to [-1, 1]$,*

$$\mathcal{R}_S(g_2) - \mathcal{R}_S(g_1) = \mathbb{E}_{\rho_\mathcal{X}}[-\eta(X)(g_2(X) - g_1(X))]. \tag{6.11}$$

In particular, under Assumption 10, for any $g : \mathcal{X} \to \mathbb{R}$,

$$\mathcal{R}_S(g) - \mathcal{R}_S(g^*) \geq \left\||\eta|^{-1}\right\|_\infty^{-1} \|\pi(g) - g^*\|_1, \tag{6.12}$$

where $g^* = \operatorname{sign}\eta$ is a minimizer of $\mathcal{R}_S$ and π is the projection of $\mathbb{R}$ on $[-1,1]$, defined as mapping $t \in \mathbb{R}$ to $\pi(t) = \operatorname{sign}(t)\min\{|t|,1\}$. Similarly, under Assumption 11, with $q = {}^{p}/_{p+1}$,

$$\mathcal{R}_S(g) - \mathcal{R}_S(g^*) \geq 2^{-1}\left\||\eta|^{-1}\right\|_{p,\infty}^{-1}\|\pi(g) - g^*\|_{q,\infty}\,. \tag{6.13}$$

Lemma 29 shows that we can set the minimizer $g^* = f^* \in \{-1,1\}^{\mathcal{X}}$. This is a useful fact as it implies that the excess of the original risk is zero as soon as $\|g - g^*\|_\infty < 1$. In essence, the only piece missing in order to prove fast convergence rates is an interpolation inequality between $L^{q,\infty}$ and L^∞. In the following, we will leverage Lemma 29 more subtly by considering a class of functions $\mathcal{G}$ and assumptions on the distribution $\rho_{\mathcal{X}}$ such that, if an estimate $g \in \mathcal{G}$ has not the same sign almost everywhere as the estimand g^*, then $\|g - g^*\|_{q,\infty}$ is bounded away from zero. By contraposition, if $g \in \mathcal{G}$ presents a small excess of surrogate risk, then $\operatorname{sign} g = \operatorname{sign} g^*$. When $\mathcal{X}$ is a metric space, one way to proceed is to assume that g is Lipschitz-continuous, together with some minimal mass assumptions. Let us begin with the minimal mass assumption. We first need the following definition.

Definition 30 (Well-behaved sets)**.** *A set $U \subset \mathcal{X}$ is said to be well-behaved with respect to ρ if there exist constants $c, r, d > 0$ such that, for any $x \in U$,*

$$\forall\,\varepsilon \in [0,r]; \qquad \rho_{\mathcal{X}}(U \cap \mathcal{B}(x,\varepsilon)) \geq c\varepsilon^d, \tag{6.14}$$

and $\mathcal{B}(x,\varepsilon)$ the ball in $\mathcal{X}$ of center x and radius ε.

The following examples show that the coefficient d that appears in (6.14) results from the dimension of the ambient space, the regularity of singularities of the border of the set, and the decay of the density when approaching the frontier of the set.

Example 14. *The set $[0,1]^p$ is well-behaved with coefficients $r = 1$, $d = p$ and $c = 2^{-d}\operatorname{vol}(\mathbb{S}^{d-1})$ with respect to the Lebesgue measure in $\mathbb{R}^d$.*

Example 15. *The set $\left\{(x,y) \in \mathbb{R}^2 \,\middle|\, x \in [0,1], y \in [0,x^{n-1}]\right\}$ is well-behaved with coefficient $r = 1$, $d = n$ and $c = n^{-1}$ with respect to the Lebesgue measure. Reciprocally, the set $[0,1]$ is well-behaved with coefficient $r = 1$, $d = n$ and $c = n^{-1}$ with respect to the measure whose density equals $p(x) = x^{n-1}$.*

Assumption 12 (Minimal mass assumption)**.** *The decision regions $\mathcal{X}_y = \{x \in \operatorname{supp}(\rho_{\mathcal{X}}) \mid f^*(x) = y\}$ for $y \in \{-1,1\}$ are well-behaved.*

Assumption 12 is a weakening of an assumption that is commonly found in the statistical learning literature. More precisely, it is often assumed that ρ is absolutely continuous according to the Lebesgue measure λ on $\mathcal{X}$ (assumed to be a Euclidean space), that its density is bounded away from zero on its support, and that its support has smooth boundary, so that $\lambda(\operatorname{supp}\rho_{\mathcal{X}} \cap \mathcal{B}(x,\varepsilon)) > c'\lambda(\mathcal{B}(x,\varepsilon))$ (see the strong density assumption in Audibert and Tsybakov, 2007).

The minimal mass requirement allows relating misclassification events to $L^{q,\infty}$ deviation.

Lemma 31. *Under Assumption 12, there exists a constant c_0 such that if g is G-Lipschitz-continuous for $G > r^{-1}$, for any $q \in (0,1]$*

$$\exists\, x \in \operatorname{supp}\rho_{\mathcal{X}}; |g(x) - g^*(x)| \geq 1 \quad \Rightarrow \quad \|g - g^*\|_{q,\infty} \geq c_0 G^{-\frac{d}{q}}. \tag{6.15}$$

Putting together Lemmas 29 and 31, we obtain the following refined calibration.

Proposition 32. *Under Assumptions 11 and 12, if g is G-Lipschitz-continuous with $G > r^{-1}$, we have*

$$\mathcal{R}_S(g) - \mathcal{R}_S(g^*) \leq 2^{-1}\left\||\eta|^{-1}\right\|_{p,\infty}^{-1} c_0 G^{-\frac{d(p+1)}{p}} \qquad \Rightarrow \qquad \mathcal{R}(\operatorname{sign} g) = \mathcal{R}(f^*). \tag{6.16}$$

6.2.2 Trade-off between estimation and approximation errors

We are now left with the research of $g_{\mathcal{D}_n}$ inside a class of Lipschitz-continuous functions such that $\mathcal{R}_S(g_{\mathcal{D}_n}) - \mathcal{R}_S(g^*)$ is sub-Gaussian (its randomness being inherited from the dataset $\mathcal{D}_n$ from which $g_{\mathcal{D}_n}$ is built). To do so, let us consider a linear class of functions

$$\mathcal{G}_{M,\sigma} = \left\{x \mapsto \left\langle \theta, \varphi(\sigma^{-1}x)\right\rangle \,\middle|\, \theta \in \mathcal{H}, \|\theta\|_{\mathcal{H}} \leq M\right\}, \tag{6.17}$$

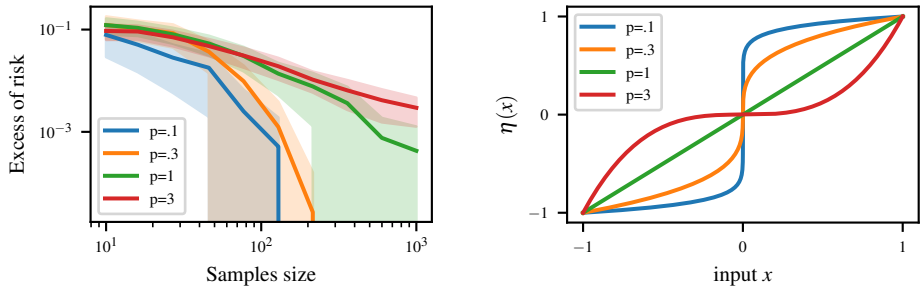

Figure 6.2: SVM generalization error as a function of the number of samples (left) for a problem where X is uniform on $[-1, .-1] \cup [.1, 1]$ and $\eta(x) = \text{sign}(x)\,|x|^p$ (right). We observe exponential convergence rates.

where $\mathcal{H}$ is a separable Hilbert space, $\varphi : \mathcal{X} \to \mathcal{H}$ is a G_φ-Lipschitz-continuous mapping, and $\sigma > 0$ is a scaling (or bandwidth) parameter. Such a class of functions can be entirely described from the kernel $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$ (see Scholkopf and Smola, 2001, for a primer on kernel methods). An example for $\mathcal{G}$ is given by the Gaussian kernel, a.k.a. radial basis function, $k(x, x') = \exp(-\|x - x'\|^2)$. Using Cauchy-Schwarz, it is easy to show that any function in $\mathcal{G}_{M,\sigma}$ is $MG_\varphi\sigma^{-1}$-Lipschitz-continuous.

In order to find a function $g_{\mathcal{D}_n}$ that is likely to minimize $\mathcal{R}_S$ without accessing the distribution ρ, but only i.i.d. samples $\mathcal{D}_n = (X_i, Y_i)_{i \le n} \sim \rho^{\otimes n}$, it is classical to consider the empirical risk minimizer

$$g_{\mathcal{D}_n} \in \operatorname*{arg\,min}_{g \in \mathcal{G}_{M,\sigma}} \frac{1}{n} \sum_{i=1}^{n} L(g(X_i), Y_i). \tag{6.18}$$

This problem is convex with respect to θ parametrizing $g \in \mathcal{G}_{M,\sigma}$, and is easily optimized with duality. We refer the curious reader to the extensive literature on SVM (see Cristianini and Shawe-Taylor, 2000; Scholkopf and Smola, 2001; Steinwart and Christmann, 2008, for books on the matter).

In order to show that $\mathcal{R}_S(g_{\mathcal{D}_n})$ is close to $\mathcal{R}_S(g^*)$, one can apply classical results from statistical learning theory, and in particular (6.4). The estimation error can be bounded using the extensive literature on Rademacher complexity for linear classes of functions on Lipschitz-continuous losses (Bartlett and Mendelson, 2002). To bound the approximation error, one needs to make additional assumptions on the problem. We refer to Steinwart and Scovel (2007); Blaschzyk and Steinwart (2018) for advanced considerations on the matter. In view of our calibration result (6.16), the following additional assumption suffices to prove exponential convergence of SVM.

Assumption 13 (Source condition)**.** *The classification problem verifies the p-margin condition 11, and there exist M, σ and a function $g \in \mathcal{G}_{M,\sigma}$ such that $\mathcal{R}_S(g) - \mathcal{R}_S(g^*) \le 4^{-1} \| |\eta|^{-1} \|_{p,\infty}^{-1} c_0 M^{-r} G_\varphi^{-r} \sigma^r$ with $r = d(p+1)/p$.*

It should be noted that Assumption 13, together with Assumption 12, implies that the decision frontier $\overline{\mathcal{X}_{-1}} \cap \overline{\mathcal{X}_1}$ (the bar notation corresponding to space closure), inherits from the regularity of g, since it is included in the set $\{x \in \mathcal{X} \mid g(x) = 0\}$. In particular, if $\mathcal{G}_{M,\sigma}$ is included in $\mathcal{C}^m$, this frontier would be in $\mathcal{C}^m$. Hence, for Assumptions 13 and 12 to hold, the boundary frontier should match the regularity implicitly defined by $\mathcal{G}_{M,\sigma}$.

We are finally ready to state our main result, establishing exponential convergence rates for SVM.

Theorem 10 (Exponential convergence rates for SVM)**.** *Under Assumptions 11, 12 and 13, there exists a constant $c > 0$ such that the empirical minimizer $g_{\mathcal{D}_n}$ defined by* (6.18) *verifies*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(\text{sign}\, g_{\mathcal{D}_n}) - \mathcal{R}(f^*) \le \exp(-cn). \tag{6.19}$$

6.2.3 Relaxing assumptions

Exponential convergence rates rely on strong assumptions in order to set the approximation error to zero. In particular, it is customary to assume that the surrogate function to learn lies in the model we have chosen,

that is, in our notation, $g^* \in \mathcal{G}_{M,\sigma}$. In our case this would be a strong assumption, since g^* is piecewise linear while $\mathcal{G}_{M,\sigma}$ is a smooth space of functions. It turns out that the assumption $g^* \in \mathcal{G}_{M,\sigma}$ is not necessary, and what we actually need is a sufficiently small risk. How small is enough is quantified by the statement of Proposition 32. From a qualitative point of view, the requirement of Assumption 13 is natural (surrogate risk must be smaller than ε), with the technical part being just a quantification of the needed behavior (bound on such ε).

To deepen the study of the approximation error, one could leverage the following geometrical characterization of the risk of misclassification. For $f : \mathcal{X} \to \{-1, 1\}$, we have

$$\mathcal{R}(f) - \mathcal{R}(f^*) = \mathbb{E}[|\eta(X)| \, \mathbf{1}_{f(x) \neq f^*(X)}] \leq \mathbb{P}(f(X) \neq f^*(X)) = \rho_{\mathcal{X}} \left(f^{-1}(\{1\}) \triangle \mathcal{X}_1 \right), \tag{6.20}$$

where $\triangle$ denotes the symmetric difference of sets, *i.e.* $A \triangle B = (A \cup B) \setminus (A \cap B)$. In particular, if the function class $\mathcal{G}_{M,\sigma}$ is rich enough, and the classes $\mathcal{X}_1$ and $\mathcal{X}_{-1}$ are separated, in the sense that the distance between any two points in each set is bounded away from zero,[2] then Assumption 13 holds, and the minimizer $g_{\mathcal{G}_{M,\sigma}}$ of the surrogate risk in $\mathcal{G}_{M,\sigma}$ verifies

$$\rho_{\mathcal{X}} \left((\operatorname{sign} g_{\mathcal{G}_{M,\sigma}})^{-1}(\{1\}) \triangle \mathcal{X}_1 \right) \leq \psi(M, \sigma), \tag{6.21}$$

for ψ a function that vanishes for sufficiently large M and small σ.

On the one hand, one could control the approximation error by assuming or deriving inequalities akin to (6.20) and (6.21), with different profiles of ψ. We conjecture that this can be done by assuming margin conditions that are well adapted to the geometric nature of SVM, such as the one proposed by Steinwart and Scovel (2007) (see also Gentile and Warmuth, 1999; Cristianini and Shawe-Taylor, 2000). On the other hand, the estimation error can be controlled by extending the ideas presented in this paper to study the worst value of the estimation error $\mathcal{R}(\operatorname{sign} g_{\mathcal{D}_n}) - \mathcal{R}(\operatorname{sign} g_{\mathcal{G}_{M,\sigma}})$ under the knowledge of $\mathcal{R}_S(g_{\mathcal{D}_n}) - \mathcal{R}_S(g_{\mathcal{G}_{M,\sigma}})$. Fitting M and σ to trade estimation and approximation error, such derivations would open the way to fast polynomial rates under less restrictive assumptions.

6.3 Numerical analysis

In this section, we provide experiments to illustrate and validate our theoretical findings. In order to be inline with the current practice of machine learning, instead of considering the hard constraint $\|\theta\| \leq M$ when minimizing a risk functional, we add a penalty $\lambda \|\theta\|^2$ to the risk to be minimized. Going from a constrained to a penalized framework does not change the nature of the statistical analysis, and one might loosely think of λ as $1/M$ (see, for example, Bach, 2023). All experiments are made with the Gaussian kernel. Precise details of the different settings are provided in Appendix 6.B.

First, we observe that the regime described in this paper kicks in when the error is already pretty small. On many real-world problems, we do not expect the generalization error as a function of the number of data used for training to exhibit a clear exponential behavior until an unusually big number of samples is used. This fact is illustrated on Figure 6.2, where for hard problems, the exponential behaviors still do not kick in after a thousand of samples.

Second, this paper shows that, in order to get exponential convergence rates for SVM, one needs the minimizer $g_{M,\sigma}$ of the surrogate risk over the selected class of functions to be a perfect classifier, *i.e.* its sign equals the sign of g^*. While this is not constraining under the cluster assumption, we inspect divergences from this condition on Figure 6.3. We observe that, even if g^* does not depend on the noise, $g_{M,\sigma}$ does. We also observe that the regularity of the decision boundary $\{x \in \mathcal{X} \mid \eta(x) = 0\}$ should match the regularity defined implicitly by the kernel k and the scale parameter σ.

Experimental comparisons of different classification approaches have been done by many people, and our goal is not to showcase the superiority of the SVM over least-squares, which might be considered as general wisdom that led to the golden age of SVM in the pre-deep-learning area (see Joachims, 1998, for example).

[2] This property is sometimes referred to as the *cluster assumption* (Rigollet, 2007) under which even weakly supervised learning techniques may exhibit exponential convergence rates (*e.g.* Cabannes et al., 2021b). In terms of practical applications, this assumption says that no one can continuously modify an input to go from a region of the space linked with one class to a region linked with another class without going through inputs that will never exist. This is typically true for well-curated image datasets such as CIFAR10: one can not continuously transform an image of a truck into an image of a horse without going through images that will never appear in the CIFAR10 dataset (Krizhevsky, 2009).

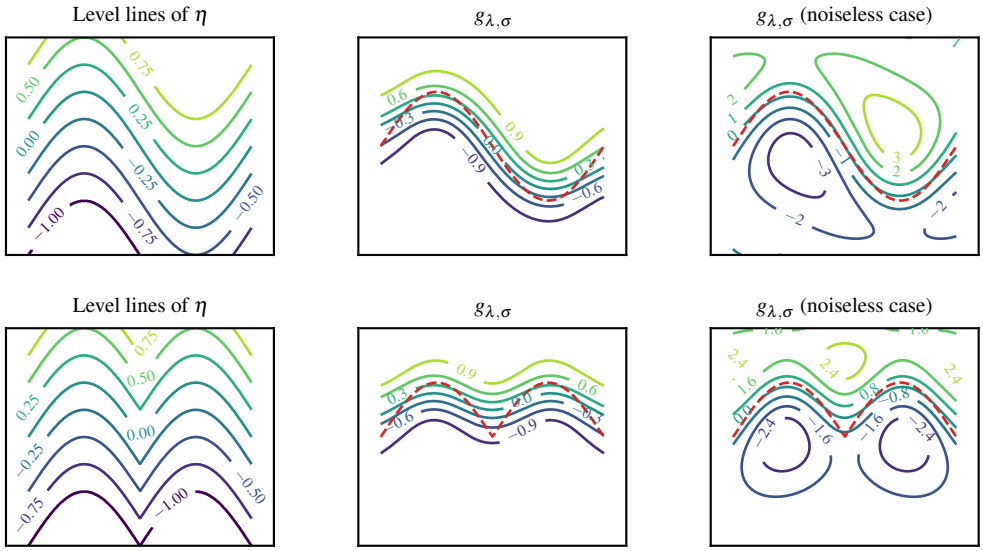

Figure 6.3: Study of the level lines of $g_{\lambda,\sigma}$ when $\eta^{-1}(0) \in \mathcal{C}^\infty$ (top) and $\eta^{-1}(0) \in \mathcal{C}^0 \setminus \mathcal{C}^1$ (bottom). The function g^* takes values -1 below the optimal decision frontier plotted in red and $+1$ above, independently of the noise. We observe that the bias error $\mathcal{R}(\operatorname{sign} g_{\lambda,\sigma}) - \mathcal{R}(f^*)$, which is bounded by the volume between the level lines $\{x \in \mathcal{X} \,|\, g_{\lambda,\sigma}(x) = 0\}$ and $\{x \in \mathcal{X} \,|\, \eta(x) = 0\}$ (plotted in red), depends on both the regularity of the latter, and on the noise level. Here, σ is taken to be of the order of 15% of the diameter of the domain, which explains the regularity of the observed level lines. The noiseless cases on the right correspond to the situations where $\mathbb{E}[Y|X] = \operatorname{sign} \eta(X)$ for η plotted on the left.

In comparison with previous works based on calibration inequalities (Rosasco et al., 2004; Steinwart, 2007), our analysis proves the robustness of SVM to noise far away from the decision boundary, in the sense that one does not need η to be bounded away from zero. This is a distinctive aspect of SVM compared to smooth surrogate methods (Nowak-Vila et al., 2020), such as softmax regression, that implicitly estimate conditional probabilities and whose performance depends on the regularity of η. We illustrate this fact graphically on Figure 6.4.

6.4 Limitations

Are surrogate methods only a proxy for classification? From a theoretical perspective, if we are only interested in the optimal mapping $f : \mathcal{X} \to \mathcal{Y}$, learning surrogate quantities can be seen as a waste of resources. In essence, this waste of resources is similar to the one occurring when we learn the full probability function $(p(y))_{y \in \mathcal{Y}}$ for some probability distribution p on $\mathcal{Y}$, while we only care about its mode. Yet, in practice, what we call a "surrogate" problem might actually be a problem of prime interest when we do not only want to predict $f^*(x)$, but we would also like to know how much we can confidently discard other potential outputs for an input x. Furthermore, assuming that a problem is exactly defined through an "original" loss that defines a clear and unique measure of error can be questioned when some practitioners evaluate methods with several metrics of performance (*e.g.* Chowdhery et al., 2022).

Do PAC-bounds provide confidence levels? Since the parameters in Assumptions 11 and 12 are hard to estimate in practice, it would be difficult to directly plug our bounds into a practical problem to derive confidence levels on how much error one might expect when deploying a model in production. Less ambitiously, we see theorems akin to Theorem 10 as providing theoretical indications that a learning method or a set of hyperparameters is sound. This is a generic downfall of probably approximately correct (PAC) generalization bounds (Valiant, 2013), which might explain why practitioners often prefer to derive error indications from test samples (see, *e.g.*, Géron, 2017). Along this line, research on conformal prediction provides interesting considerations to obtain useful confidence information from test samples (Vovk and Shafer, 2008). Finally, all these statistical methods to get confidence intervals assume representative (if not

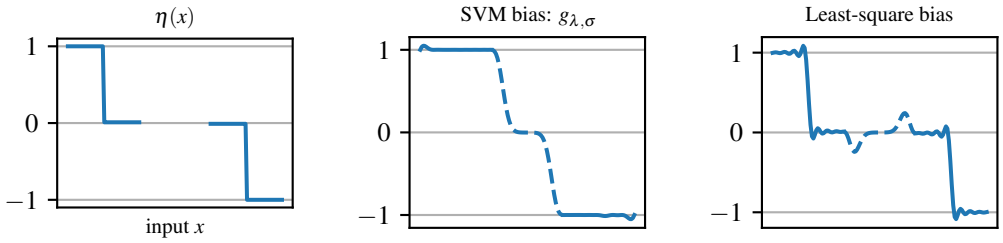

Figure 6.4: Comparison of the regularized risk (*i.e.* $\mathcal{R}_S(g_\theta) + \lambda \|\theta\|^2$) minimizer for the hinge loss surrogate (middle) and the least-squares surrogate (right), when η is not regular (left). In this setting, the hinge loss is minimized for $g = \text{sign}(\eta)$, which can be chosen regular, while the least-squares loss is minimized for $g = \eta$, which can not be chosen regular. The reconstruction is made with σ about 3% of the domain diameter, and λ relatively small. We assume no density in the middle of the domain, explaining the absence of definition of η and the dashed lines of the right figures. The oscillation on the later figures is related to the Gibbs phenomenon (Wilbraham, 1848). This phenomenon prevents the regularized least-squares solution from being a perfect classifier.

i.i.d.) data, an assumption sometimes hard to meet in practice, which is a problem that has found echoes in the civil society (*e.g.* Benjamin, 2019).

6.5 Conclusion

In this work, we were keen to illustrate a simple mechanism to get exponential convergence rates for a loss that is quite popular and whose understanding can not be easily reduced to existing work. In particular, we show that the hard margin condition is not crucial in order to derive exponential convergence rates for the SVM.

This provides a crucial step to better understand convergence rates on classification problems. An extension to generic discrete output problems could be made by considering polyhedral losses, and deriving variants of Lemma 29 (see Frongillo and Waggoner, 2021, for calibration inequalities for such losses). An important follow-up would be to provide a more global picture of fast polynomial rates for SVM under relaxations of Assumptions 12 and 13.

Finally, Chizat and Bach (2020) have made a link between two-layer wide neural networks in the interpolation regime (which implies Assumption 10 with $\eta_0 = 1$) and max-margin classifiers over specific linear classes of functions. As a consequence, we could directly plug in our analysis to prove exponential convergence rates for those small neural networks in this noiseless setting. Studying rates, constants and hyperparameter tuning in this setting would be of particular interest if it was to provide practical guidelines to deep learning practitioners in the spirit of Yang et al. (2021).

Appendix

6.A Proofs

Notation. This paper makes use of the following standard notations. We used the simplex notation Δ_A to denote the space of probability measure over a Polish space A. We also used the notation $\mathcal{Y}^{\mathcal{X}}$ to denote functions from $\mathcal{X}$ to $\mathcal{Y}$ (which can be seen as a sequence of elements in $\mathcal{Y}$ indexed by $\mathcal{X}$).

Equations. Some standard derivations in the fast rates literature were omitted in the main paper. In particular, under Assumption 10, we know that when $|g(x) - \eta(x)| < |\eta(x)|$ then $\operatorname{sign} g(x) = \operatorname{sign} \eta(x) = f^*(x)$, hence $\mathcal{R}(\operatorname{sign} g) - \mathcal{R}(f^*) \leq \mathbb{P}_X(\operatorname{sign} g(X) \neq f^*(X)) \leq 1_{\|g-\eta\|_\infty \geq \eta_0}$. As a consequence, $\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\operatorname{sign} g_n)] - \mathcal{R}(f^*) \leq \mathbb{P}_{\mathcal{D}_n}(\|g - \eta\|_\infty \geq \eta_0)$.

Similarly, the outline exposition follows from the fact that

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\operatorname{sign} g_{\theta_n})] - \mathcal{R}(f^*) &= \mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\operatorname{sign} g_{\theta_n}) - \mathcal{R}(\operatorname{sign} g_{\theta^*})] \leq \mathbb{E}_{\mathcal{D}_n}[1_{\operatorname{sign} g_{\theta_n} \neq (\operatorname{sign} g_{\theta^*})}] \\ &\leq \mathbb{E}_{\mathcal{D}_n}[1_{\|\theta_n - \theta^*\| \geq \varepsilon}] \leq \mathbb{P}_{\mathcal{D}_n}(\|\theta_n - \theta^*\| \geq \varepsilon) \\ &\leq \mathbb{P}_{\mathcal{D}_n}(\mathcal{R}_S(g_{\theta_n}) - \mathcal{R}(g_{\theta^*}) \geq \varepsilon).\end{aligned}$$

6.A.1 Proof of Lemma 29

The first part follows by integration of a pointwise result. Consider the function $h_p : \mathbb{R} \to \mathbb{R}; q \mapsto p(1-q)_+ + (1-p)(1+q)_+$, where $p \in (0,1)$ represents $\mathbb{P}(Y = 1|X)$ and q represents $g(x)$. The function h_p has a slope equal to $-p$ for $q < -1$, then slope $1 - 2p$ for $q \in (-1, 1)$, and $1 - p$ for $q > 1$. Therefore, when $q_1, q_2 \in (-1, 1)$, we have

$$h_p(q_2) - h_p(q_1) = (1 - 2p)(q_2 - q_1).$$

Taking $p = \mathbb{P}(Y = 1|X)$, $q_2 = g_2(X)$ and $q_1 = g_1(X)$, we get $1 - 2p = -\mathbb{E}[Y|X] = -\eta(X)$. By integration, we obtain the claim. From the previous slope considerations, it also follows that h_p is minimized by $q = \operatorname{sign}(2p - 1)$, meaning that one can take $g^*(X) = \operatorname{sign} \eta(X)$.

The second part follows from the fact that projecting on $[-1, 1]$ can only reduce the value of the hinge loss, that $\eta(x)(\pi(g)(x) - g^*(x))$ is always negative, and the reverse Hölder inequality:

$$\mathcal{R}_S(g) - \mathcal{R}(g^*) \geq \mathcal{R}_S(\pi(g)) - \mathcal{R}(g^*) = \|\eta(\pi(g) - g^*)\|_1 \geq \|\pi(g) - g^*\|_q \left\||\eta|^{-1}\right\|_p^{-1}.$$

A Hölder inequality also holds for weak Lebesgue spaces (see Castillo and Rafeiro, 2016, Theorem 5.23), whence

$$\mathcal{R}_S(g) - \mathcal{R}(g^*) \geq \|\eta(\pi(g) - g^*)\|_1 \geq \|\eta(\pi(g) - g^*)\|_{1,\infty} \geq \frac{1}{2}\left\||\eta|^{-1}\right\|_{p,\infty}^{-1} \|\pi(g) - g^*\|_{\frac{p}{p+1},\infty}.$$

This completes the proof.

6.A.2 Proof of Lemma 31

Assume without restrictions that there exists $x \in \mathcal{X}_1$ such that $|g^*(x) - g(x)| \geq 1$. For any event $A = A(X)$, by the law of total probability we have

$$\mathbb{P}(A) = \rho_{\mathcal{X}}(\mathcal{X}_1)\,\mathbb{P}\left(A \mid X \in \mathcal{X}_1\right) + \rho_{\mathcal{X}}(\mathcal{X}_{-1})\,\mathbb{P}\left(A \mid X \in \mathcal{X}_{-1}\right) \geq \rho_{\mathcal{X}}(\mathcal{X}_1)\,\mathbb{P}\left(A \mid X \in \mathcal{X}_1\right).$$

Hence, since $g^*(\mathcal{X}_1) = \{1\}$,

$$\|g - g^*\|_{q,\infty}^q = \sup_{t>0} t^q \, \mathbb{P}(|g(X) - g^*(X)| > t) \geq \sup_{t>0} t^q \, \mathbb{P}\left(|g(X) - 1| > t \mid X \in \mathcal{X}_1\right) \rho_{\mathcal{X}}(\mathcal{X}_1).$$

Using the triangular inequality, the G-Lipschitz continuity of g, and the definition of x, we have that, for any $x' \in \mathcal{X}$,

$$|g(x') - 1| \geq |g(x) - 1| - |g(x') - g(x)| \geq 1 - Gd(x, x').$$

As a consequence,

$$\mathbb{P}\left(|g(X) - 1| > t \mid X \in \mathcal{X}_1\right) \geq \mathbb{P}\left(X \in \mathcal{B}\left(x, \frac{1-t}{G}\right) \mid X \in \mathcal{X}_1\right) = \frac{\rho_{\mathcal{X}}\left(\mathcal{X}_1 \cap \mathcal{B}\left(x, \frac{1-t}{G}\right)\right)}{\rho_{\mathcal{X}}\left(\mathcal{X}_1\right)}.$$

Combined with the previous facts, we get

$$\|g - g^*\|_{q,\infty}^q \geq \sup_{t>0} t^q \rho_{\mathcal{X}}\left(\mathcal{X}_1 \cap \mathcal{B}\left(x, \frac{1-t}{G}\right)\right).$$

Thanks to Assumption 12, there exists (c, r, d) such that (6.14) holds for $\mathcal{X}_1$. Hence, when $G^{-1} < r$, we get the following lower bound:

$$\|g - g^*\|_{q,\infty}^q \geq cG^{-d} \sup_{t \in [0,1]} t^q (1-t)^d = cG^{-d} \frac{q^q d^d}{(d+q)^{d+q}}.$$

This proves the statement in the lemma.

6.A.3 Proof of Proposition 32

Suppose $\mathcal{R}(\operatorname{sign} g) > \mathcal{R}(f^*)$. Then, observing that $\operatorname{sign}(\pi(t)) = \operatorname{sign}(t)$ for all $t \in \mathbb{R}$, and taking $g^* = f^*$, we know there must be $x \in \operatorname{supp} \rho_{\mathcal{X}}$ such that $|\pi(g(x)) - g^*(x)| \geq 1$. Hence, by Lemma 31, we get $\|\pi(g) - g^*\|_{q,\infty} \geq c_0 G^{-\frac{d}{q}}$, and therefore, by Lemma 29, $\mathcal{R}_S(g) - \mathcal{R}_S(g^*) \geq 2^{-1} c_0 G^{-\frac{d}{q}} \left\||\eta|^{-1}\right\|_{p,\infty}^{-1}$. Thus, the proposition is proved.

6.A.4 Proof of Theorem 10

From Proposition 32 and Assumption 13, we get, with $\tilde{L} = \max\left\{MG_\varphi \sigma^{-1}, r^{-1}\right\}$ and $q = {}^{p}/_{p+1}$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\operatorname{sign} g_{\mathcal{D}_n})] - \mathcal{R}(f^*) &\leq \mathbb{P}_{\mathcal{D}_n}\left(\mathcal{R}_S(\pi \circ g_{\mathcal{D}_n}) - \mathcal{R}_S(g^*) \geq 2^{-1} \left\||\eta|^{-1}\right\|_{p,\infty}^{-1} c_0 \tilde{L}^{-\frac{d}{q}}\right) \\
&\leq \mathbb{P}_{\mathcal{D}_n}\left(\mathcal{R}_S(\pi \circ g_{\mathcal{D}_n}) - \mathcal{R}_S(g_{M,\sigma}) \geq 4^{-1} \left\||\eta|^{-1}\right\|_{p,\infty}^{-1} c_0 \tilde{L}^{-\frac{d}{q}}\right).
\end{aligned}$$

To deal with this last quantity, we proceed by usign the fact that

$$\mathcal{R}_{S,\mathcal{D}_n}(\pi \circ g_{\mathcal{D}_n}) \leq \mathcal{R}_{S,\mathcal{D}_n}(g_{\mathcal{D}_n}) \leq \mathcal{R}_{S,\mathcal{D}_n}(g_{M,\sigma}),$$

where $\mathcal{R}_{S,\mathcal{D}_n}$ denotes the empirical surrogate risk, to deduce that

$$\mathcal{R}_S(\pi \circ g_{\mathcal{D}_n}) - \mathcal{R}_S(g_{M,\sigma}) \leq \mathcal{R}_S(\pi \circ g_{\mathcal{D}_n}) - \mathcal{R}_{S,\mathcal{D}_n}(\pi \circ g_{\mathcal{D}_n}) + \mathcal{R}_{S,\mathcal{D}_n}(g_{M,\sigma}) + \mathcal{R}_S(g_{M,\sigma}).$$

Hence, we get the following union bound

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\operatorname{sign} g_{\mathcal{D}_n})] - \mathcal{R}(f^*) \leq{}& \mathbb{P}_{\mathcal{D}_n}\left(\mathcal{R}_S(\pi \circ g_{\mathcal{D}_n}) - \mathcal{R}_{S,\mathcal{D}_n}(g_{\mathcal{D}_n}) \geq 8^{-1} \left\||\eta|^{-1}\right\|_{p,\infty}^{-1} c_0 \tilde{L}^{-\frac{d}{q}}\right) \\
&+ \mathbb{P}_{\mathcal{D}_n}\left(\mathcal{R}_S(\pi \circ g_{M,\sigma}) - \mathcal{R}_{S,\mathcal{D}_n}(g_{M,\sigma}) \geq 8^{-1} \left\||\eta|^{-1}\right\|_{p,\infty}^{-1} c_0 \tilde{L}^{-\frac{d}{q}}\right).
\end{aligned}$$

Regarding the first term, we can reuse the literature on Rademacher complexity for linear models on convex risks (Bartlett and Mendelson, 2002), which ensures that

$$\mathbb{E}_{\mathcal{D}_n}\left[\sup_{g \in \mathcal{G}_{M,\sigma}} \left|\mathcal{R}_S(\pi \circ g) - \mathcal{R}_{S,\mathcal{D}_n}(\pi \circ g)\right|\right] \leq M \|\varphi\|_\infty n^{-1/2}.$$

Note that Assumption 12 implies that $\operatorname{supp}\rho_{\mathcal{X}}$ is compact, hence, if φ is Lipschitz-continuous, it is bounded on $\operatorname{supp}\rho_{\mathcal{X}}$. This allows us to use McDiarmid inequality to get the same type of bound on the deviation of $\mathcal{R}_S(g_{\mathcal{D}_n})$ around its mean. Let $H(\mathcal{D}_n) = \sup_{g\in\mathcal{G}_{M,\sigma}} \mathcal{R}_S(\pi\circ g) - \mathcal{R}_{\mathcal{D}_n}(\pi\circ g)$. Let us decompose $\mathcal{D}_n = ((x_1,y_1),\cdots,(x_n,y_n))$. We would like to show that if $\mathcal{D}'_n$ is equal to $\mathcal{D}_n$ for each datapoint but for (x_i,y_i) that becomes (x'_i,y'_i) then $H(\mathcal{D}_n) - H(\mathcal{D}'_n)$ is bounded. We have

$$\begin{aligned}
H(\mathcal{D}_n) - H(\mathcal{D}'_n) &= \sup_{g\in\mathcal{G}_{M,\sigma}} \mathcal{R}_S(\pi\circ g) - \mathcal{R}_{S,\mathcal{D}_n}(\pi\circ g) - \sup_{g'\in\mathcal{G}_{M,\sigma}} \mathcal{R}_S(\pi\circ g') - \mathcal{R}_{S,\mathcal{D}'_n}(\pi\circ g') \\
&\leq \sup_{g\in\mathcal{G}_{M,\sigma}} \mathcal{R}_{S,\mathcal{D}_n}(\pi\circ g) - \mathcal{R}_{S,\mathcal{D}'_n}(\pi\circ g) \\
&= n^{-1} \sup_{g\in\mathcal{G}_{M,\sigma}} L(\pi\circ g(x'_i),y'_i) - L(\pi\circ g(x_i),y_i) \leq n^{-1}
\end{aligned}$$

Using McDiarmid's inequality, we get

$$\mathbb{P}(H(\mathcal{D}_n) - \mathbb{E}[H(\mathcal{D}_n)] \geq t) \leq \exp(-2nt^2).$$

In other terms, when adding the control we have on the expectation, we get

$$\mathbb{P}_{\mathcal{D}_n}\left(\sup_{g\in\mathcal{G}_{M,\sigma}} \mathcal{R}_S(\pi\circ g) - \mathcal{R}_{S,\mathcal{D}_n}(\pi\circ g) > t + M\left\|\varphi\right\|_\infty n^{-1/2}\right) \leq \exp\left(-2nt^2\right). \tag{6.22}$$

When $8^{-1}\left\||\eta|^{-1}\right\|_{p,\infty}^{-1} c_0\tilde{L}^{-\frac{d}{q}} \geq \left\|\varphi\right\|_\infty Mn^{-1/2}$, this leads to

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_n}\left(\mathcal{R}_S(\pi\circ g_{\mathcal{D}_n}) - \mathcal{R}_{S,\mathcal{D}_n}(g_{\mathcal{D}_n}) \geq 8^{-1}\left\||\eta|^{-1}\right\|_{p,\infty}^{-1} c_0\tilde{L}^{-\frac{d}{q}}\right) \\
&\leq \exp\left(-\frac{n}{8}\left(8^{-1}\left\||\eta|^{-1}\right\|_{p,\infty}^{-1} c_0\tilde{L}^{-\frac{d}{q}} - M\left\|\varphi\right\|_\infty n^{-1/2}\right)^2\right) \\
&\leq \exp\left(-\frac{c_0^2\sigma^{\frac{2d(p+1)}{p}}}{512\left\||\eta|^{-1}\right\|_{p,\infty}^2 (MG_\varphi)^{\frac{2d(p+1)}{p}}}\cdot n + \frac{c_0\left\|\varphi\right\|_\infty}{32\left\||\eta|^{-1}\right\|_{p,\infty} M^{\frac{d(p+1)}{p}-1} G_\varphi^{\frac{d(p+1)}{p}}}\cdot n^{-1/2} - \frac{M^2\left\|\varphi\right\|_\infty^2}{8}\right).
\end{aligned}$$

Regarding the second term, we can use classical concentration of $\mathcal{R}_{\mathcal{D}_n}(g_{M,\sigma})$ around its mean. For example, using the fact that Assumption 12 implies that $\rho_{\mathcal{X}}$ is compact, and using the fact that L and $g_{\sigma,M}$ are Lipschitz, we deduce that $L(g_{\sigma,M},Y)$ is bounded, hence one can applies Hoeffding's inequality to get the same type of exponential control on this term.

The result follows form those concentration inequality and the fact that $\mathcal{R}$ is bounded by one and that any $\min(1, a\exp(-bn))$ for $a,b>0$ and $n>1$ can be bounded by $\exp(-cn)$ for a $c>0$).

6.B Experimental details

In our experiments, we used the SVM implementation of Chang and Lin (2011) through its *Scikit-learn* wrapper (Pedregosa et al., 2011) in *Python*. We used *Numpy* (Harris et al., 2020) to reduce our work to high-level array instructions, and *Matplotlib* for visualization (Hunter, 2007). Randomness in experiments was controlled with the random seed provided by *Numpy*, which we initialized at zero.

Figures 6.2 and 6.5 are derived by averaging 100 trials of the following procedure. We draw uniformly at random n independent samples uniformly distributed on $\mathcal{X} \in \{[-1,1],[-1,-.1]\cup[.1,1]\}$. We draw randomly one output y_i for each input x_i, according to $\eta(x_i)$. We consider the Gaussian kernel $k(x,x') = \exp(-\left\|x-x'\right\|^2/2\sigma^2)$ for $\sigma = .2$, and solve the empirical risk minimization associated to the hinge loss with the penalization $\lambda\left\|\theta\right\|^2$ (rather than the hard constraint $\left\|\theta\right\| < M$) for $\lambda = 10^{-4}$. The generalization error is measured through the formula $\mathbb{E}[\left\|\eta(x)\right\|\mathbf{1}_{f(x)\neq f^*(x)}]$, with an empirical approximation of this sum with the points $(x_i)_{i\leq n}$ chosen such that $\rho_{\mathcal{X}}([x_i,x_{i+1}]) = 1/n$ and $\rho_{\mathcal{X}}([x_n,+\infty)) = 1/n$, with $n = 10^4$ (which makes sure that the exponential behavior observed is not due to the lack of testing samples). For each x, the height of each dark part corresponds to one standard deviation of the generalization error computed from the 100 trials, and the solid line corresponds to the empirical average. The fact that the dark parts are not centered around the averages is due to the fact that we have drawn log-plots but centered the interval for linearly-scaled plots.

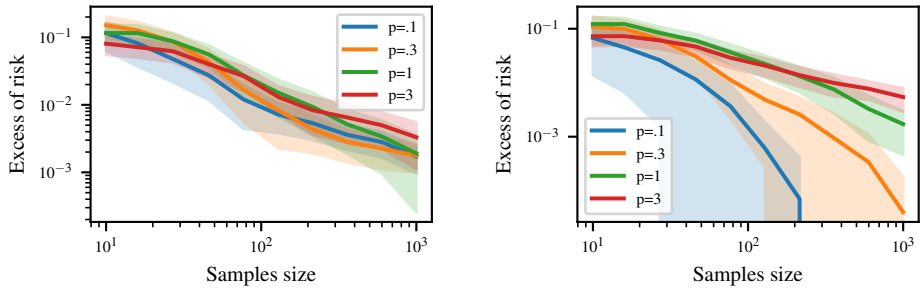

Figure 6.5: (Left) Similar setting as Figure 6.2 but with X uniform on $[-1, 1]$. The behavior of the excess of risk is quite different without the separation in $\mathcal{X}$: no exponential convergence rate is kicking in after a thousand of samples. (Right) Similar setting as Figure 6.2, using kernel ridge regression with the least-squares surrogate. Exponential convergence rates are observed with a slight delay compared to the hinge loss, and are explained by the hard-margin condition 10.

Figure 6.3 is obtained by considering $\mathcal{X} = [0, 1]^2$ with uniform input distribution, the Gaussian kernel with $\sigma = .2$, and the penalty parameter $\lambda = 10^{-3}$ (instead of a hard constraint leading to a parameter M as in the main text derivations). We take $n = 10^4 = 100^2$ points uniformly spread out on $\mathcal{X}$ (on the regular lattice $\frac{1}{\sqrt{n}} \cdot \mathbb{Z}^2 \cap \mathcal{X}$) to approximate $g_{\lambda,\sigma}$ with empirical risk minimization on this curated dataset. We consider $\eta(x) = \pi_{[-1,1]}(2x_2 - .5\sin(2\pi x_1) - 1)$, and assign to each x in the dataset a sample $(x, 1)$ weighted by $\mathbb{P}(Y = 1 \mid X = x) = (\eta(x) - 1)/2$, and a sample $(x, -1)$, weighted by $\mathbb{P}(Y = -1 \mid X = x)$. The "noiseless" setting denotes the setting where $(Y \mid X)$ is deterministic, but with the same decision frontier between the classes $\mathcal{X}_1$ and $\mathcal{X}_{-1}$ characterized by $\{(x, .5 + .25\sin(2\pi x)) \mid x \in [0, 1]\}$. Once we fit the support vector machine with this dataset, we test it with $n = 2.5 \cdot 10^5 = 500^2$ data points uniformly spread out on $\mathcal{X}$, and use *Matplotlib* to automatically draw level lines.

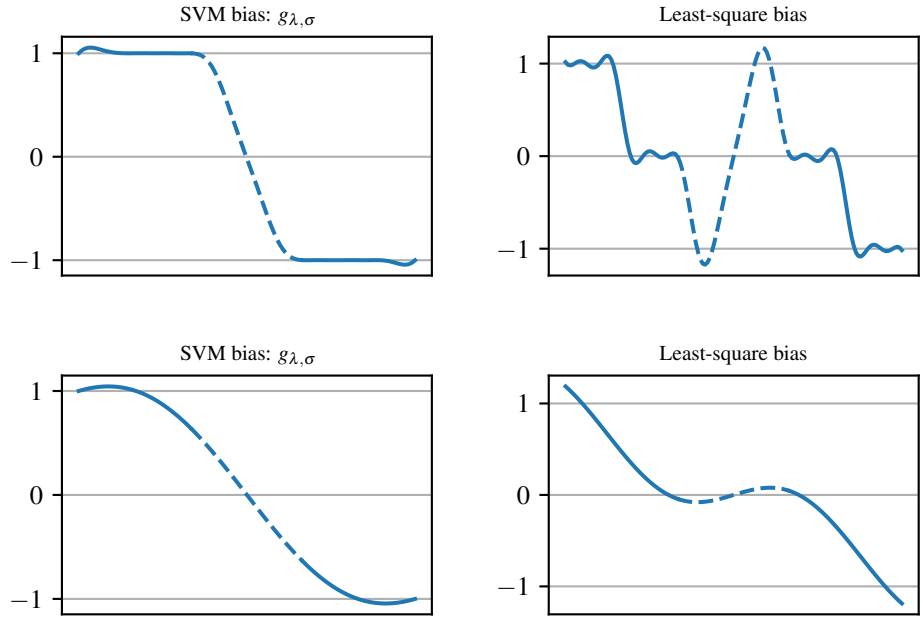

Figure 6.6: Same setting as Figure 6.4, with $\sigma = .2$ and $\lambda = 10^{-6}$ (top), and with $\sigma = 1$ and $\lambda = 10^{-3}$ (bottom).

Figures 6.4 and 6.6 correspond to $\mathcal{X} = [0, 3]$ with the input distribution uniform on $[0, 1] \cup [2, 3]$. Figure 6.4 is obtained with $\sigma = .1$ and $\lambda = 10^{-6}$. We derive it by considering $n = 100$ points uniformly spread out on the domain of η, solving the equivalent curated empirical risk minimization, that approximates

both

$$g_{\lambda,\sigma} = \underset{g:\mathcal{X}\to\mathbb{R}}{\arg\min}\, \mathbb{E}_\rho[(0, 1 - Y\left\langle \theta, \varphi\left(\frac{x}{\sigma}\right)\right\rangle)_+] + \lambda \|\theta\|^2 , \tag{6.23}$$

$$g_{(\mathrm{LS})} = \underset{g:\mathcal{X}\to\mathbb{R}}{\arg\min}\, \mathbb{E}_\rho[\left\|\left\langle \theta, \varphi\left(\frac{x}{\sigma}\right)\right\rangle - Y\right\|^2] + \lambda \|\theta\|^2 . \tag{6.24}$$

The robustness of SVM might be understood from its geometrical definition: when trying to find the maximum separating margin, infinitesimal modifications that change the regularity properties of η do not really matter. The picture is different for the least-squares surrogate with kernel methods, where from few point evaluations, the system reconstructs a function by assuming regularity and inferring information on high-order derivatives. This is similar to the Runge phenomenon with Hermite interpolation. More precisely, the Gaussian kernel is linked to a space of functions with rapidly decreasing Fourier coefficients (see, for example, Bach, 2023, for a more precise link). The function η that needs to be approximated on Figure 6.4 is similar to the Heaviside function, whose Fourier coefficients are of the form $(\frac{1}{i\pi k})_{k\in\mathbb{N}^*}$ and do not decrease fast enough to be all reconstructed. This leads to some high-frequency oscillations missing in the reconstruction as it appears on Figure 6.4.

Part III

Learning with Partial Supervision

Chapter 7

Infimum Loss

The following is a reproduction of Cabannes et al. (2020b).

Annotating datasets is one of the main costs nowadays in supervised learning. The goal of weak supervision is to enable models to learn while using only forms of labeling which are cheaper to collect, as partial labeling. This is a type of incomplete annotation where, for each data point, supervision is cast as a set of labels containing the real one. The problem of supervised learning with partial labeling has been studied for specific instances such as classification, multi-label, ranking or segmentation, but a general framework is still missing. This paper provides a unified framework based on structured prediction and on the concept of *infimum loss* to deal with partial labeling over a wide family of learning problems and loss functions. The framework leads naturally to explicit algorithms that can be easily implemented and for which proved statistical consistency and learning rates. Experiments confirm the superiority of the proposed approach over commonly used baselines.

7.1 Introduction

Fully supervised learning demands tight supervision of large amounts of data, a supervision that can be quite costly to acquire and constrains the scope of applications. To overcome this bottleneck, the machine learning community is seeking to incorporate weaker sources of information in the learning framework. In this paper, we address those limitations through partial labeling: *e.g.*, giving only partial ordering when learning user preferences over items, or providing the label "flower" for a picture of Arum Lilies, instead of spending a consequent amount of time to find the exact taxonomy.

Partial labeling has been studied in the context of classification (Cour et al., 2011; Nguyen and Caruana, 2008), multilabeling (Yu et al., 2014), ranking (Hüllermeier et al., 2008; Korba et al., 2018), as well as segmentation (Verbeek and Triggs, 2008; Papandreou et al., 2015), however a generic framework is still missing. Such a framework is a crucial step toward understanding how to learn from weaker sources of information, and widening the spectrum of machine learning beyond rigid applications of supervised learning. Some interesting directions are provided by Cid-Sueiro et al. (2014); van Rooyen and Williamson (2017), to recover the information lost in a corrupt acquisition of labels. Yet, they assume that the corruption process is known, which is a strong requirement that we want to relax.

In this paper, we make the following contributions:

- We provide a principled framework to solve the problem of learning with partial labeling, via *structured prediction*. This approach naturally leads to a variational framework built on the *infimum loss*.
- We prove that the proposed framework is able to recover the original solution of the supervised learning problem under identifiability assumptions on the labeling process.
- We derive an explicit algorithm which is easy to train and with strong theoretical guarantees. In particular, we prove that it is consistent, and we provide generalization error rates.
- Finally, we test our method against some simple baselines, on synthetic and real examples. We show that for certain partial labeling scenarios with symmetries, our infimum loss performs similarly to a simple baseline. However, in scenarios where the acquisition process of the labels is more adversarial in nature, the proposed algorithm performs consistently better.

7.2 Partial labeling with infimum loss

In this section, we introduce a statistical framework for partial labeling, and we show that it is characterized naturally in terms of risk minimization with the infimum loss. First, let's recall some elements of fully supervised and weakly supervised learning.

Fully supervised learning consists in learning a function $f \in \mathcal{Y}^{\mathcal{X}}$ between an input space $\mathcal{X}$ and an output space $\mathcal{Y}$, given a joint distribution $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X} \times \mathcal{Y}$, and a loss function $\ell \in \mathbb{R}^{\mathcal{Y}\times\mathcal{Y}}$, that minimizes the risk

$$\mathcal{R}(f;\rho) = \mathbb{E}_{(X,Y)\sim\rho}\left[\ell(f(X),Y)\right], \tag{7.1}$$

given observations $(x_i, y_i)_{i\leq n} \sim \rho^{\otimes n}$. We will assume that the loss ℓ is proper, *i.e.* it is continuous non-negative and is zero on, and only on, the diagonal of $\mathcal{Y} \times \mathcal{Y}$, and strictly positive outside. We will also assume that $\mathcal{Y}$ is compact.

In *weakly supervised learning*, given $(x_i)_{i\leq n}$, one does not have direct observations of $(y_i)_{i\leq n}$ but weaker information. The goal is still to recover the solution $f \in \mathcal{Y}^{\mathcal{X}}$ of the fully supervised problem (7.1). In *partial labeling*, also known as *superset learning* or as *learning with ambiguous labels*, which is an instance of weak supervision, information is cast as closed sets $(S_i)_{i\leq n}$ in $\mathcal{S}$, where $\mathcal{S} \subset 2^{\mathcal{Y}}$ is the space of closed subsets of $\mathcal{Y}$, containing the true labels $(y_i \in S_i)$. In this paper, we model this scenario by considering a data distribution $\tau \in \Delta_{\mathcal{X}\times\mathcal{S}}$, that generates the samples (x_i, S_i). We will denote τ as *weak distribution* to distinguish it from ρ. Capturing the dependence on the original problem, τ must be compatible with ρ, a matching property that we formalize with the concept of eligibility.

Definition 33 (Eligibility)**.** *Given a probability measure τ on $\mathcal{X} \times \mathcal{S}$, a probability measure ρ on $\mathcal{X} \times \mathcal{Y}$ is said to be eligible for τ (denoted by $\rho \vdash \tau$), if there exists a probability measure π over $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$ such that ρ is the marginal of π over $\mathcal{X} \times \mathcal{Y}$, τ is the marginal of π over $\mathcal{X} \times \mathcal{S}$, and, for $y \in \mathcal{Y}$ and $S \in \mathcal{S}$*

$$y \notin S \qquad \Rightarrow \qquad \mathbb{P}_\pi\left(S \,|\, Y = y\right) = 0.$$

We will alternatively say that τ is a weakening *of ρ, or that ρ and τ are* compatible*.*

7.2.1 Disambiguation principle

According to the setting described above, the problem of partial labeling is completely defined by a loss and a weak distribution (ℓ, τ). The goal is to recover the solution of the original supervised learning problem in (7.1) assuming that the original distribution verifies $\rho \vdash \tau$. Since more than one ρ may be eligible for τ, we would like to introduce a guiding principle to identify a $\rho^\star$ among them. With this goal we define the concept of *non-ambiguity* for τ, a setting in which a natural choice for $\rho^\star$ appears.

Definition 34 (Non-ambiguity)**.** *For any $x \in \mathcal{X}$, denote by $\tau|_x$ the conditional probability of τ given x, and define the set S_x as*

$$S_x = \bigcap_{S\in\mathrm{supp}(\tau|_x)} S.$$

The weak distribution τ is said non-ambiguous *if, for every $x \in \mathcal{X}$, S_x is a singleton. Moreover, we say that τ is* strictly non-ambiguous *if it is non-ambiguous and there exists $\eta \in (0, 1)$ such that, for all $x \in \mathcal{X}$ and $z \notin S_x$*

$$\mathbb{P}_{S\sim\tau|_x}(z \in S) \leq 1 - \eta.$$

This concept is similar to the one by Cour et al. (2011), but more subtle because this quantity only depends on τ, and makes no assumption on the original distribution ρ describing the fully supervised process that we can not access. In this sense, it is also more general.

When τ is non-ambiguous, we can write $S_x = \{y_x\}$ for any x, where y_x is the only element of S_x. In this case it is natural to identify $\rho^\star$ as the one satisfying $\rho^\star|_x = \delta_{y_x}$. Actually, such a $\rho^\star$ is characterized without S_x as the only deterministic distribution that is eligible for τ. Because deterministic distributions are characterized as minimizing the minimum risk (7.1), we introduce the following *minimum variability principle* to disambiguate between all eligible ρ's, and identify $\rho^\star$,

$$\rho^\star \in \operatorname*{arg\,min}_{\rho\vdash\tau} \mathcal{E}(\rho), \qquad \mathcal{E}(\rho) = \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f;\rho). \tag{7.2}$$

The quantity $\mathcal{E}$ can be identified as a variance, since if f_ρ is the minimizer of $\mathcal{R}(f;\rho)$, $f_\rho(x)$ can be seen as the mean of $\rho|_x$ and ℓ the natural distance in $\mathcal{Y}$. Indeed, when $\ell = \ell_2$ is the mean square loss, this is exactly the case. The principle above recovers exactly $\rho^\star|_x = \delta_{y_x}$, when τ is non-ambiguous, as stated by Theorem 35, proven in Appendix 7.A.1.

Proposition 35 (Non-ambiguity determinism)**.** *When τ is non-ambiguous, the solution $\rho^\star$ (7.2) exists and satisfies that, for any $x \in \mathcal{X}$, $\rho^\star|_x = \delta_{y_x}$, where y_x is the only element of S_x.*

Proposition 35 provides a justification for the usage of the minimum variability principle. Indeed, under non-ambiguity assumption, following this principle will allow us to build an algorithm that recovers the original fully supervised distribution. Therefore, given samples (x_i, S_i), it is of interest to test if τ is non-ambiguous. Such tests should leverage other regularity hypotheses on τ, which we will not address in this work.

Now, we characterize the minimum variability principle in terms of a variational optimization problem that we can tackle in Section 7.3 via empirical risk minimization.

7.2.2 Variational formulation via the infimum loss

Given a partial labeling problem (ℓ, τ), define the solutions based on the minimum variability principle as the functions minimizing the recovered risk

$$f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f;\rho^\star). \tag{7.3}$$

for $\rho^\star$ a distribution solving (7.2). As shown in Theorem 11 below, proven in Appendix 7.A.2, the proposed disambiguation paradigm naturally leads to a variational framework involving the *infimum loss*.

Theorem 11 (Infimum loss (*IL*))**.** *The functions f^* defined in* (7.3) *are characterized as*

$$f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}_S(f),$$

where the risk $\mathcal{R}_S$ is defined as

$$\mathcal{R}_S(f) = \mathbb{E}_{(X,S)\sim\tau}\left[L(f(X),S)\right], \tag{7.4}$$

and L is the infimum loss

$$L(z,S) = \inf_{y\in S} \ell(z,y). \tag{7.5}$$

The infimum loss, also known as the ambiguous loss (Luo and Orabona, 2010; Cour et al., 2011), or as the optimistic superset loss (Hüllermeier, 2014), captures the idea that, when given a set S, this set contains the good label y but also a lot of bad ones, that should not be taken into account when retrieving f. In other terms, f should only match the best guess in S. Indeed, if ℓ is seen as a distance, L is its natural extension to sets.

7.2.3 Recovery of the fully supervised solutions

In this subsection, we investigate the setting where an original fully supervised learning problem ρ_0 has been weakened due to incomplete labeling, leading to a weak distribution τ. The goal here is to understand under which conditions on τ and ℓ it is possible to recover the original fully supervised solution based on the infimum loss framework. Denote f_0 the function minimizing $\mathcal{R}(f;\rho_0)$. The theorem below, proven in Appendix 7.A.3, shows that under non-ambiguity and deterministic conditions, it is possible to fully recover the function f_0 also from τ.

Theorem 12 (Supervision recovery)**.** *For an instance (ℓ, ρ_0, τ) of the weakened supervised problem, if we denote by f_0 the minimizer of* (7.1)*, we have the under the conditions that (1) τ is not ambiguous (2) for all $x \in \mathcal{X}$, $S_x = \{f_0(x)\}$; the infimum loss recovers the original fully supervised solution,* i.e. *the f^* defined in* (7.3) *verifies $f^* = f_0$.*

Furthermore, when ρ_0 is deterministic and τ not ambiguous, the $\rho^\star$ defined in (7.2) *verifies $\rho^\star = \rho_0$.*

At a comprehensive level, this theorem states that under non-ambiguity of the partial labeling process, if the labels are a deterministic function of the inputs, the infimum loss framework makes it possible to recover the solution of the original fully supervised problem while only accessing weak labels. In the next subsection, we will investigate which is the relation between the two problems when dealing with an estimator f of f^*.

7.2.4 Comparison inequality

In the following, we want to characterize the error performed by $\mathcal{R}(f;\rho^\star)$ with respect to the error performed by $\mathcal{R}_S(f)$. This will be useful since, in the next section, we will provide an estimator for f^* based on structured prediction, that minimizes the risk $\mathcal{R}_S$. First, we introduce a measure of discrepancy for the loss function.

Definition 36 (Discrepancy of the loss ℓ)**.** *Given a loss function ℓ, the* discrepancy degree *ν of ℓ is defined as*

$$\nu = \log \sup_{y,z' \neq z} \frac{\ell(z,y)}{\ell(z,z')}.$$

$\mathcal{Y}$ will be said discrete for ℓ when $\nu < +\infty$, which is always the case when $\mathcal{Y}$ is finite.

Now we are ready to state the comparison inequality that generalizes a result on classification with the $0-1$ loss from Cour et al. (2011) to arbitrary losses and output spaces.

Proposition 37 (Comparison inequality)**.** *When $\mathcal{Y}$ is discrete and τ is strictly non-ambiguous for a given $\eta \in (0,1)$, then the following holds*

$$\mathcal{R}(f;\rho^\star) - \mathcal{R}(f^*;\rho^\star) \leq C(\mathcal{R}_S(f) - \mathcal{R}_S(f^*)), \tag{7.6}$$

for any measurable function $f \in \mathcal{Y}^\mathcal{X}$, where C does not depend on τ, f, and is defined as follows and always finite

$$C = \eta^{-1} e^{\nu}.$$

When ρ_0 is deterministic, since we know from Theorem 12 that $\rho^\star = \rho_0$, this theorem allows bounding the error made on the original fully supervised problem with the error measured with the infimum loss on the weakly supervised one.

Note that the constant presented above is the product of two independent terms, the first measuring the ambiguity of the weak distribution τ, and the second measuring a form of discrepancy for the loss. In the appendix, we provide a more refined bound for C, that is $C = C(\ell,\tau)$, that shows a more elaborated interaction between ℓ and τ. This may be interesting in situations where it is possible to control the labeling process and may suggest strategies to active partial labeling, with the goal of minimizing the costs of labeling while preserving the properties presented in this section and reducing the impact of the constant C in the learning process. An example is provided in the Appendix 7.A.5.

7.3 Consistent algorithm for partial labeling

In this section, we provide an algorithmic approach based on structured prediction to solve the weak supervised learning problem expressed in terms of infimum loss from Theorem 11. From this viewpoint, we could consider different structured prediction frameworks as structured SVM (Tsochantaridis et al., 2005), conditional random fields (Lafferty et al., 2001) or surrogate mean estimation (Ciliberto et al., 2016). For example, Luo and Orabona (2010) used a margin maximization formulation in a structured SVM fashion, Hüllermeier and Cheng (2015) went for nearest neighbors, and Cour et al. (2011) design a surrogate method specific to the 0-1 loss, for which they show consistency based on Bartlett et al. (2006).

In the following, we will use the structured prediction method of Ciliberto et al. (2016); Nowak-Vila et al. (2019), which allows us to derive an explicit estimator, easy to train and with strong theoretical properties, in particular, consistency and finite sample bounds for the generalization error. The estimator is based on the pointwise characterization of f^* as

$$f^*(x) \in \operatorname*{arg\,min}_{z \in \mathcal{Y}} \mathbb{E}_{S \sim \tau|_x} \left[\inf_{y \in S} \ell(z,y) \right],$$

and weights $\alpha_i(x)$ that are trained on the dataset such that $\hat{\tau}_{|x} = \sum_{i=1}^n \alpha_i(x) \delta_{S_i}$ is a good approximation of $\tau|_x$. Plugging this approximation in the precedent equation leads to our estimator, that is defined explicitly as follows

$$f_n(x) \in \operatorname*{arg\,min}_{z \in \mathcal{Y}} \inf_{y_i \in S_i} \sum_{i=1}^n \alpha_i(x) \ell(z, y_i). \tag{7.7}$$

Among possible choices for α, we will consider the following kernel ridge regression estimator to be learned at training time

$$\alpha(x) = (K + n\lambda)^{-1} v(x),$$

with $\lambda > 0$ a regularizer parameter and $K = (k(x_i, x_j))_{i,j} \in \mathbb{R}^{n \times n}, v(x) = (k(x, x_i))_i \in \mathbb{R}^n$ where $k \in \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive-definite kernel (Scholkopf and Smola, 2001) that defines a similarity function between input points (*e.g.*, if $\mathcal{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$ a commonly used kernel is the Gaussian kernel $k(x, x') = e^{-\|x-x'\|^2}$). Other choices can be done to learn α, beyond kernel methods, a particularly appealing one is harmonic functions, incorporating a prior on low density separation to boost learning (Zhu et al., 2003; Zhou et al., 2003; Bengio et al., 2006). Here we use the kernel estimator since it allows deriving strong theoretical results, based on kernel conditional mean estimation (Muandet et al., 2017).

7.3.1 Theoretical guarantees

In this following, we want to prove that f_n converges to f^* as n goes to infinity, and we want to quantify it with finite sample bounds. The intuition behind this result is that as the number of data points tends toward infinity, $\hat{\tau}$ concentrates toward τ, making our algorithm in (7.7) converging to a minimizer of (7.4) as explained more in detail in Appendix 7.A.6.

Theorem 13 (Consistency)**.** *Let $\mathcal{Y}$ be finite and τ be a non-ambiguous probability. Let k be a bounded continuous universal kernel,* e.g. *the Gaussian kernel (see Micchelli et al., 2006, for details), and f_n the estimator in* (7.7) *trained on $n \in \mathbb{N}$ examples and with $\lambda = n^{-1/2}$. Then, holds with probability* 1

$$\lim_{n \to \infty} \mathcal{R}(f_n; \rho^\star) = \mathcal{R}(f^*; \rho^\star).$$

In the next theorem, instead we want to quantify how fast f_n converges to f^* depending on the number of examples. To obtain this result, we need a finer characterization of the infimum loss L as:

$$L(z, S) = \langle \psi(z), \varphi(S) \rangle ,$$

where $\mathcal{H}$ is a Hilbert space and $\psi : \mathcal{Y} \to \mathcal{H}, \varphi : 2^{\mathcal{Y}} \to \mathcal{H}$ are suitable maps. Such a decomposition always exists in finite case (as for the infimum loss over $\mathcal{Y}$ finite) and many explicit examples for losses of interest are presented by Nowak-Vila et al. (2019). We now introduce the conditional expectation of $\varphi(S)$ given x, defined as

$$\begin{array}{rccc} g: & \mathcal{X} & \to & \mathcal{H} \\ & x & \to & \mathbb{E}_\tau \left[\varphi(S) \mid X = x \right]. \end{array}$$

The idea behind the proof is that the distance between f_n and f is bounded by the distance of g_n an estimator of g that is implicitly computed via α. If g has some form of regularity, *e.g.* $g \in \mathcal{G}$, with $\mathcal{G}$ the space of functions representable by the chosen kernel (see Scholkopf and Smola, 2001), then it is possible to derive explicit rates, as stated in the following theorem.

Theorem 14 (Convergence rates)**.** *In the setting of Theorem 13, if τ is η-strictly non-ambiguous for $\eta \in (0, 1)$, and if $g \in \mathcal{G}$, then there exists a $\tilde{C}$, such that, for any $\delta \in (0, 1)$ and $n \in \mathbb{N}$, holds with probability at least $1 - \delta$,*

$$\mathcal{R}(f_n; \rho^\star) - \mathcal{R}(f^*; \rho^\star) \leq \tilde{C} \log \left(\frac{8}{\delta} \right)^2 n^{-1/4}. \tag{7.8}$$

Those last two theorem are proven in Appendix 7.A.6 and combines the consistency and learning results for kernel ridge regression (Caponnetto and De Vito, 2006; Smale and Zhou, 2007), with a comparison inequality of Ciliberto et al. (2016) which relates the excess risk of the structured prediction problem with the one of the surrogate loss $\mathcal{R}_S$, together with our Theorem 37, which relates the error $\mathcal{R}$ to $\mathcal{R}_S$.

Those results make our algorithm the first algorithm for partial labeling that, to our knowledge, is applicable to a generic loss ℓ and has strong theoretical guarantees as consistency and learning rates. In the next section we will compare with the state of the art and other variational principles.

7.4 Previous works and baselines

Partial labeling was first approached through discriminative models, proposing to learn $(Y \mid X)$ among a family of parametrized distributions by maximizing the log likelihood based on expectation-maximization scheme (Jin and Ghahramani, 2002), eventually integrating knowledge on the partial labeling process (Grandvalet, 2002; Papandreou et al., 2015). In the meanwhile, some applications of clustering methods have involved special instances of partial labeling, like segmentation approached with spectral method (Weiss, 1999), semi-supervision approached with max-margin (Xu et al., 2004). Also, initially geared toward clustering, Bach and Harchaoui (2007) considered the infimum principle on the mean square loss, and this was generalized to weakly supervised problems (Joulin et al., 2010). The infimum loss as an objective to minimize when learning from partial labels was introduced by Cour et al. (2011) for the classification instance and used by Luo and Orabona (2010); Hüllermeier (2014) in generic cases. Compared to those last two, we provide a framework that derives the use of infimum loss from first principles and from which we derive an explicit and easy to train algorithm with strong statistical guarantees, which were missing in previous work. In the rest of the section, we will compare the infimum loss with other variational principles that have been considered in the literature, in particular the supremum loss (Guillaume et al., 2017) and the average loss (Denoeux, 2013).

Average loss (*AC*). A simple loss to deal with uncertainty is to average over all potential candidates, assuming S discrete,

$$L_{ac}(z, S) = \frac{1}{|S|} \sum_{y \in S} \ell(z, y).$$

It is equivalent to a fully supervised distribution ρ_{ac} by sampling Y uniformly at random among S

$$\rho_{ac}(y) = \int_{\mathcal{S}} \frac{1}{|S|} \mathbf{1}_{y \in S} \, \mathrm{d}\tau(S).$$

This directly follows from the definition of L_{ac} and of the risk $\mathcal{R}(z; \rho_{ac})$. However, as soon as the loss ℓ has discrepancy, *i.e.* $\nu > 0$, the average loss will implicitly advantage some labels, which can lead to inconsistency, even in the deterministic not ambiguous setting of Theorem 37 (see Appendix 7.A.7 for more details).

Supremum loss (*SP*). Another loss that has been considered is the supremum loss (Wald, 1945; Madry et al., 2018), bounding from above the fully supervised risk in (7.1). It is widely used in the context of robust risk minimization and reads

$$R_{sp}(f) = \sup_{\rho \vdash \tau} \mathbb{E}_{(X,Y) \sim \rho} \left[\ell(f(x), S) \right].$$

Similarly to the infimum loss in Theorem 11, this risk can be written from the loss function

$$L_{sp}(z, S) = \sup_{y \in S} \ell(z, y).$$

Yet, this adversarial approach is not consistent for partial labeling, even in the deterministic non-ambiguous setting of Theorem 37, since it finds the solution that best agrees with *all* the elements in S and not only the true one (see Appendix 7.A.7 for more details).

7.4.1 Instance showcasing superiority of our method

In the rest of this section, we consider a pointwise example to showcase the underlying dynamics of the different methods. It is illustrated in Figure 7.1. Consider $\mathcal{Y} = \{a, b, c\}$ and a proper symmetric loss function such that $\ell(a, b) = \ell(a, c) = 1$, $\ell(b, c) = 2$. The simplex $\Delta_{\mathcal{Y}}$ is naturally split into decision regions, for $e \in \mathcal{Y}$,

$$R_e = \left\{ \rho \in \Delta_{\mathcal{Y}} \,\middle|\, e \in \arg\min_{z \in \mathcal{Y}} \mathbb{E}_\rho[\ell(z, Y)] \right\}.$$

Both *IL* and *AC* solutions can be understood geometrically by looking at where $\rho^\star$ and ρ_{ac} fall in the partition of the simplex $(R_e)_{e \in \mathcal{Y}}$. Consider a fully supervised problem with distribution δ_c, and a weakening τ of ρ defined by $\tau(\{a, b, c\}) = \frac{5}{8}$ and $\tau(\{c\}) = \tau(\{a, c\}) = \tau(\{b, c\}) = \frac{1}{8}$. This distribution can be represented

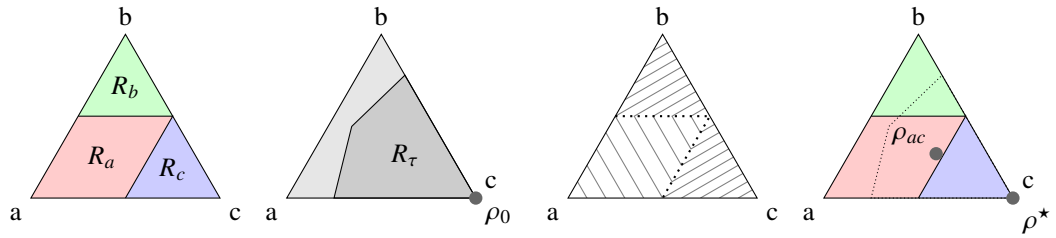

Figure 7.1: Simplex $\Delta_\mathcal{Y}$. (Left) Decision frontiers. (Middle left) Full and weak distributions. (Middle right) Level curves of the piecewise linear objective $\mathcal{E}$ (7.2), to optimize when disambiguating τ into $\rho^\star$. (Right) Disambiguation of *AC* and *IL*.

on the simplex in terms of the region $R_\tau = \{\rho \in \Delta_\mathcal{Y} \,|\, \rho \vdash \tau\}$. Finding $\rho^\star$ correspond to minimizing the piecewise linear function $\mathcal{E}(\rho)$ (7.2) inside R_τ. On this example, it is minimized for $\rho^\star = \delta_c$, which we know from Theorem 37. Now note that if we use the average loss, it disambiguates ρ as

$$\rho_{ac}(c) = \frac{11}{24} = \frac{1}{3}\frac{5}{8} + \frac{1}{8} + 2 \cdot \frac{1}{2}\frac{1}{8}, \quad \rho_{ac}(b) = \rho_{ac}(a) = \frac{13}{48}.$$

This distribution falls in the decision region of a, which is inconsistent with the real label $y = c$. For the supremum loss, one can show, based on $\mathcal{R}_{sp}(a) = \ell(a, c) = 1$, $\mathcal{R}_{sp}(b) = \ell(b, c) = 2$ and $\mathcal{R}_{sp}(c) = 3/2$, that the supremum loss is minimized for $z = a$, which is also inconsistent. Instead, by using the infimum loss, we have $f^* = f_0 = c$, and moreover that $\rho^\star = \rho_0$ that is the optimal one.

7.4.2 Algorithmic considerations for AC, SP

The averaging candidates principle, approached with the framework of quadratic surrogates (Ciliberto et al., 2016), leads to the following algorithm

$$\begin{aligned} f_{ac}(x) &\in \arg\min_{z\in\mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \frac{1}{|S_i|} \sum_{y\in S_i} \ell(z, y) \\ &= \arg\min_{z\in\mathcal{Y}} \sum_{y\in\mathcal{Y}} \left(\sum_{i=1}^{n} \mathbf{1}_{y\in S_i} \frac{\alpha_i(x)}{|S_i|} \right) \ell(z, y). \end{aligned}$$

This estimator is computationally attractive because the inference complexity is the same as the inference complexity of the original problem when approached with the same structured prediction estimator. Therefore, one can directly reuse algorithms developed to solve the original inference problem (Nowak-Vila et al., 2019). Finally, with a similar approach to the one in Section 7.3, we can derive the following algorithm for the supremum loss

$$f_{sp}(x) \in \arg\min_{z\in\mathcal{Y}} \sup_{y_i\in S_i} \sum_{i=1}^{n} \alpha_i(x) \ell(z, y_i).$$

In the next section, we will use the average candidates as baseline to compare with the algorithm proposed in this paper, as the supremum loss consistently performs worth, as it is not fitted for partial labeling.

7.5 Applications and experiments

In this section, we will apply (7.7) to some synthetic and real datasets from different prediction problems and compared with the average estimator presented in the section above, used as a baseline. Code is available online.[1]

[1] https://github.com/VivienCabannes/partial_labelling

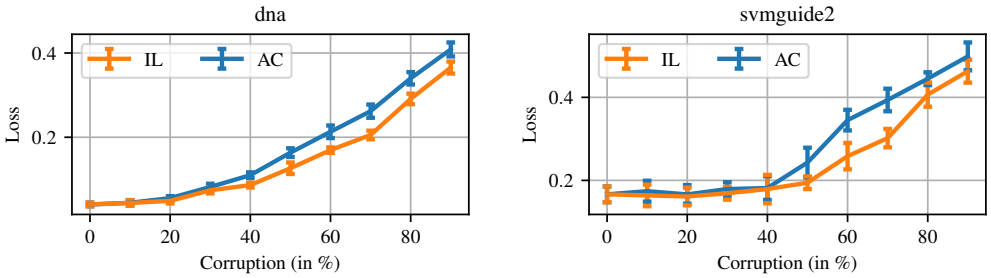

Figure 7.2: Classification. Testing risks (from (7.1)) achieved by *AC* and *IL* on the "DNA" and "svmguide2" datasets from *LIBSVM* as a function of corruption parameter c, when the corruption is as follows: for y being the most present labels of the dataset, and $z' \neq z$, $\mathbb{P}(z' \in S \,|\, Y = z) = c \cdot \mathbf{1}_{z=y}$. Plotted intervals show the standard deviation on eight-fold cross-validation. Experiments were done with the Gaussian kernel. See all experimental details in (7.B).

7.5.1 Classification

Classification consists in recognizing the most relevant item among m items. The output space is isomorphic to the set of indices $\mathcal{Y} = [\![1, m]\!]$, and the usual loss function is the 0-1 loss

$$\ell(z, y) = \mathbf{1}_{y \neq z}.$$

It has already been widely studied with several approaches that are calibrated in non-ambiguous deterministic settings, notably by Cour et al. (2011). The infimum loss reads $L(z, S) = \mathbf{1}_{z \notin S}$, and its risk in (7.4) is minimized for

$$f(x) \in \arg\max_{z \in \mathcal{Y}} \mathbb{P}(z \in S \,|\, X = x).$$

Based on data $(x_i, S_i)_{i \leq n}$, our estimator (7.7) reads

$$f_n(x) = \arg\max_{z \in \mathcal{Y}} \sum_{i; z \in S_i} \alpha_i(x).$$

For this instance, the supremum loss is really conservative, only learning from set that are singletons $L_{sp}(z, S) = \mathbf{1}_{S \neq \{z\}}$, while the average loss is similar to the infimum one, adding an evidence weight depending on the size of S, $L_{ac}(z, S) \simeq \mathbf{1}_{z \notin S}/|S|$.

Real data experiment. To compare *IL* and *AC*, we used *LIBSVM* datasets (Chang and Lin, 2011) on which we corrupted labels to simulate partial labeling. When the corruption is uniform, the two methods perform the same. Yet, when labels are unbalanced, such as in the "DNA" and "svmguide2" datasets, and we only corrupt the most frequent label $y \in \mathcal{Y}$, the infimum loss performs better as shown in Figure 7.2.

7.5.2 Ranking

Ranking consists in ordering m items based on an input x that is often the conjunction of a user u and a query q, $(x = (u, q))$. An ordering can be thought of as a permutation, that is, $\mathcal{Y} = \mathfrak{S}_m$. While designing a loss for ranking is intrinsically linked to a voting system (Arrow, 1950), making it a fundamentally hard problem; Kemeny (1959) suggested approaching it through pairwise disagreement, which is current machine learning standard (Duchi et al., 2010), leading to the Kendall embedding

$$\varphi(y) = \big(\operatorname{sign}(y_i - y_j)\big)_{i<j \leq m},$$

and the Kendall loss (Kendall, 1938), with $C = m(m-1)/2$

$$\ell(y, z) = C - \varphi(y)^T \varphi(z).$$

Supervision often comes as partial order on items, *e.g.*,

$$S = \big\{ y \in \mathfrak{S}_m \,\big|\, y_i > y_j > y_k, y_l > y_m \big\}.$$

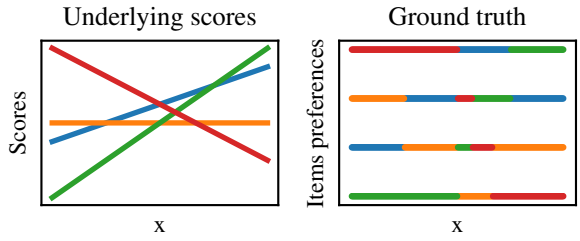

Figure 7.3: Ranking, experimental setting. Colors represent four different items to rank. Each item is associated with a utility function of x shown on the left figure. From those scores, an ordering y of the items is retrieved as represented on the right.

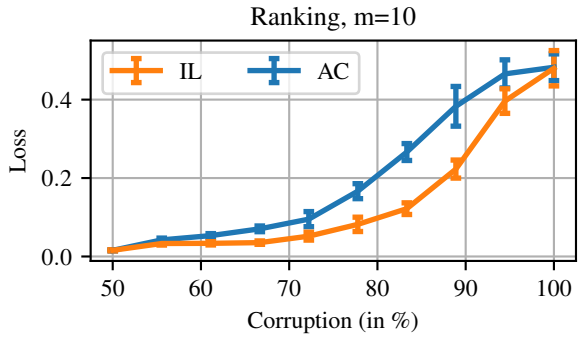

Figure 7.4: Ranking, results. Testing risks (from (7.1)) achieved by *AC* and *IL* as a function of corruption parameter c. When $c = 1$, both risks are similar at 0.5. The simulation setting is the same as in Figure 7.2. The error bars are defined as for Figure 7.2, after cross-validation over eight folds. *IL* clearly outperforms *AC*.

It corresponds to fixing some coordinates in the Kendall embedding. In this setting, *AC* and *SP* are not consistent, as one can recreate a similar situation to the one in Section 7.4, considering $m = 3$, $a = (1, 2, 3)$, $b = (2, 1, 3)$ and $c = (1, 3, 2)$ (permutations being represented with $(\sigma^{-1}(i))_{i \leq m}$), and supervision being most often $S = (1 > 3) = \{a, b, c\}$ and sometimes $S = (1 > 3 > 2) = \{c\}$.

Minimum feedback arc set. Dealing with Kendall's loss requires solving problem of the form,

$$\arg\min_{y \in S} \langle c, \varphi(y) \rangle ,$$

for $c \in \mathbb{R}^{m^2}$, and constraints due to partial ordering encoded in $S \subset \mathcal{Y}$. This problem is an instance of the constrained minimum feedback arc set problem. We provide a simple heuristic to solve it in Appendix 7.C, which consists of approaching it as an integer linear program. Such heuristics are analyzed and refined for analysis purposes by Ailon et al. (2005); van Zuylen et al. (2007).

Algorithm specification. At inference, the infimum loss requires solving:

$$f_n(x) = \arg\max_{z \in \mathcal{Y}} \sup_{(y_i) \in S_i} \sum_{i=1}^{n} \alpha_i(x) \, \langle \varphi(z), \varphi(y_i) \rangle . \tag{7.7}$$

It can be approached with alternate minimization, initializing $\varphi(y_i) \in \mathrm{Conv}(\varphi(S_i))$, by putting 0 on unseen observed pairwise comparisons, then, iteratively, solving a minimum feedback arc set problem in z, then solving several minimum feedback arc set problems with the same objective, but different constraints in (y_i). This is done efficiently using warm start on the dual simplex algorithm.

Synthetic experiments. Let us consider $\mathcal{X} = [0, 1]$ embodying some input features. Let $\{1, \dots, m\}$, $m \in \mathbb{N}$ be abstract items to order, each item being linked to a utility function $v_i \in \mathbb{R}^{\mathcal{X}}$, that characterizes

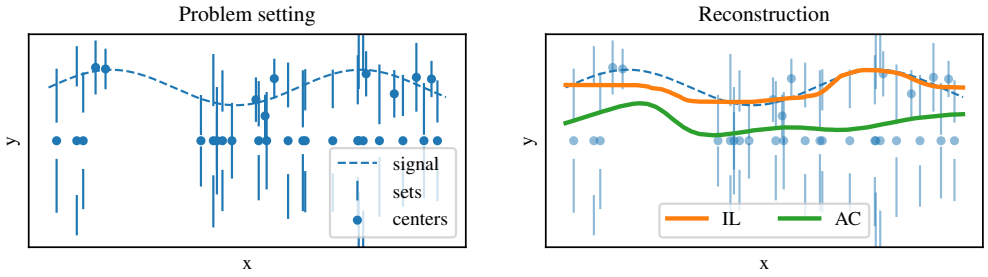

Figure 7.5: Partial regression on $\mathbb{R}$. In this setting we aim at recovering a signal $y(x)$ given upper and lower bounds on its amplitude, and in thirty percent of case, information on its phase, or equivalently in $\mathbb{R}$, its sign. *IL* clearly outperforms the baseline. Indeed, *AC* is a particularly ill-fitted method on such a problem, since it regresses on the barycenter of the resulting sets.

the value of i for x as $v_i(x)$. Labels $y(x) \in \mathcal{Y}$ are retrieved by sorting $(v_i(x))_{i \leq m}$. To simulate a problem instance, we set v_i as $v_i(x) = a_i \cdot x + b_i$, where a_i and b_i follow a standard normal distribution. Such a setting is illustrated in Figure 7.3.

After sampling x uniformly on $[0, 1]$ and retrieving the ordering y based on scores, we simulate partial labeling by randomly losing pairwise comparisons. The comparisons are formally defined as coordinates of the Kendall's embedding $(\varphi(y)_{jk})_{jk \leq m}$. To create non-symmetric perturbations we corrupt more often items whose scores differ a lot. In other words, we suppose that the partial labeling focuses on pairs that are hard to discriminate. The corruption is set upon a parameter $c \in [0, 1]$. In fact, for $m = 10$, until $c = 0.5$, our corruption is fruitless since it can most often be inverted based on transitivity constraints in ordering, while the problem becomes non-trivial with $c \geq 0.5$. In the latter setting, *IL* clearly outperforms *AC* on Figure 7.4.

7.5.3 Partial regression

Partial regression is an example of non-discrete partial labeling problem, where $\mathcal{Y} = \mathbb{R}^m$ and the usual loss is the Euclidean distance

$$\ell(y, z) = \|y - z\|^2 .$$

This partial labeling problem consists of regression where observations are sets $S \subset \mathbb{R}^m$ that contain the true output y instead of y. Among others, it arises for example in economical models, where bounds are preferred over approximation when acquiring training labels (Tobin, 1958). As an example, we will illustrate how partial regression could appear for some phase problems arising with physical measurements. Suppose a physicist wants to measure the law between a vector quantity Y and some input parameters X. Suppose that, while she can record the input parameters x, her sensors do not exactly measure y but render an interval in which the amplitude $\|y\|$ lays and only occasionally render its phase $y/\|y\|$, in a fashion that leads to a set of candidates S for y. The geometry over ℓ^2 makes it a perfect example to showcase superiority of the infimum loss as illustrated in Figure 7.5.

In this figure, we consider $\mathcal{Y} = \mathbb{R}$ and suppose that Y is a deterministic function of X as shown by the dotted blue line signal. If, for a given x_i, measurements only provides that $|y_i| \in [1, 2]$ without the sign of y_i, a situation where the phase is lost, this corresponds to the set $S_i = [-2, -1] \cup [1, 2]$, explaining the shape of observed sets that are symmetric around the origin. Whenever the acquired data has no phase, which happens seventy percent of the time in our simulation, AC will target the set centers, explaining the green curve. On the other hand, IL is aiming at passing by each set, which explains the orange curve, crossing all blue bars.

7.6 Conclusions

In this paper, we deal with the problem of weakly supervised learning, beyond standard regression and classification, focusing on the more general case of arbitrary loss functions and structured prediction. We provide a principled framework to solve the problem of learning with partial labeling, from which a natural variational approach based on the infimum loss is derived. We prove that under some identifiability assumptions on the labeling process the framework is able to recover the solution of the original supervised learning problem. The resulting algorithm is easy to train and with strong theoretical guarantees. In particular,

we prove that it is consistent, and we provide generalization error rates. Finally, the algorithm is tested on simulated and real datasets, showing that when the acquisition process of the labels is more adversarial in nature, the proposed algorithm performs consistently better than baselines. This paper focuses on the problem of partial labeling, however the resulting mathematical framework is quite flexible in nature, and it is interesting to explore the possibility to extend it to tackle also other weakly supervised problems, as imprecise labels from non-experts (Dawid and Skene, 1979), more general constraints over the set $(y_i)_{i\leq n}$ (Quadrianto et al., 2009) or semi-supervision (Chapelle et al., 2006).

Appendix

7.A Proofs

In the paper, we have implicitly considered $\mathcal{X}, \mathcal{Y}$ separable and completely metrizable topological spaces, *i.e.* Polish spaces, which allows considering probabilities. Moreover, we assumed that $\mathcal{Y}$ is compact, to have minimizers well-defined. The observation space was considered to be the set of closed subsets of $\mathcal{Y}$ endowed with the Hausdorff distance, $\mathcal{S} = \mathrm{Cl}(\mathcal{Y}), d_H$. As such, $\mathcal{S}$ is also a Polish metric space, inheriting this property from $\mathcal{Y}$ (Beer, 1993). In the following, we will show that the closeness of sets is important in order to switch from the minimum variability principle to the infimum loss.

In terms of notations, we use the simplex notation $\Delta_{\mathcal{A}}$ to denote the space of Borel probability measures over the space $\mathcal{A}$. In particular, $\Delta_{\mathcal{X}\times\mathcal{Y}}$, $\Delta_{\mathcal{X}\times\mathcal{S}}$ and $\Delta_{\mathcal{X}\times\mathcal{Y}\times\mathcal{S}}$ are endowed with the weak-* topology and are Polish, inheriting the properties from original spaces (Aliprantis and Border, 2006). The fact that such spaces are Polish allows defining the conditional probabilities given $x \in \mathcal{X}$. We will denote this conditional probability $\rho|_x$ when, for example, $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$. Finally, we will denote by $\rho_{\mathcal{X}}$ the marginal distributions of ρ over $\mathcal{X}$.

Before diving into proofs, we would like to point out that many of our results are pointwise results. At an intuitive level, we only leverage the structure of the loss on the output space and aggregate those results over $\mathcal{X}$.

Remark 38 (Going pointwise)**.** *The learning frameworks in* (7.1)*,* (7.2) *and* (7.4) *are* pointwise separable *as their solutions can be written as aggregation of pointwise solutions (Devroye et al., 1996). More exactly, the partial labeling risk (and similarly the fully supervised one) can be expressed as*

$$\mathcal{R}_S(f) = \mathbb{E}_X\left[\mathcal{R}_{S,X}(f(X))\right],$$

where the conditional risk reads,

$$\mathcal{R}_{S,x}(z) = \mathbb{E}_{S\sim\tau|_x}\left[L(z,S)\right],$$

with $\tau|_x$ the conditional distribution of $(S \mid X = x)$. Thus, minimizing $\mathcal{R}_S$ globally for $f \in \mathcal{Y}^{\mathcal{X}}$ is equivalent to minimizing locally $\mathcal{R}_{S,x}$ for $f(x)$ for almost all x. Similarly, for (7.2)*,*

$$\mathcal{E}(\rho) = \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_\rho\left[\ell(f(X),Y)\right] = \mathbb{E}_X\left[\inf_{z\in\mathcal{Y}} \mathbb{E}_{Y\sim\rho|_x}\left[\ell(z,Y) \mid X = x\right]\right].$$

Therefore, studies on risk can be done pointwise on instances $(\ell, \rho|_x, \tau|_x)$, before integrating along $\mathcal{X}$. Actually, Theorems 35, 11, 12 and 37 are pointwise results.

7.A.1 Proof of Theorem 35

Here we want to prove that when τ is non-ambiguous, then it is possible to define an optimal $\rho^\star$ that is deterministic on $\mathcal{Y}$, and that this $\rho^\star$ is characterized by solving (7.2).

Lemma 39. *When τ is non-ambiguous, and there is one, and only one, deterministic distribution eligible for τ. More exactly, if we write, for any $x \in \mathcal{X}$ in the support of $\tau_{\mathcal{X}}$, based on Definition 34, $S_x = \{y_x\}$, then this deterministic distribution is characterized as $\rho|_x = \delta_{y_x}$ almost everywhere.*

Proof. Let us consider a probability measure $\tau \in \Delta_{\mathcal{X}\times\mathcal{S}}$. We begin by working on the concept of eligibility. Consider $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ eligible for τ and a suitable π as defined in Definition 33. First of all, the condition that, for $y \in S$, $\mathbb{P}_\pi(S \mid Y = y) = 0$, can be stated formally in terms of measure as

$$\pi(\{(x,y,S) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{S} \mid y \notin S\}) = 0,$$

from which we deduced that, for $y \in \mathcal{Y}$ and $x \in \mathcal{X}$,

$$\begin{aligned}\rho|_x(y) = \pi|_x(\{y\} \times \mathcal{S}) &= \pi|_x(\{y\} \times \{S \in \mathcal{S} \mid y \in S\}) \\ &\leq \pi|_x(\mathcal{Y} \times \{S \in \mathcal{S} \mid y \in S\}) = \tau|_x(\{S \in \mathcal{S} \mid y \in S\}).\end{aligned}$$

It follows that when ρ is deterministic, if we write $\rho|_x = \delta_{y_x}$, then we have $\tau|_x(\{S \in \mathcal{S} \mid y_x \in S\}) = 1$, which means that y_x is in all sets that are in the support of $\tau|_x$, or that, using notations of Definition 34, $y_x \in S_x$. So far, we have proved that if there exists a deterministic distribution, $\rho|_x = \delta_{y_x}$, that is eligible for $\tau|_x$, we have $y_x \in S_x$. Reciprocally, one can do the reverse derivations, to show that if $\rho|_x = \delta_{y_x}$, with $y_x \in S_x$, for all $x \in \mathcal{X}$, then ρ is eligible for τ When τ is non-ambiguous, S_x is a singleton and therefore, there could be only one deterministic eligible distribution for τ, that is characterized in the lemma. □

Now we use the characterization of deterministic distribution through the minimization of the risk (7.1).

Lemma 40 (Deterministic characterization)**.** *When $\mathcal{Y}$ is compact and ℓ proper, deterministic distribution are exactly characterized by minimum variability* (7.2) *as*

$$\mathcal{E}(\rho) = \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_\rho\left[\ell(f(X), Y)\right] = 0.$$

Proof. Let's consider $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$, because $\mathcal{Y}$ is compact and ℓ continuous, we can consider f_ρ a minimizer of $\mathcal{R}(f;\rho)$. Let's now suppose that $\mathcal{R}(f_\rho;\rho) = 0$, since ℓ is non-negative, it means that almost everywhere

$$\mathbb{E}_{Y\sim\rho|_x}\left[\ell(f_\rho(x), Y)\right] = 0.$$

Suppose that $\rho|_x$ is not deterministic, then there is at least two points y and z in $\mathcal{Y}$ in its support, then, because ℓ is proper, we come to the absurd conclusion that

$$\mathbb{E}_{Y\sim\rho|_x}\left[\ell(f_\rho(x), Y)\right] \geq \rho|_x(y)\ell(f_\rho(x), y) + \rho|_x(z)\ell(f_\rho(x), z) > 0.$$

So $\mathcal{R}(f_\rho;\rho) = 0$ implies that ρ is deterministic. Reciprocally, when ρ is deterministic it is easy to show that the risk is minimized at zero. □

7.A.2 Proof of Theorem 11

At a comprehensive level, Theorem 11 is composed of two parts:

- A double minimum switch, to take the minimum over ρ before the minimum over f, and for which we need some compactness assumption to consider the joint minimum.
- A minimum-expectation switch, to take the minimum over $\rho \vdash \tau$ as a minimum $y \in S$ before the expectation to compute the risk, and for which we need some measure properties.

We begin with the minimum-expectation switch. To proceed with derivations, we need first to reformulate the concept of eligibility in Definition 33 in terms of measures.

Lemma 41 (Measure eligibility)**.** *Given a probability τ over $\mathcal{X} \times \mathcal{S}$, the space of probabilities over $\mathcal{X} \times \mathcal{Y}$ satisfying $\rho \vdash \tau$ is characterized by all probability measures of the form*

$$\rho(C) = \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{S}} \mathbf{1}_C(x, y)\, \mathrm{d}\pi|_{x,S}(y)\, \mathrm{d}\tau(x, S),$$

for any C a closed subset of $\mathcal{X} \times \mathcal{Y}$, and where π is a probability measure over $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$ that satisfies $\pi_{\mathcal{X}\times\mathcal{S}} = \tau$ and $\pi|_{x,S}(S) = 1$ for any (x, S) in the support of τ.

Proof. For any ρ that is eligible for τ there exists a suitable π on $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$ as specified by Definition 33. Actually, the set of π leading to an eligible $\rho := \pi_{\mathcal{X}\times\mathcal{Y}}$ is characterized by satisfying $\pi_{\mathcal{X}\times\mathcal{S}} = \tau$ and

$$\pi(\{(x, y, S) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{S} \mid y \notin S\}) = 0.$$

This last property can be reformulated with the complementary space as

$$\pi(\{(x, y, S) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{S} \mid y \in S\}) = 1,$$

which equivalently reads, that for any (x, S) in the support of τ, we have

$$\pi|_{x,S}(S) = \pi|_{x,S}(\{y \in \mathcal{Y} \mid y \in S\}) = 1.$$

Finally, using the conditional decomposition we have that, for C a closed subset of $\mathcal{X} \times \mathcal{Y}$

$$\rho(C) = \pi_{\mathcal{X}\times\mathcal{Y}}(C) = \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{S}} \mathbf{1}_C(x, y)\, \mathrm{d}\pi(x, y, S) = \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{S}} \mathbf{1}_C(x, y)\, \mathrm{d}\pi|_{x,S}(y)\, \mathrm{d}\pi_{\mathcal{X}\times\mathcal{S}}(x, S),$$

which ends the proof since $\tau = \pi_{\mathcal{X}\times\mathcal{S}}$. □

We are now ready to state the minimum-expectation switch.

Lemma 42 (Minimum-Expectation switch)**.** *For a probability measure $\tau \in \Delta_{\mathcal{X}\times\mathcal{S}}$, and measurable functions $\ell \in \mathbb{R}^{\mathcal{Y}\times\mathcal{Y}}$ and $f \in \mathcal{Y}^{\mathcal{X}}$, the infimum of eligible expectations of ℓ is the expectation of the infimum of f over S where S is distributed according to τ. Formally*

$$\inf_{\rho \vdash \tau} \mathbb{E}_{(X,Y)\sim\rho} \left[\ell(f(X), Y)\right] = \mathbb{E}_{(X,S)\sim\tau} \left[\inf_{y\in S} \ell(f(X), y)\right].$$

Proof. Before all, note that $(x, S) \to \inf_{y\in S} \ell(f(x), y)$ inherit measurability from f allowing to consider such an expectation (see Theorem 18.19 of Aliprantis and Border, 2006, and references therein for details). Moreover, let us use Lemma 41 to reformulate the right hand side problem as

$$\inf_{\rho \vdash \tau} \mathbb{E}_{(X,Y)\sim\rho} \left[\ell(f(X), Y)\right] = \inf_{\pi\in\mathcal{M}} \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{S}} \ell(f(x), y)\, \mathrm{d}\pi_{x,S}(y)\, \mathrm{d}\tau(x, S).$$

Where we denote by $\mathcal{M} \subset \Delta_{\mathcal{X}\times\mathcal{Y}\times\mathcal{S}}$ the space of probability measures π that satisfy the assumption of Lemma 41. We will now prove the equality by showing that both quantities bound the other one.

(≥). To proceed with the first bound, notice that for $x \in \mathcal{X}$ and $S \in \mathcal{S}$, when $\pi|_{x,S} \in \Delta_{\mathcal{Y}}$ only charge S, *i.e.* if $\pi \in \mathcal{M}$, then

$$\int_{\mathcal{Y}} \ell(f(x), y)\, \mathrm{d}\pi_{x,S}(y) \geq \inf_{y\in S} \ell(f(x), y).$$

The first bound is then obtained by taking the expectation over τ of this pointwise property.

(≤). For the second bound, we consider the function $Y \in \mathcal{Y}^{\mathcal{X}\times\mathcal{S}}$ define as

$$Y(x, S) = \arg\min_{y\in S} \ell(f(x), y).$$

Such a function is well-defined since S is compact due to the fact that $\mathcal{Y}$ is compact and $\mathcal{S}$ is the set of closed set. However, in more general cases, one can consider a sequence that minimizes $\ell(f(x), y)$ rather than the argmin to show the same as what we are going to show. Now, if we define $\pi^{(f)}$ with $\pi^{(f)}_{\mathcal{X}\times\mathcal{S}} := \tau$ and $\pi^{(f)}|_{x,S} := \delta_{Y(x,S)}$, because $Y(x, S)$ is in S, we have that $\pi^{(f)}$ is in $\mathcal{M}$, so, for $x \in \mathcal{X}$ and $S \in \mathcal{S}$

$$\inf_{\pi\in\mathcal{M}} \int_{\mathcal{Y}} \ell(f(x), y)\, \mathrm{d}\pi_{x,S}(y) \leq \int_{\mathcal{Y}} \ell(f(x), y)\, \mathrm{d}\pi^{(f)}_{x,S}(y) = \ell(f(x), Y(x, S)) = \inf_{y\in S} \ell(f(x), y).$$

We end the proof by integrating this over τ. □

Now, we will move on to the minimum switch. First, we make sure that the infimum loss minimizer is well-defined.

Lemma 43 (Infimum loss minimizer)**.** *When $\mathcal{Y}$ is compact and the observed set are closed, there exists a measurable function $f_S \in \mathcal{Y}^{\mathcal{X}}$ that minimize the infimum loss risk*

$$\mathcal{R}_S(f_S) = \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}_S(f), \qquad \textit{where} \qquad \mathcal{R}_S(f) = \int \min_{y\in S} \ell(f(x), y)\, \mathrm{d}\tau(x, S).$$

The infimum on the right hand side is a minimum because S is a closed subset of $\mathcal{Y}$ compact, and therefore, is compact.

Proof. First note that $d(y, y') = \sup_{z\in\mathcal{Y}} |\ell(z, y) - \ell(z, y')|$ is a metric on $\mathcal{Y}$ when ℓ is a proper loss: the triangular inequality holds trivially; when $y = y'$ then $d(y, y') = 0$; when $y \neq y'$, by properness we have $\ell(y, y) = 0$ and $d(y, y') \geq \ell(y, y') > 0$. Moreover, note that $L(z, S) = \min_{y\in S} \ell(z, y)$ is continuous and 1-Lipschitz with respect to the topology induced by the Hausdorff distance d_H based on d, indeed given two sets $S, S' \in \mathcal{S}$

$$|L(z, S) - L(z, S')| \leq \max\left\{\max_{y\in S}\min_{y'\in S'} |\ell(z, y) - \ell(z, y')|\,,\ \max_{y'\in S'}\min_{y\in S} |\ell(z, y) - \ell(z, y')|\right\}$$
$$\leq \max\left\{\max_{y\in S}\min_{y'\in S'} d(y, y'),\ \max_{y'\in S'}\min_{y\in S} d(y, y')\right\} = d_H(S, S').$$

The result of existence of a measurable f_S minimizing $\mathcal{R}_S(f) = \int L(f(x), S)d\tau(x, S)$ follows by the compactness of $\mathcal{Y}$, the continuity of $L(z, S)$ in the first variable with respect to the topology induced by d, in the second with respect to the topology induced by d_H and measurability of $\tau|_x$ in x, via Berge maximum theorem (see Theorem 18.19 of Aliprantis and Border, 2006, and references therein). □

We can state the minimum switch now.

Lemma 44 (Minimum switch)**.** *When $\mathcal{Y}$ is compact, and observed sets are closed, solving the partial labeling through the minimum variability principle*

$$f^* \in \operatorname*{arg\,min}_{f\in\mathcal{Y}^{\mathcal{X}}} \mathbb{E}_{\rho^\star}\left[\ell(f(X), Y)\right], \qquad \textit{with} \qquad \rho^\star \in \operatorname*{arg\,min}_{\rho\vdash\tau} \inf_{f\in\mathcal{Y}^{\mathcal{X}}} \mathbb{E}_{\rho}\left[\ell(f(X), Y)\right].$$

can be done jointly in f and ρ, and rewritten as

$$f^* \in \operatorname*{arg\,min}_{f\in\mathcal{Y}^{\mathcal{X}}} \inf_{\rho\vdash\tau} \mathbb{E}_{\rho}\left[\ell(f(X), Y)\right].$$

Proof. When $(\rho^\star, f^*)$ is a minimizer of the top problem, it also minimizes the joint problem $(\rho, f) \to \mathcal{R}(f; \rho)$, and we can switch the infimum order. The hard part is to show that when f_S minimizes the bottom risk, the infimum over ρ is indeed a minimum. Indeed, we know from Lemma 42 that f_S is characterized as a minimizer of the infimum risk $\mathcal{R}_S$, those are well-defined as shown in precedent lemma. To f_S, we can associate $\rho_S := \pi^{(f)}$ as defined in the proof of Lemma 42, which is due to the closeness of sets in $\mathcal{S}$ and the compactness of $\mathcal{Y}$. Indeed, (f_S, ρ_S) minimize jointly the objective $\mathcal{R}(f, \rho)$, so we have that

$$\rho_S \in \operatorname*{arg\,min}_{\rho\vdash\tau} \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f; \rho), \qquad \text{and} \qquad f_S \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f; \rho_S).$$

From which we deduced that ρ_S can be written as a $\rho^\star$ and f_S as an f^*. □

Remark 45 (A counter example when sets are not closed.)**.** *The minimum switch relies on compactness assumption, which can be violated when the observed sets in $\mathcal{S}$ are not closed. Let us consider the case where $\mathcal{Y} = \mathbb{R}$, $\ell = \ell_2$ is the mean square loss. Consider the pointwise weak supervision*

$$\tau = \frac{1}{2}\delta_{\mathbb{Q}} + \frac{1}{2}\delta_{\sqrt{2}\mathbb{Q}},$$

In this case, we have $\rho^\star = \delta_0$. Yet, for any z, we do have $\mathcal{R}_{S,x}(z) = 0$ for any $z \in \mathbb{R}$. For example, if $z = \sqrt{2}$, one can consider

$$\rho_n = \frac{1}{2}\delta_{\sqrt{2}} + \frac{1}{2}\delta_{\frac{\lfloor 10^n\sqrt{2}\rfloor}{10^n}},$$

to show that $z \in \arg\min_{z\in\mathcal{Y}} \inf_{\rho\vdash\tau} \mathcal{R}(z, \rho)$. As one can see this is counter example is based on the fact that $\{\rho \,|\, \rho \vdash \tau\}$ is not complete, so that there exists infimum of $\mathcal{R}_x(z, \rho)$ that are not minimum such as $\mathcal{R}_x(\sqrt{2}, \delta_{\sqrt{2}})$.

7.A.3 Proof of Theorem 12

If τ is not ambiguous, then, almost surely for $x \in \mathcal{X}$, if y_x is the only element in S_x of Definition 34, we know that $\rho^\star|_x = \delta_{y_x}$, and consequently we derive $f^*(x) = y_x$, so for it to be consistent with f_0, we need that $f_0(x) = y_x$.

Moreover, because τ is a weakening of ρ_0, ρ_0 is eligible for τ. When ρ_0 is deterministic, we know from considerations in the proof of Lemma 39, that it is $\rho^\star$, the only deterministic distribution eligible for τ. Thus, in fact, the condition $S_x = \{f_0(x)\}$ is implied by ρ_0 deterministic.

7.A.4 Proof of Theorem 37

When τ is not ambiguous, we know from Theorem 35, that $\rho^\star$ is deterministic. Let us write $\rho^\star|_x = \delta_{y_x}$, we have $f^*(x) = y_x$, and $\mathcal{R}_x(f^*) = 0$, moreover, because y_x is in every S in the support of $\tau|_S$, then $\mathcal{R}_{S,x}(f^*) = 0$. Similarly to the bound given by Cour et al. (2011) for the 0-1 loss, we have

$$\begin{aligned}\mathcal{R}_{S,x}(z) = \mathbb{E}_{S\sim\tau|_x}[\inf_{z'\in S}\ell(z,z')] &= \sum_{S;z\notin S}\inf_{z'\notin S}\ell(z,z')\,\mathbb{P}_{S\sim\tau|_x}(S)\\ &\geq \inf_{z'\neq z}\ell(z,z')\,\mathbb{P}_{S\sim\tau|_x}(z\notin S) \geq \inf_{z'\neq z}\ell(z,z')\eta,\end{aligned}$$

while $\mathcal{R}_x(z) = \ell(z,y)$, so we deduce locally

$$\begin{aligned}\mathcal{R}_x(z;\rho^\star|_x) - \mathcal{R}_x(f^*(x);\rho^\star|_x) &\leq \frac{\ell(z,y)}{\inf_{z'\neq z}\ell(z,z')}\eta^{-1}\left(\mathcal{R}_{S,x}(z) - \mathcal{R}_{S,x}(f^*(x))\right)\\ &\leq e^\nu\eta^{-1}\left(\mathcal{R}_{S,x}(z) - \mathcal{R}_{S,x}(f^*(x))\right).\end{aligned}$$

Integrating over x this last equation gives us the bound in Theorem 37.

7.A.5 Refined bound analysis of Theorem 37

The constant C that appears in Theorem 37 is the result of controlling separately the corruption process and the discrepancy of the loss. Indeed, they can be controlled together, leading to a better constant. To relates the two risk $\mathcal{R}$ and $\mathcal{R}_S$, we will consider the pointwise setting $\tau \in \Delta_{2^\mathcal{Y}}$ and $\rho_0 \in \Delta_\mathcal{Y}$ that satisfies $\rho_0 \vdash \tau$, we will also consider a prediction $z \in \mathcal{Y}$.

Proposition 46 (Bound refinement)**.** *When $\mathcal{Y}$ is discrete and τ not ambiguous, the best C that verifies* (7.6) *in the pointwise setting $\tau \in \Delta_{2^\mathcal{Y}}$ is the maximum of λ^{-1}, for $\lambda \in [0,1]$ such that there exists a point $z \neq y$ and signed measured σ that verify $\mathcal{R}(z;\sigma) = 0$ and such that $\sigma + \lambda\delta_y + (1-\lambda)\delta_z$ is a probability measure that is eligible for τ.*

Proof. First, let's extend our study to the space $\mathcal{M}_\mathcal{Y}$ of signed measure over $\mathcal{Y}$. We extend the risk definition in (7.1) to any signed measure $\mu \in \mathcal{M}_\mathcal{Y}$, with

$$\mathcal{R}_x(z;\mu) = \int_\mathcal{Y}\ell(z,y)\,\mathrm{d}\mu(y).$$

Note that the risk is a linear function of the distribution μ. Two spaces are going to be of particular interest: the one of measures of mass one $\mathcal{M}_{\mathcal{Y},1}$, and the one of measures of mass null $\mathcal{M}_{\mathcal{Y},0}$, where

$$\mathcal{M}_{\mathcal{Y},p} = \{\mu \in \mathcal{M} \mid \mu(\mathcal{Y}) = p\}.$$

Let's now relates for a ρ_0, τ and z, the risk $\mathcal{R}_x(z;\rho_0)$ and $\mathcal{R}_{S,x}(z)$. To do so, we introduce the space of signed measures of null mass, that could be said orthogonal to $(\ell(z,y))_{y\in\mathcal{Y}}$, formally

$$D_z = \left\{\mu \in \mathcal{M}_{\mathcal{Y},0}\,\middle|\,\mathcal{R}_x(z;\mu) = 0\right\}.$$

There are two alternatives: (1) either $\mathcal{R}_x(z;\rho_0) = 0$, and so $\mathcal{R}_{S,x}(z) = 0$ too, and we have related the two risk; (2) either $\mathcal{R}_x(z,\rho_0) \neq 0$, and the space $\mathcal{M}_{\mathcal{Y},1}$ can be decomposed as

$$\mathcal{M}_{\mathcal{Y},1} = D_z + \{\lambda\rho_0 + (1-\lambda)\delta_z \mid \lambda \in \mathbb{R}\}.$$

To prove it take $\mu \in \mathcal{M}_{\mathcal{Y},1}$, and use linearity of the risk after writing

$$\mu = \lambda\rho_0 + (1-\lambda)\delta_z + (\mu - (\lambda\rho_0 + (1-\lambda)\delta_z)), \qquad \text{with} \qquad \lambda = \frac{\mathcal{R}_x(z,\mu)}{\mathcal{R}_x(z,\rho_0)}.$$

For such a μ, using the linearity of the risk, and the properness of the loss, if we denote by d_z the part in D_z of the last decomposition, we have

$$\mathcal{R}_x(z;\mu) = \lambda\mathcal{R}_x(z;\rho_0) + (1-\lambda)\mathcal{R}_x(z;\delta_z) + \mathcal{R}_x(z;d_z) = \lambda\mathcal{R}_x(z;\rho_0)$$

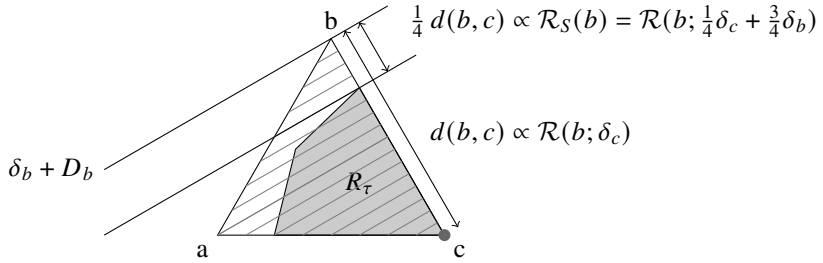

Figure 7.6: Geometrical understanding of Proposition 46, showing the link between the infimum and the fully supervised risk. The drawing is set in the affine span of the simplex $\mathcal{M}_{\mathcal{Y},1}$, where we identify a with δ_a. The underlying instance (ℓ, τ) is taken from Section 7.4, and can be linked to the setting of Proposition 46 with $z = b$, $y = c$. Are represented in the simplex the level curves of the function $\rho \to \mathcal{R}(z;\rho)$. Based on this drawing, one can recover $\mathcal{R}_S(b) = \mathcal{R}(b)/4$, which is better than the bound given in Theorem 37.

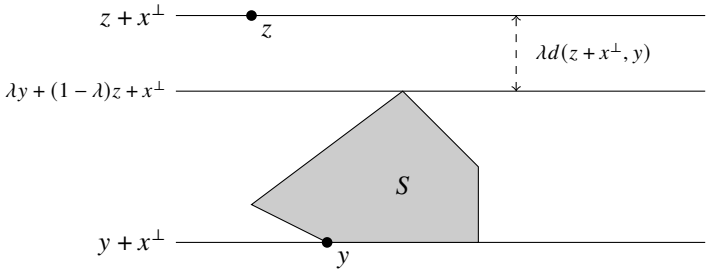

Figure 7.7: A variant of Thales theorem.

If we denote by $R_\tau = \{\rho \in \Delta_{\mathcal{Y}} \mid \rho \vdash \tau\}$, we can conclude that

$$\frac{\mathcal{R}_{S,x}(z)}{\mathcal{R}_x(z;\rho_0)} = \inf\left\{\lambda \mid (\lambda\rho_0 + (1-\lambda)\delta_z) \in R_\tau + D_z\right\}.$$

Finally, when τ is not ambiguous, we know that $\rho^\star$ is deterministic, and if ρ_0 is deterministic then $\rho_0 = \rho^\star$. In this case, there exists a y such that $\rho_0 = \delta_y$, and we can suppose this y is different from z otherwise $\mathcal{R}_x(z;\rho_0) = 0$. In this case, we also have $\mathcal{R}_x(z^*) = \mathcal{R}_{S,x}(z^*) = 0$ with $z^* = y$, and thus the excess of risk to relates in (7.6) is indeed the relation between the two risks.

□

Remark 47 (Proposition 46 as a variant of Thales theorem)**.** *Proposition 46 can be seen as a variant of the Thales theorem. Indeed, with the geometrical embedding π of the simplex in $\mathbb{R}^{\mathcal{Y}}$, $\pi(\rho) = (\rho(y))_{y\in\mathcal{Y}}$, one can have, with d the Euclidean distance*

$$\frac{\mathcal{R}_{S,x}(z)}{\mathcal{R}_x(z;\rho_0)} = \frac{d(\pi(\delta_z + D_z), \pi(R_\tau))}{d(\pi(\delta_z + D_z), \pi(\rho_0))}.$$

And conclude by using the following variant of Thales theorem, that can be derived from Figure 7.7: For $x, y, z \in \mathbb{R}^d$, and $S \subset \mathbb{R}^d$, with d the Euclidean distance, if $y \in S$, $d(z + x^\perp, S) = \gamma d(z + x^\perp, y)$, where

$$\gamma = \min\left\{|\lambda| \,\middle|\, \lambda \in \mathbb{R}, (\lambda y + (1-\lambda)z + x^\perp) \cap S \neq \emptyset\right\}.$$

Moreover, notice that if S is contained in the half-space that contains y regarding the cut with the hyperplane $z + x^\perp$, λ can be restricted to be in $[0, 1]$.

Remark 48 (Active labeling)**.** *When annotating data, as a partial labeler, you could ask yourself how to optimize your labeling. For example, suppose that you want to poll a population to retrieve preferences among a set of presidential candidates. Suppose that for a given polled person, you can only ask her to compare between four candidates. Which candidates would you ask her to compare? According to the questions you are asking, you will end up with different sets of potential weak distribution τ. If aware of the problem ℓ that your dataset is intended to tackle, and aware of a constant $C = C(\ell, \tau)$ that verifies (7.6), you might want to design your questions in order to maximize on average over potential τ, the quantity $C(\ell, \tau)$. An example where τ is not well-designed according to ℓ is given in Figure 7.8.*

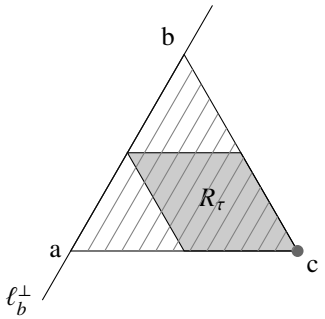

Figure 7.8: Example of a bad link between τ and ℓ. Same representation as Figure 7.6 with a different instance where $\tau = \frac{1}{2}\delta_{\{a,c\}} + \frac{1}{2}\delta_{\{b,c\}}$ and $\ell(b,a) = 0$, $\ell(b,c) = 1$. In this example $C_\ell(\tau) = +\infty$, and the infimum loss is 0 on $\mathcal{Y}$ and therefore not consistent. Given the loss structure, partial labeling acquisition should focus on specifying sets that do not intersect $\{a,b\}$. Note that this instance violates the proper loss assumption, explaining its inconsistency.

7.A.6 Proof of Theorems 13 and 14

First, note that, since $\mathcal{R}_S(f)$ is characterized by $\mathcal{R}_S(f) = \mathbb{E}_{(x,S)\sim\tau} \min_{u\in S} \ell(f(x),u)$, then the problem

$$f^* = \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}_S(f) = \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(x,S)\sim\tau}\left[\min_{y\in S} \ell(f(x),y)\right].$$

can be considered as an instance of structured prediction with loss $L(z,S) = \min_{y\in S} \ell(f(x),y)$. The framework for structured prediction presented in Ciliberto et al. (2016), and extended in Ciliberto et al. (2020), provides consistency and learning rates in terms of the excess risk $\mathcal{R}_S(f_n) - \mathcal{R}_S(f^*)$ when f^* is estimated via f_n defined as in (7.7) and when the structured loss L admits the decomposition

$$L(z,S) = \langle \psi(z), \varphi(S)\rangle_{\mathcal{H}},$$

for a separable Hilbert space $\mathcal{H}$ and two maps $\psi : \mathcal{Y} \to \mathcal{H}$ and $\varphi : \mathcal{S} \to \mathcal{H}$. Note that since $\mathcal{Y}$ is finite L always admits the decomposition, indeed the cardinality of $\mathcal{Y}$ is finite, *i.e.*, $|\mathcal{Y}| < \infty$ and $|\mathcal{S}| = 2^{|\mathcal{Y}|}$. Choose an ordering for the elements in $\mathcal{Y}$ and in $\mathcal{S}$ and denote them respectively $o_\mathcal{Y} : \mathbb{N} \to \mathcal{Y}$ and $o_\mathcal{S} : \mathbb{N} \to \mathcal{S}$. Let $n_\mathcal{Y} : \mathcal{Y} \to \mathbb{N}$ be the inverse of $o_\mathcal{Y}$, *i.e.* $o_\mathcal{Y}(n_\mathcal{Y}(y)) = y$ and $n_\mathcal{Y}(o_\mathcal{Y}(i)) = i$ for $y \in \mathcal{Y}$ and $i \in 1,\dots,|\mathcal{Y}|$, define analogously $n_\mathcal{S}$. Now let $\mathcal{H} = \mathbb{R}^{|\mathcal{Y}|}$ and define the matrix $B \in \mathbb{R}^{|\mathcal{Y}|\times 2^{|\mathcal{Y}|}}$ with element $B_{i,j} = L(o_\mathcal{Y}(i), o_\mathcal{S}(j))$ for $i = 1,\dots,|\mathcal{Y}|$ and $j = 1,\dots,2^{|\mathcal{Y}|}$, then define

$$\psi(z) = e^{|\mathcal{Y}|}_{n_\mathcal{Y}(z)}, \quad \varphi(S) = B e^{2^{|\mathcal{Y}|}}_{n_\mathcal{S}(S)},$$

where e^k_i is the i-th element of the canonical basis of $\mathbb{R}^k$. We have that

$$\langle \psi(z), \varphi(S)\rangle_\mathcal{H} = \langle e^{|\mathcal{Y}|}_{n_\mathcal{Y}(z)}, B e^{2^{|\mathcal{Y}|}}_{n_\mathcal{S}(S)}\rangle_{R^{|\mathcal{Y}|}} = B_{n_\mathcal{Y}(z),n_\mathcal{S}(S)} = L(i_\mathcal{Y}(n_\mathcal{Y}(z)), i_\mathcal{S}(n_\mathcal{S}(S))) = L(z,S),$$

for any $z \in \mathcal{Y}, S \in \mathcal{S}$. So we can apply Theorem 4 and 5 of Ciliberto et al. (2016) (see also their extended forms in Theorem 4 and 5 of Ciliberto et al., 2020). The last step is to connect the excess risk on $\mathcal{R}_S$ with the excess risk on $\mathcal{R}(f,\rho^\star)$, which is done by our comparison inequality in Theorem 37.

Remark 49 (Illustrating the consistency in a discrete setting)**.** *Suppose that $\tau_{|x}$ has been approximate, as a signed measure $\hat\tau_{|x} = \sum_{i=1}^n \alpha_i(x)\delta_{S_i}$. After renormalization, one can represent it as a region $R_{\hat\tau_{|x}}$ in the affine span of $\Delta_\mathcal{Y}$. Retaking the settings of Section 7.4, suppose that*

$$\hat\tau(\{a,b\}) = \frac{1}{2}, \qquad \hat\tau(\{c\}) = \frac{1}{2}, \qquad \hat\tau(\{a,c\}) = \frac{1}{4}, \qquad \hat\tau(\{a,b,c\}) = -\frac{1}{4}.$$

This corresponds to the region $R_{\hat\tau}$ represented in Figure 7.9. It leads to a disambiguation $\hat\rho$ that minimizes $\mathcal{E}$ (7.2), inside this space as

$$\hat\rho(a) = \frac{1}{2}, \qquad \hat\rho(b) = -\frac{1}{4}, \qquad \hat\rho(c) = \frac{3}{4},$$

and to the right prediction $\hat z = c$, since $\hat\rho$ felt in the decision region R_c. As the number of data augments, $R_{\hat\tau}$ converges toward R_τ, so does $\hat\rho$ toward $\rho^\star$ and the risk $\mathcal{R}(\hat f)$ toward its minimum.

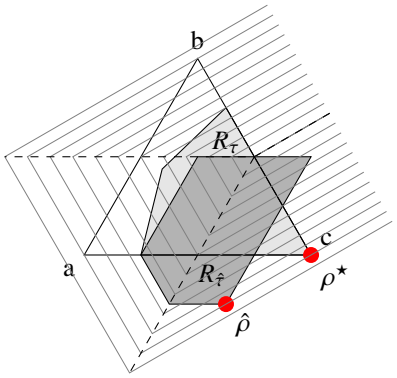

Figure 7.9: Understanding convergence of the algorithm (7.7). Our method is approximating τ as a signed measured $\hat{\tau}$, which leads to $R_{\hat{\tau}}$ in dark gray compared to the ground truth R_{τ} in light gray. The disambiguation of $\hat{\rho}$ and $\rho^\star$ is done on those two domains with the same objective $\mathcal{E}$ (7.2), which level curves are represented with light lines.

7.A.7 Understanding of the average and the supremum loss

For the average loss, if there is discrepancy in the loss $\nu > 0$, then there exists a, b, c such that $\ell(b, c) = (1 + \varepsilon)\ell(a, b)$, for some $\varepsilon > 0$. In this case, one can recreate the example of Section 7.4 by considering $\rho_0 = \rho^\star = \delta_c$ and

$$\tau = \lambda\delta_{\{c\}} + (1 - \lambda)\delta_{\{a,b,c\}}, \qquad \text{with} \qquad \lambda = \frac{1}{2}\frac{\varepsilon}{3\ell(a, b) + \varepsilon},$$

to show the inconsistency of the average loss. Similarly, supposing, without loss of generality that $\ell(a, c) \in [\ell(a, b), \ell(b, c)]$, the case where $\rho_0 = \rho^\star = \delta_b$ and

$$\tau = \lambda\delta_{\{b\}} + (1 - \lambda)\delta_{\{a,b,c\}}, \qquad \text{with} \qquad \lambda = \frac{1}{2}\min\left(\frac{\varepsilon}{1 + \varepsilon}, \frac{1 + \varepsilon - x}{2 + \varepsilon - x}\right), \qquad x = \frac{\ell(a, c)}{\ell(a, b)},$$

will fail the supremum loss, which will recover $z^* = a$, instead of $z^* = b$.

7.B Experiments

7.B.1 Classification

Let us consider the classification setting of Section 7.5.1. The infimum loss reads $L(z, S) = \mathbf{1}_{z \notin S}$. Given a weak distribution τ, the infimum loss is therefore solving for

$$f(x) \in \operatorname*{arg\,min}_{z \in \mathcal{Y}} \mathbb{E}_{S \sim \tau|_x}\left[L(z, S)\right] = \operatorname*{arg\,min}_{z \in \mathcal{Y}} \mathbb{E}_{S \sim \tau|_x}\left[\mathbf{1}_{z \notin S}\right] = \operatorname*{arg\,min}_{z \in \mathcal{Y}} \mathbb{P}_{S \sim \tau|_x}(z \notin S) = \operatorname*{arg\,max}_{z \in \mathcal{Y}} \mathbb{P}_{S \sim \tau|_x}(z \in S).$$

Given data, (z_i, S_i) our estimator consists in approximating the conditional distributions $\tau|_x$ as

$$\hat{\tau}|_x = \sum_{i=1}^{n} \alpha_i(x)\delta_{S_i},$$

from which we deduce the inference formula, that we could also be derived from (7.7),

$$\hat{f}(x) \in \operatorname*{arg\,max}_{z \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x)\mathbf{1}_{z \in S_i} = \operatorname*{arg\,max}_{z \in \mathcal{Y}} \sum_{i; z \in S_i} \alpha_i(x).$$

Complexity analysis

The complexity of our algorithm (7.7) can be split in two parts:

- a training part, where given (x_i, S_i) we precompute quantities that will be useful at inference.
- an inference part, where given a new x, we compute the corresponding prediction $\hat{f}(x)$.

In the following, we will review the time and space complexity of both parts. We give this complexity in terms of n the number of data and m the number of items in $\mathcal{Y}$. Results are summed up in Table 7.1.

Table 7.1: Complexity of our algorithm for classification.

Complexity	Time	Space
Training	$\mathcal{O}(n^2(n+m))$	$\mathcal{O}(n(n+m))$
Inference	$\mathcal{O}(nm)$	$\mathcal{O}(n+m)$

Training. Let us suppose that computing $L(y, S) = \mathbf{1}_{y \notin S}$ can be done in a constant cost that does not depend on m. We first compute the following matrices in $\mathcal{O}(nm)$ and $\mathcal{O}(n^2)$ in time and space.

$$L = (L(y, S_i))_{i \leq n, y \in \mathcal{Y}} \in \mathbb{R}^{n \times m}, \qquad K_\lambda = (k(x_i, x_j) + n\lambda\delta_{i=j})_{ij} \in \mathbb{R}^{n \times n}.$$

We then solve the following, based on the `_gesv` routine of Lapack, in $\mathcal{O}(n^3 + n^2 m)$ in time and $\mathcal{O}(n(n+m))$ in space (see Golub and Loan, 1996, for details)

$$\beta = K_\lambda^{-1} L \in \mathbb{R}^{n \times m}.$$

Inference. At inference, we first compute in $\mathcal{O}(n)$ in both time and space

$$v(x) = (k(x, x_i))_{i \leq n} \in \mathbb{R}^n.$$

Then we do the following multiplication in $\mathcal{O}(nm)$ in time and $\mathcal{O}(m)$ in space,

$$\mathcal{R}_{S,x} = v(x)^T \beta \in \mathbb{R}^m.$$

Finally, we take the minimum of $\mathcal{R}_{S,x}(z)$ over z in $\mathcal{O}(m)$ in time and $\mathcal{O}(1)$ in space.

Baselines

The average loss is really similar to the infimum loss, it reads

$$L_{ac}(z, S) = \frac{1}{|S|} \sum_{y \in S} \ell(z, y) = 1 - \frac{\mathbf{1}_{z \in S}}{|S|} \simeq \frac{1}{|S|} \cdot \mathbf{1}_{z \notin S} = \frac{1}{|S|} L(z, S).$$

Following similar derivations to the one for the infimum loss, given a distribution τ, one can show that the average loss is solving for

$$f_{ac}(x) \in \operatorname*{arg\,max}_{z \in \mathcal{Y}} \sum_{S; z \in S} \frac{1}{|S|} \tau|_x(S),$$

which is consistent when τ is not ambiguous. The difference with the infimum loss is due to the term in $|S|$. It can be understood as an evidence weight, giving less importance to big sets that do not allow discriminating efficiently between candidates. Given data (x_i, S_i), it leads to the estimator

$$\hat{f}_{ac}(x) \in \operatorname*{arg\,min}_{z \in \mathcal{Y}} \sum_{i; z \in S_i} \frac{\alpha_i(x)}{|S_i|}.$$

The supremum loss is really conservative since

$$L_{sp}(z, S) = \sup_{y \in S} \ell(y, z) = \sup_{y \in S} \mathbf{1}_{y \neq z} = \mathbf{1}_{S \neq \{z\}}.$$

It is solving for

$$f(x) \in \operatorname*{arg\,max}_{z \in \mathcal{Y}} \tau|_x(\{z\}),$$

which empirically correspond to discarding all the set with more than one element

$$\hat{f}_{sp}(x) \in \operatorname*{arg\,min}_{z \in \mathcal{Y}} \sum_{i; S_i = \{z\}} \alpha_i(x).$$

Note that τ could be not ambiguous while charging no singleton, in this case, the supremum loss is not informative, as its risk is the same for any prediction.

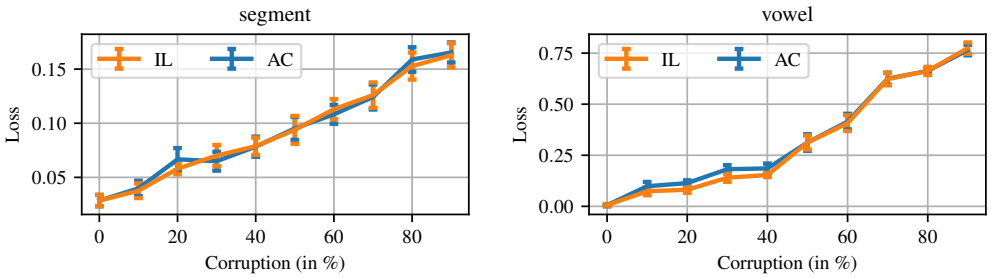

Figure 7.10: Classification. Testing risks (7.1) achieved by *AC* and *IL* on the "segment" and "vowel" datasets from *LIBSVM* as a function of corruption parameter c, when the corruption is uniform, as described in Section 7.B.1.

Table 7.2: *LIBSVM* datasets characteristics, showing the number of data, of classes, of input features, and the proportion of the most present class when labels are unbalanced.

Dataset	Data (n)	Classes (m)	Features (d)	Balanced	Most present
DNA	2000	3	180	×	52.6%
Svmguide2	391	3	20	×	56.5%
Segment	2310	7	19	✓	-
Vowel	528	11	10	✓	-

Corruptions on the *LIBSVM* datasets

To illustrate the dynamic of our method versus the average baseline, we used *LIBSVM* datasets (Chang and Lin, 2011), that we corrupted by artificially adding false class candidates to transform fully supervised pairs (x, y) into weakly supervised ones (x, S). We experiment with two types of corruption processes.

- A uniform one, reading, with the μ of Definition 33, for $z \neq y$,

$$\mathbb{P}_{(Y,S)\sim\mu|_{\mathcal{Y}\times 2^{\mathcal{Y}}}}(z \in S \,|\, Y = y) = c.$$

  with c a corruption parameter that we vary between zero and one. In this case, the average loss and the infimum one works the same as shown on Figure 7.10.
- A skewed one, where we only corrupt the pair (x, y) when y is the most present class in the dataset. More exactly, if y is the most present class in the dataset, for $z \in \mathcal{Y}$, and $z' \neq z$, our corruption process reads

$$\mathbb{P}_{(Y,S)\sim\mu|_{\mathcal{Y}\times 2^{\mathcal{Y}}}}(z' \in S \,|\, Y = z) = c \cdot \mathbf{1}_{z=y}.$$

  In unbalanced dataset, such as the "DNA" and "svmguide2" datasets, where the most present class represents more than fifty percent of the labels as shown Table 7.2, this allows fooling the average loss as shown Figure 7.2. Indeed, this corruption was designed to fool the average loss since we knew of the evidence weight $\frac{1}{|S|}$ appearing in its solution.

Reproducibility specifications

All experiments were run with *Python*, based on the *NumPy* library. Randomness was controlled by instantiating the random seed of *NumPy* to 0 before doing any computations. Results of Figures 7.2 and 7.10 were computed by using eight folds, and trying out several hyperparameters, before keeping the set of hyperparameters that hold the lowest mean error over the eight folds. Because we used a Gaussian kernel, there were two hyperparameters, the Gaussian kernel parameter σ, and the regularization parameter λ. We search for the best hyperparameters based on the heuristic

$$\sigma = c_\sigma d, \qquad \lambda = c_\lambda n^{-1/2},$$

where d is the dimension of the input $\mathcal{X}$ (or the number of features), and where the Gaussian kernel reads

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right).$$

We tried $c_\sigma \in \{10, 5, 1, .5, .1, .01\}$ and $c_\lambda \in \left\{10^i \,\middle|\, i \in [\![3, -3]\!]\right\}$.

7.B.2 Ranking

Consider the ranking setting of Section 7.5.2, where $\mathcal{Y} = \mathfrak{S}_m$, φ is the Kendall's embedding and the loss is equivalent to $\ell(z, y) = -\varphi(y)^T \varphi(z)$.

Complexity analysis

Given data (x_i, S_i), our algorithm is solving at inference for

$$f(x) \in \arg\min_{z \in \mathcal{Y}} \inf_{y_i \in S_i} -\sum_{i=1}^{n} \alpha_i(x) \varphi(z)^T \varphi(y_i) = \arg\max_{z \in \mathcal{Y}} \sup_{y_i \in S_i} \sum_{i=1}^{n} \alpha_i(x) \varphi(z)^T \varphi(y_i)$$

We solved it through alternate minimization, by iteratively solving in z for

$$\varphi(z)^{(t+1)} = \arg\max_{\xi \in \varphi(\mathcal{Y})} \left\langle \xi, \sum_{i=1}^{n} \alpha_i(x) \varphi(y_i)^{(t)} \right\rangle,$$

and solving for each y_i for

$$\varphi(y_i)^{(t+1)} = \arg\max_{\xi \in \varphi(S_i)} \alpha_i(x) \left\langle \xi, \varphi(z) \right\rangle.$$

We initialize the problem with the coordinates of $\varphi(y_i)$ put to 0 when not specified by the constraint $y_i \in S_i$.[2] Those two problems are minimum feedback arc set problems, that are *NP*-hard in m, meaning that one has to check for all potential solutions, and there is $m!$ of them, which is the cardinal of $\mathfrak{S}_m$. We suggest solving them using an integer linear programming (ILP) formulation that we relax into linear programming as explained in Appendix 7.C. All the problems in y_i share the same objective, up to a change in sign, but different constraint $\xi \in \varphi(S_i)$, such a setting is particularly suited for warm start on the dual simplex algorithm to solve efficiently one after the other the linear programs associated to each y_i.

To give numbers, at training time, we compute the inverse K_λ^{-1} in $\mathcal{O}(n^3)$ in time and $\mathcal{O}(n^2)$ in space, and at inference we compute $\alpha(x) K_\lambda^{-1} v(x)$ in $\mathcal{O}(n^2)$ in time and $\mathcal{O}(n)$ in space, before solving iteratively n *NP*-hard problem in m of complexity $nNP(m)$, that cost nm^2 in space to represent using *Cplex* (IBM, 2017), if we allow our self e iterations, the inference complexity is $\mathcal{O}(n^2 + e\, n\, NP(m))$ in time and $\mathcal{O}(nm^2)$ in space.

Baselines

The supremum loss is really similar to the infimum loss, only changing an infimum by a supremum. However, algorithmically, this change leads to solving for a local saddle point rather than solving for a local minimum. While the latter are always defined, there might be instances where no saddle point exists. In this case, the supremum optimization might stall without getting to any stable solution, and the user might consider stopping the optimization after a certain number of iteration and outputting the current state as a solution. The average loss, despite its simple formulation, does not lead to an easy implementation either. Indeed, when given a set S, the average loss is implicitly computing the center of this set $c(S)$, and replacing $L_{ac}(z, S)$ by $\ell(z, c(S))$, more exactly

$$L_{ac}(z, S) \simeq -\frac{1}{|S|} \sum_{y \in S} \varphi(z)^T \varphi(y) = -\varphi(y)^T \left(\frac{1}{|S|} \sum_{y \in S} \varphi(y) \right).$$

To compute the center $\left(\frac{1}{|S|} \sum_{y \in S} \varphi(y) \right)$, we sample $c_k \sim \mathcal{N}(0, I_{m^2})$, solve the resulting minimum feedback arc set problem, with the constraint $y \in S$, and end up with solutions $\varphi(y_k)$. After removing duplicates, we estimate the average with the empirical one. Note that this work is done at training, leading the average loss to have a quite good inference complexity in $\mathcal{O}(nm + NP(m))$ in time.

[2] Coordinates of the Kendall's embedding correspond to pairwise comparison between two items j and k, so we put to 0 the coordinates for which we can not infer preferences from S between items j and k.

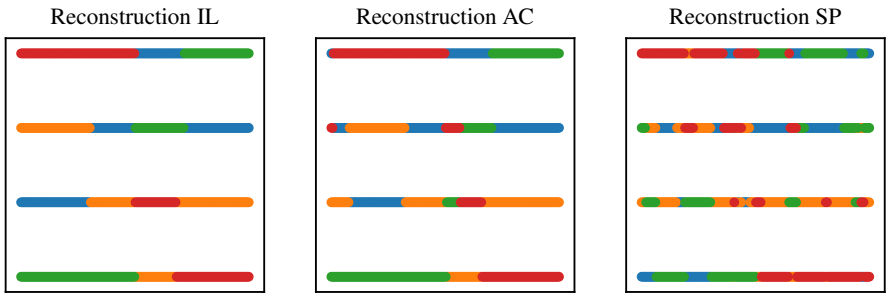

Figure 7.11: Reconstruction of the problem of Figure 7.3, given $n = 50$ random points $(x_i, y_i)_{i \leq n}$, after losing at random fifty percent of the coordinates $(\varphi(y_i))_{i \leq n}$, leading to sets $(S_i)_{i \leq n}$ of potential candidates. Hyperparameter were chosen as $\sigma = 1$ for the Gaussian kernel and $\lambda = 10^{-3} n^{-1/2}$ for the regularization parameter. The percentage of error in the reconstructed Kendall's embedding is 3% for *IL*, 4% for *AC* and 13% for *SP*. As for classification, with such a random corruption process, *AC* and *IL* show similar behaviors.

Synthetic example: ordering lines

In the following, we explain our synthetic example of Section 7.5.2. It corresponds of choosing $\mathcal{X} = [0, 1]$, choose m a number of items, simulate $a, b \sim \mathcal{N}(0, I_m)$, compute scores $v_i(x) = ax + b$, and order items according to their scores as shown on Figure 7.3. For Figure 7.4, we chose $m = 10$, as this is the biggest m for which can rely on our minimum feedback arc set heuristic to recover the real minimum feedback arc set solution and therefore not to play a role in what our algorithm will output. The corruption process was defined as losing coordinates in the Kendall's embedding, more exactly given a point $x \in \mathcal{X}$, we have a score $(v_i(x))_{i \leq m}$ and an ordering $y \in \mathcal{Y}$. To create a skewed corruption, we first compute the normalized distance between scores as

$$d_{ij} = \frac{\left| v_i - v_j \right|}{\max_{k,l} |v_k - v_l|} \in [0, 1]$$

and remove the pairwise comparison for which $d_{ij} > c$, where c is a corruption parameter between 0 and 1, formally

$$S = \left\{ z \in \mathcal{Y} \,\middle|\, \forall\, (j,k) \in I,\ \varphi(z)_{jk} = \varphi(y)_{jk} \right\}, \qquad \text{where} \qquad I = \left\{ (j,k) \,\middle|\, d_{(j,k)} < c \right\},$$

Because of the transitivity constraint, when c is small the comparison that we lost can be found back using transitivity between comparisons.

Reproducibility specification

To get Figure 7.4, we generate eight problems that correspond to ordering $m = 10$ lines, seen as eight folds. We only cross validated results with the same heuristics as in Appendix 7.B.1, yet, because computations were expensive we only tried $c_\sigma \in \{1, .5\}$, and $c_\lambda \in \left\{10^3, 1, 10^{-3}\right\}$. Again, randomness was controlled by instantiating random seeds to 0. Solving the linear program behind our minimum feedback arc set was done using *Cplex* (IBM, 2017), which is the fastest linear program solver we are aware of.

7.B.3 Multilabel

Multilabel is another application of partial labeling that we did not mention in our experiment section in the core paper. This omission was motivated by the fact that, under natural weak supervision, the three losses (infimum, average and supremum) are basically the same. However, we will provide, now, an explanation of this problem and our algorithm to solve it.

Multilabel prediction consists in finding which are the relevant tags (possibly more than one) among m potential tags. In this case, one can represent $\mathcal{Y} = \{-1, 1\}^m$, with $y_i = 1$ (resp. $y_i = -1$), meaning that tag i is relevant (resp. not relevant). The classical loss is the Hamming loss, which is the decoupled sum of errors for each label:

$$\ell(y, z) = \sum_{i=1}^{m} \mathbf{1}_{y_i \neq z_i}.$$

Table 7.3: Complexity of our algorithm for multilabels.

Complexity	Time	Space
Training	$\mathcal{O}(n^2(n+m))$	$\mathcal{O}(n(n+m))$
Inference	$\mathcal{O}(nm)$	$\mathcal{O}(n+m)$
Inference top-k	$\mathcal{O}(nm + m\log(m))$	$\mathcal{O}(n+m)$

Natural weak supervision consists in mentioning only a few relevant or irrelevant tags. This is the setting of Yu et al. (2014). This leads to sets S that are built from a set P of relevant items, and a set N of irrelevant items.

$$S = \{y \in \mathcal{Y} \,|\, \forall\, i \in P, y_i = 1, \forall\, i \in N, y_i = -1\}\,.$$

In this case, the infimum loss reads,

$$L(z, S) = \sum_{i \in P} \mathbf{1}_{z_i = -1} + \sum_{i \in N} \mathbf{1}_{z_i = 1}.$$

For such supervision, the infimum, the average and the supremum loss are intrinsically the same, they only differs by constants, due to the fact that for each unseen labels, the infimum loss pays 0, the average loss $1/2$ and the supremum loss 1.

When considering data $(x_i, S_i)_{i \leq n}$, where (S_i) is built from (N_i, P_i), our algorithm in (7.7) reads $\hat{f}(x) = (\mathrm{sign}(\hat{f}_j(x)))_{j \leq m}$, based on the scores

$$\hat{f}_j(x) = \sum_{i; j \in P_i} \alpha_i(x) - \sum_{i; j \in N_i} \alpha_i(x).$$

Tackling positive bias.

In the precedent development, we implicitly assumed that the ratio between positive and negative labels given by the weak supervision reflects the one of the full distribution. An assumption that is often violated in practice. It is common that partial labeling only mentions a subset of the relevant tags (*i.e.*, $N = \emptyset$). This case is ill-conditioned as always outputting all tags ($y = \mathbf{1}$) will minimize the infimum loss. To solve this problem, we can constrain the prediction space to the top-k space $\mathcal{Y}_k = \left\{y \in \mathcal{Y} \,\middle|\, \sum_{i=1}^m \mathbf{1}_{y_j = 1} = k\right\}$, which will lead to taking the top-k over the score $(\hat{f}_j)_{j \leq m}$. We can also break the loss symmetry and add a penalization with $\varepsilon > 0$,

$$\ell_\varepsilon(z, y) = \ell(z, y) + \varepsilon \sum_{i=1}^m \mathbf{1}_{z_i = 1}.$$

In this case, the inference algorithm will threshold scores at ε rather than 0.

Complexity analysis

The complexity analysis is similar to the one for classification. At training, we compute $L = (\mathbf{1}_{j \in P_i} - \mathbf{1}_{j \in N_i})$, and we solve for $\beta = K_\lambda^{-1} L$ in $\mathbb{R}^{n \times m}$. At testing, we compute $v(x)$ and $\beta^T v(x)$ in $\mathbb{R}^m$, before thresholding it or taking the top-k in either $\mathcal{O}(m)$ or $\mathcal{O}(m\log(m))$. As such, complexity reads similarly as for the classification case. Yet notice that, for multilabeling, the dimension of $\mathcal{Y}$ is not m but 2^m, meaning we do not scale with $\#\mathcal{Y}$ but with the intrinsic dimension.

Corruptions on the *MULAN* datasets

When sets are given by few positive and negative tags, all losses are the same. Yet, under other types of supervision, such as when the sets come as Hamming balls, defined by

$$B(z, r) = \{y \in \mathcal{Y} \,|\, \ell(z, y) \leq r\}\,,$$

the methods will not behave the same. We experiment on MULAN datasets provided by Tsoumakas et al. (2011). Because supervision with Hamming balls does not lead to efficient implementation, we went for

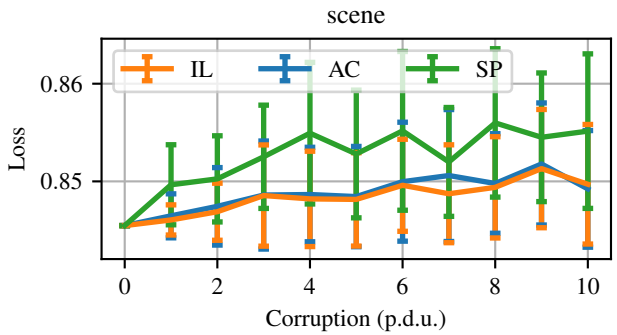

Figure 7.12: Multilabeling. Testing risks (from (7.1)) achieved by *AC* and *IL* on the "scene" dataset from MULAN as a function of corruption parameter c, shown in procedure defined unit, when the supervision is given as Hamming balls, as described in Section 7.B.3.

extensive grid search for the best solution, which reduces our ability to consider large m. Among MULAN datasets, we went for the "scene" one, with $m = 6$ tags, and $n = 2407$ data. When given a pair (x, y), we add corruption on y, by first sampling a radius parameter $r \sim \mathcal{U}(() [0, c * (m+1)])$, with c a corruption parameter. We then sample, with replacement, $\lfloor r \rfloor$ coordinates to modify to pass from y to a center c. We then consider the supervision $S = B(c, r)$. For such randomness, somehow uniform corruption, the infimum loss works slightly better than the average loss that both outperform the supremum loss as shown on Figure 7.12.

Reproducibility specification

To get Figure 7.12, we follow the same cross-validation scheme as for classification and ranking. More exactly, we cross-validated over eight folds with the same heuristics for σ, the Gaussian kernel parameter, and λ, the regularization one, with $c_\sigma \in \{10, 5, 1, .5, .1, .01\}$, and $c_\lambda \in \left\{10^i \,\middle|\, i \in [\![-3, 3]\!]\right\}$.

7.B.4 Partial regression

Partial regression is the regression instance of partial labeling. When supervision comes as intervals, it is known as interval regression, and known as censored regression when sets come as half-lines. Note that for censored regression, neither the average, nor the supremum loss can be properly defined.

Baselines

Given a bounded set S, learning with the average loss corresponds to considering the center of this set, since, for $z \in \mathcal{Y}$, with λ the Lebesgue measure

$$
\begin{aligned}
L_{ac}(z, S) &= \frac{1}{\lambda(S)} \int_S \|z - y\|^2 \, \lambda(\mathrm{d}y) = \|z\|^2 - 2 \left\langle z, \frac{1}{\lambda(S)} \int_S y \lambda(\mathrm{d}y) \right\rangle + \frac{1}{\lambda(S)} \int_S \|y\|^2 \, \lambda(\mathrm{d}y) \\
&= \left\| z - \frac{1}{\lambda(S)} \int_S y \lambda(\mathrm{d}y) \right\|^2 + \frac{1}{\lambda(S)} \int_S \|y\|^2 \, \lambda(\mathrm{d}y) - \left\| \frac{1}{\lambda(S)} \int_S y \lambda(\mathrm{d}y) \right\|^2 = \|z - c(S)\|^2 + C_S,
\end{aligned}
$$

where $c(S) = \frac{1}{\lambda(S)} \int_S y \lambda(\mathrm{d}y)$ is the center of S. As such, the average loss is always convex. As the supremum of convex function, the supremum loss is also convex.

Reproducibility specification

To compute Figure 7.5, for both *AC* and *IL*, we consider σ, the Gaussian kernel parameter, and λ, the regularization parameter, achieving the best risk when measured with the fully supervised distribution (7.1). We tried over $\sigma \in \{1, .5, .1, .05, .01\}$ and $\lambda \in \left\{10^3, 1, 10^{-3}\right\}$. Randomness was controlled by instantiating random seeds.

7.B.5 Beyond

Beyond the examples showcased previously, advances in dealing with weak supervision could be beneficial for several problems. Supervision on *image segmentation* problems usually comes as partial pixel annotation. This problem is often tackled through conditional random fields (Verbeek and Triggs, 2008), making it a perfect mix between partial labeling and structured prediction. *Action retrieval* on instructional video, where partial supervision is retrieved from the audio track is another interesting application (Alayrac, 2018).

7.C Minimum feedback arc set

7.C.1 Formulation

Consider a directed weighted graph with vertices $[\![1, m]\!]$ and edges $\{i \to j\}$ with weights $(w_{ij})_{i,j\leq m} \in \mathbb{R}_+^{m^2}$. The goal is to find a directed acyclic graph $G = (V, E)$ that maximizes the weights on remaining edges

$$\arg\max_{E} \sum_{i\to j\in E} w_{ij}.$$

This directed acyclic graph can be seen as a preference graph, item j being preferred over item i. Since w_{ij} are non-negative, the underlying ordering in G is necessarily total, and therefore can be written based on a score function, that can be embedded in the permutation of $[\![1, m]\!]$, $\sigma \in \mathfrak{S}_m$, with $\sigma(j) > \sigma(i)$ meaning that j is preferred over i. Thus the problem reads equivalently

$$\begin{aligned}\arg\max_{\sigma\in\mathfrak{S}_m} \sum_{i,j\leq m} w_{ij}\mathbf{1}_{\sigma(j)>\sigma(i)} &= \arg\max_{\sigma\in\mathfrak{S}_m} \sum_{i<j\leq m} c_{ij}\mathbf{1}_{\sigma(j)>\sigma(i)} = \arg\max_{\sigma\in\mathfrak{S}_m} \sum_{i<j\leq m} c_{ij}\,\mathrm{sign}\,(\sigma(j)-\sigma(i)) \\ &= \arg\min_{\sigma\in\mathfrak{S}_m} \sum_{i<j\leq m} c_{ij}\,\mathrm{sign}\,(\sigma(i)-\sigma(j)) = \arg\min_{\sigma\in\mathfrak{S}_m} \sum_{i<j\leq m} c_{ij}\mathbf{1}_{\sigma(i)>\sigma(j)}\end{aligned}$$

with $c_{ij} = w_{ij} - w_{ji}$. This last formulation is the one usually encountered for ranking algorithms in machine learning (Duchi et al., 2010).

We are going to study in depth this problem under the formulation

$$\arg\min_{\sigma\in\mathfrak{S}_m} \sum_{i<j\leq m} c_{ij}\,\mathrm{sign}\,(\sigma(i)-\sigma(j)) \tag{7.9}$$

7.C.2 Integer linear programming

Definition 50 (Kendall's embedding)**.** *For $\sigma \in \mathfrak{S}_m$, define Kendall's embedding, with $m_e = m(m-1)/2$,*

$$\varphi(\sigma) = \mathrm{sign}\,(\sigma(i)-\sigma(j))_{i<j\leq m} \in \{-1, 1\}^{m_e}.$$

We associate it to the Kendall's polytope of order m, $\mathrm{Conv}\,(\varphi(\mathfrak{S}_m))$.

The Kendall's embedding, Definition 50, cast the minimum feedback arcset problem (7.9) as a linear program

$$\begin{array}{ll}\text{minimize} & \langle c, x\rangle \\ \text{subject to} & x \in \mathrm{Conv}\,(\varphi(\mathfrak{S}_m))\,.\end{array}$$

Since the objective is linear, the solution is known to lie on a vertex of the constraint polytope, which is the set of Kendall's embeddings of permutations. Yet, how to describe Kendall's polytope?

Definition 51 (Transitivity polytope)**.** *The transitivity polytope of order m is defined in $\mathbb{R}^{m_e}$ as*

$$\mathcal{M} = \left\{x \in \mathbb{R}^{m_e} \,\middle|\, \forall\, i < k < j; -1 \leq x_{ij} + x_{jk} - x_{ik} \leq 1\right\}$$

This polytope encodes the transitivity constraints of Kendall's embeddings, Definition 50.

The transitivity polytope, Definition 53, will be used to approximate Kendall's polytope based on the following property.

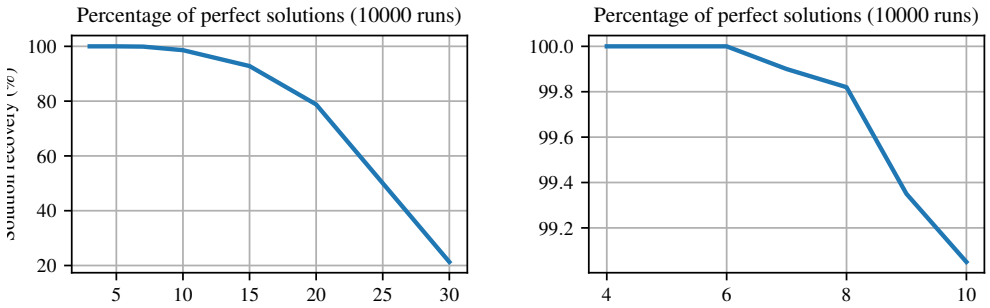

Figure 7.13: Evaluating the percentage of exact solutions of the ILP relaxation as m grows large. Evaluation is done by choosing an objective $c \sim \mathcal{N}(0, I_{m_e})$, solving the ILP relaxation, Definition 53, and evaluating if the solution is in $\{-1, 1\}^{m_e}$. The experience is repeated several times to estimate how often, on average, the original solution of (7.9) is returned by the ILP.

Proposition 52 (Relaxed polytope)**.** *The intersection between the transitivity polytope and the vertex of the hypercube is exactly the set of Kendall's embeddings of permutations. Mathematically*

$$\varphi(\mathfrak{S}_m) = \mathcal{M} \cap \{-1, 1\}^{m_e} .$$

Proof. First of all it is easy to show that $\varphi(\mathfrak{S}_m) \subset \{-1, 1\}^{m_e}$, and that, $\varphi(\mathfrak{S}_m) \subset \mathcal{M}$.

Let's now consider $x \in \mathcal{M} \cap \{-1, 1\}^{m_e}$. Let's associate to x the symmetric embedding

$$\tilde{x}_{ij} = \begin{cases} x_{ij} & \text{if} \quad i < j \\ 0 & \text{if} \quad i = j \\ -x_{ji} & \text{if} \quad j < i \end{cases}$$

Let's consider the permutation σ resulting from the ordering of $\sum_k \tilde{x}_{ik}$

$$\sigma^{-1}(1) = \underset{i \in [\![1,m]\!]}{\arg\min} \sum_{k=1}^{m} \tilde{x}_{ik} \qquad \text{and} \qquad \sigma^{-1}(i) = \underset{i \in [\![1,m]\!] \setminus \sigma^{-1}([\![1,i-1]\!])}{\arg\min} \sum_{k=1}^{m} \tilde{x}_{ik}.$$

Let's now show that $\varphi(\sigma) = x$, or equivalently that $\tilde{\varphi}(\sigma) = (\text{sign}(\sigma(i) - \sigma(j)))_{i,j \leq m} = \tilde{x}$. First, one can show that $\tilde{x}$ verify the transitivity constraints

$$\forall\, i, j, k \leq m, \qquad -1 \leq \tilde{x}_{ij} + \tilde{x}_{jk} - \tilde{x}_{ik} \leq 1.$$

This can be proven for any ordering of i, j, k based on the fact that $x \in \mathcal{M}$. For example, if $i < k < j$, we have

$$[-1, 1] \ni x_{ik} + x_{kj} - x_{ij} = \tilde{x}_{ik} - \tilde{x}_{jk} - \tilde{x}_{ij}.$$

which leads to

$$\tilde{x}_{ij} + \tilde{x}_{jk} - \tilde{x}_{ik} \in -[-1, 1] = [-1, 1] .$$

Now suppose, without loss of generality, that $\tilde{x}_{ij} = 1$ (if $\tilde{x}_{ij} = -1$, just consider $\tilde{x}_{ji} = 1$). The transitivity constraints tell us that $\tilde{x}_{ik} \geq \tilde{x}_{jk}$ for all k, therefore

$$\sum_{k \notin \{i,j\}} \tilde{x}_{ik} \geq \sum_{k \notin \{i,j\}} \tilde{x}_{jk}, \qquad \Rightarrow \qquad \sum_{k=1}^{m} \tilde{x}_{ik} > \sum_{k=1}^{m} \tilde{x}_{jk}. \qquad \Rightarrow \qquad \sigma(i) > \sigma(j).$$

This shows that $\tilde{\varphi(\sigma)}_{ij} = 1 = \tilde{x}_{ij}$. Thus, we have shown that $x \in \varphi(\mathfrak{S}_m)$, which concludes the proof. □

Definition 53 (ILP relaxation)**.** *Based on Proposition 52, we define the canonical polytope $\mathcal{C} = \mathcal{M} \cap [-1, 1]^{m_e}$, and relax the problem* (7.9) *into*

$$\begin{array}{ll} \textit{minimize} & \langle c, x \rangle \\ \textit{subject to} & x \in \mathcal{C} \end{array}$$

As soon as the solution x is in $\{-1, 1\}^{m_e}$, Proposition 52 tells us that x recover the exact minimum feedback arc set solution (7.9)*.*

In small dimensions, the canonical polytope $\mathcal{C}$ is the same as the Kendall's one, and the ILP relaxation gives the right solution. Yet, as shown in Figure 7.13, as soon as $m > 5$, there exists vertex in $\mathcal{C}$ that does not correspond to a permutation embedding. For small dimensions, proving that $\mathcal{C}$ is exactly the Kendall's polytope is done with a simple drawing for $m = 3$, using unimodularity of the transitivity constraint matrix is enough for $m = 4$ (Hoffman and Kruskal, 2010). The case $m = 5$ is also provable, based on several tricks that we will not discuss here.

Remark 54 (Low noise consistency)**.** *Remark that the low-noise setting considered by Duchi et al. (2010) correspond to having* $\operatorname{sign}(c) = -\varphi(y)$ *for a* $y \in \mathcal{Y}$*, in this case our algorithm is consistent and does recover the best solution* $z = y$*.*

7.C.3 Sorting heuristics

When formatting and solving the integer linear program takes too much time, one can go for a simple sorting heuristic, mainly based on a heuristic to compare items two by two and using quick sorting. A review of some heuristic with guarantees is provided by Ailon et al. (2005), Similar study when in presence of constraints on the resulting total order can be found in van Zuylen et al. (2007).

Chapter 8

Disambiguation Framework

The following is a reproduction of Cabannes et al. (2021b).

Machine learning approached through supervised learning requires expensive annotation of data. This motivates weakly supervised learning, where data are annotated with incomplete yet discriminative information. In this paper, we focus on partial labeling, an instance of weak supervision where, from a given input, we are given a set of potential targets. We review disambiguation principles to recover full supervision from weak supervision, and propose an empirical disambiguation algorithm. We prove exponential convergence rates of our algorithm under classical learnability assumptions, and we illustrate the usefulness of our method on practical examples.

8.1 Introduction

In many applications of machine learning, such as recommender systems, where an input x characterizing a user should be matched with a target y representing an ordering of a large number m of items, accessing fully supervised data (x, y) is not an option. Instead, one should expect weak information on the target y, which could be a list of previously taken (if items are online courses), watched (if items are plays), *etc.*, items by a user characterized by the feature vector x. This motivates *weakly supervised learning*, aiming at learning a mapping from inputs to targets in such a setting where tools from supervised learning can not be applied off-the-shelves.

Recent applications of weakly supervised learning showcase impressive results in solving complex tasks such as action retrieval on instructional videos (Miech et al., 2019), image semantic segmentation (Papandreou et al., 2015), salient object detection (Wang et al., 2017), 3D pose estimation (Dabral et al., 2018), text-to-speech synthesis (Jia et al., 2018), to name a few. However, those applications of weakly supervised learning are usually based on clever heuristics, and theoretical foundations of learning from weakly supervised data are scarce, especially when compared to statistical learning literature on supervised learning (Vapnik, 1995; Boucheron et al., 2005; Steinwart and Christmann, 2008). We aim to provide a step in this direction.

In this paper, we focus on partial labeling, a popular instance of weak supervision, approached with a structured prediction point of view Ciliberto et al. (2020). We detail this setup in Section 8.2. Our contributions are organized as follows.

- In Section 8.3, we introduce a disambiguation algorithm to retrieve fully supervised samples from weakly supervised ones, before applying off-the-shelf supervised learning algorithms to the completed dataset.
- In Section 8.4, we prove exponential convergence rates of our algorithm, in terms of the fully supervised excess of risk, given classical learnability assumptions.
- In Section 8.5, we explain why disambiguation algorithms are intrinsically non-convex, and provide guidelines based on well-grounded heuristics to implement our algorithm.

We end this paper with a review of literature in Section 8.6, before showcasing the usefulness of our method on practical examples in Section 8.7, and opening on perspectives in Section 8.8.

8.2 Disambiguation of partial labeling

In this section, we review the supervised learning setup, introduce the partial labeling problem along with a principle to tackle this instance of weak supervision.

Algorithms can be formalized as mapping an input x to a desired output y, respectively belonging to an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. Machine learning consists in automating the design of the mapping $f : \mathcal{X} \to \mathcal{Y}$, based on a joint distribution $\mu \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ over input/output pairings (x, y) and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, measuring the error cost of outputting $f(x)$ when one should have output y. The optimal mapping is defined as satisfying

$$f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,Y)\sim\mu} \left[\ell(f(X), Y)\right]. \tag{8.1}$$

In *supervised learning*, it is assumed that one does not have access to the full distribution μ, but only to independent samples $(X_i, Y_i)_{i\leq n} \sim \mu^{\otimes n}$. In practice, accessing such samples means building a dataset of examples. While input data (x_i) are usually easily accessible, getting output pairings (y_i) generally requires careful annotation, which is both time-consuming and expensive. For example, in image classification, (x_i) can be collected by scrapping images over the Internet. Subsequently, a "data labeler" might be asked to recognize a rare feline y_i on an image x_i. While getting y_i will be hard in this setting, recognizing that it is a feline and describing elements of color and shape is easy, and already helps to determine what outputs $f(x_i)$ are acceptable. A second example is given when pooling a known population (x_i) to get estimation of their political orientation (y_i), one might get information from recent election of percentage of voters across the political landscape, leading to global constraints that (y_i) should verify. A supervision that gives information on $(y_i)_{i\leq n}$ without giving its precise value is called *weak supervision*.

Partial labeling, also known as "superset learning", is an instance of weak supervision, in which, for an input x, we do not access the precise label y but only a set s of potential labels, $y \in s \subset \mathcal{Y}$. For example, on a caracal image x, one might not get the label "caracal" y, but the set s "feline", containing all the labels y corresponding to felines. It is modelled through a distribution $\nu \in \Delta_{\mathcal{X}\times 2^{\mathcal{Y}}}$ over $\mathcal{X} \times 2^{\mathcal{Y}}$ generating samples (X, S), which should be compatible with the fully supervised distribution $\mu \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ as formalized by the following definition.

Definition 55 (Compatibility, Cabannes et al. (2020b))**.** *A fully supervised distribution $\mu \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ is* compatible *with a weakly supervised distribution $\nu \in \Delta_{\mathcal{X}\times 2^{\mathcal{Y}}}$, denoted by $\mu \vdash \nu$ if there exists an underlying distribution $\pi \in \Delta_{\mathcal{X}\times\mathcal{Y}\times 2^{\mathcal{Y}}}$, such that μ, and ν, are the respective marginal distributions of π over $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{X} \times 2^{\mathcal{Y}}$, and such that $y \in s$ for any tuple (x, y, s) in the support of π (or equivalently $\pi|_s \in \Delta_s$, with $\pi|_s$ denoting the conditional distribution of π given s).*

This definition means that a weakly supervised sample $(X, S) \sim \nu$ can be thought as proceeding from a fully supervised sample $(X, Y) \sim \mu$ after losing information on Y according to the sampling of $S \sim \pi|_{X,Y}$. The goal of partial labeling is still to learn f^* from (8.1), yet without accessing a fully supervised distribution $\mu \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ but only the weakly supervised distribution $\nu \in \Delta_{\mathcal{X}\times 2^{\mathcal{Y}}}$. As such, this is an ill-posed problem, since ν does not discriminate between all μ compatible with it. Following *lex parsimoniae*, Cabannes et al. (2020b) have suggested looking for μ such that the labels are the most deterministic function of the inputs, which they measure with a loss-based "variance", leading to the disambiguation

$$\mu^* \in \operatorname*{arg\,min}_{\mu\vdash\nu} \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,Y)\sim\mu} \left[\ell(f(X), Y)\right], \tag{8.2}$$

and to the definition of the optimal mapping $f^* : \mathcal{X} \to \mathcal{Y}$

$$f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,Y)\sim\mu^*} \left[\ell(f(X), Y)\right]. \tag{8.3}$$

This principle is motivated by Theorem 1 of Cabannes et al. (2020b) showing that f^* in (8.3) is characterized by $f^* \in \arg\min_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,S)\sim\nu} \left[\inf_{y\in S} \ell(f(X), y)\right]$, matching a prior formulation based on infimum loss (Cour et al., 2011; Luo and Orabona, 2010; Hüllermeier, 2014). In practice, it means that if $(S|X = x)$ has probability 50% to be the set "feline" and 50% the set "orange with black stripes", $(Y|X = x)$ should be considered as 100% "tiger", rather than 20% "cat", 30% "lion" and 50% "orange car with black stripes", which could also explain $(S|X = x)$. Similarly to supervised learning, partial labeling consists in retrieving f^* without accessing ν but only samples $(X_i, S_i)_{i\leq n} \sim \nu^{\otimes n}$.

Remark 56 (Measure of determinism)**.** (8.2) *is not the only variational way to push toward distribution where labels are deterministic function of the inputs. For example, one could minimize entropy (*e.g.*, Berthelot et al., 2019; Lienen and Hüllermeier, 2021). However, a loss-based principle is appreciable since the loss usually encodes structures of the output space (Ciliberto et al., 2020), which will allow sample and computational complexity of consequent algorithms to scale with an intrinsic dimension of the space rather than the real one,* e.g.*, m rather than m! when $\mathcal{Y} = \mathfrak{S}_m$ and ℓ is a suitable ranking loss (see Section 8.5.4 or Nowak-Vila et al., 2019).*

8.3 Learning algorithm

In this section, given weakly supervised samples, we present a disambiguation algorithm to retrieve fully supervised samples based on an empirical expression of (8.2), before learning a mapping from $\mathcal{X}$ to $\mathcal{Y}$ based on those fully supervised samples, according to (8.3).

Given a partially labeled dataset $\mathcal{D}_n = (x_i, s_i)_{i\leq n}$, sampled accordingly to $\nu^{\otimes n}$, we retrieve fully supervised samples, based on the following empirical version of (8.2), with $C_n = \prod_{i\leq n} s_i \subset \mathcal{Y}^n$

$$(\hat{y}_i)_{i\leq n} \in \operatorname*{arg\,min}_{(y_i)_{i\leq n}\in C_n} \inf_{(z_i)_{i\leq n}\in\mathcal{Y}^n} \sum_{i,j=1}^{n} \alpha_j(x_i)\ell(z_i, y_j), \tag{8.4}$$

where $(\alpha_i(x))_{i\leq n}$ is a set of weights measuring how much one should base its prediction for x on the observations made at x_i. This formulation is motivated by the Bayes approximate rule proposed by Stone (1977), which can be seen as the approximation of μ by $n^{-1}\sum_{i,j=1}^{n}\alpha_j(x_i)\delta_{x_i}\otimes\delta_{y_j}$ in (8.2).

Once fully supervised samples $(x_i, \hat{y}_i)$ have been recollected, one can learn $f_n : \mathcal{X} \to \mathcal{Y}$, approximating f^*, with classical supervised learning techniques. In this work, we will consider the structured prediction estimator introduced by Ciliberto et al. (2016), defined as

$$f_n(x) \in \operatorname*{arg\,min}_{z\in\mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x)\ell(z, \hat{y}_i). \tag{8.5}$$

Weighting scheme α. For the weighting scheme α, several choices are appealing. Laplacian diffusion is one of them as it incorporates a prior on low density separation to boost learning (Zhu et al., 2003; Zhou et al., 2003; Bengio et al., 2006; Hein et al., 2007). Kernel ridge regression is another due to its theoretical guarantees (Ciliberto et al., 2020). In the theoretical analysis, we will use nearest neighbors. Assuming $\mathcal{X}$ is endowed with a distance d, and assuming, for readability' sake, that ties to define nearest neighbors do not happen, it is defined as

$$\alpha_i(x) = \begin{cases} k^{-1} & \text{if} \quad \sum_{j=1}^{n}\mathbf{1}_{d(x,x_j)\leq d(x,x_i)} \leq k \\ 0 & \text{otherwise,} \end{cases}$$

where k is a parameter fixing the number of neighbors. Our analysis, leading to Theorem 15, also holds for other local averaging methods such as partitioning or Nadaraya-Watson estimators.

8.4 Consistency result

In this section, we assume $\mathcal{Y}$ finite, and prove the convergence of f_n toward f^* as n, the number of samples, grows to infinity. To derive such a consistency result, we introduce a surrogate problem that we relate to the risk through a calibration inequality. We later assume that weights are given by nearest neighbors and review classical assumptions, that we work to derive exponential convergence rates.

In the following, we are interested in bounding the expected generalization error, defined as

$$\mathcal{E}(f_n) = \mathbb{E}_{\mathcal{D}_n}\mathcal{R}(f_n) - \mathcal{R}(f^*), \tag{8.6}$$

where $\mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\mu^*}\left[\ell(f(X), Y)\right]$, by a quantity that goes to zero, when n goes to infinity. This implies, under boundedness of ℓ, convergence in probability (the randomness being inherited from $\mathcal{D}_n$) of $\mathcal{R}(f_n)$ toward $\inf_{f:\mathcal{X}\to\mathcal{Y}}\mathcal{R}(f)$, which is referred as *consistency* of the learning algorithm.[1] We first introduce a few objects.

[1] If $\mathbb{E}|X| < +\infty$ and $\mathbb{E}[|X_n - X|] \to 0$, $X_n \to X$ in probability.

Disambiguation ground truth (y_i^*)**.** Introduce $\pi^* \in \Delta_{\mathcal{X}\times\mathcal{Y}\times 2^{\mathcal{Y}}}$ expressing the compatibility of μ^* and ν as in Definition 55. Given samples $(x_i, s_i)_{i\leq n}$ forming a dataset $\mathcal{D}_n$, we enrich this dataset by sampling $y_i^* \sim \pi^*|_{x_i,s_i}$, which build an underlying dataset (x_i, y_i^*, s_i) sampled accordingly $(\pi^*)^{\otimes n}$. Given $\mathcal{D}_n$, while *a priori*, y_i^* are random variables, sampled accordingly to $\pi^*|_{x_i,s_i}$, because of the definition of μ^* (8.2), under basic definition assumptions, they are actually deterministic, defined as $y_i^* = \arg\min_{y\in s_i} \ell(f^*(x_i), y)$. As such, they should be seen as ground truth for $\hat{y}_i$.

Surrogate estimates. The approximate Bayes rule was successfully analyzed recently through the prism of plug-in estimators by Ciliberto et al. (2020). While we will not cast our algorithm as a plug-in estimator, we will leverage this surrogate approach, introducing two mappings φ and ψ from $\mathcal{Y}$ to a Hilbert space $\mathcal{H}$ such that

$$\forall\, z, y \in \mathcal{Y}, \qquad \ell(z, y) = \langle \psi(z), \varphi(y) \rangle\,, \tag{8.7}$$

Such mappings always exist when $\mathcal{Y}$ is finite, and have been used to encode "problem structure" defined by the loss ℓ (Nowak-Vila et al., 2019). We introduce three surrogate quantities that will play a major role in the following analysis, they map $\mathcal{X}$ to $\mathcal{H}$ as

$$g^*(x) = \mathbb{E}_{\mu^*}\left[\varphi(Y) \mid X = x\right], \qquad g_n(x) = \sum_{i=1}^n \alpha_i(x)\varphi(\hat{y}_i),$$
$$g_n^*(x) = \sum_{i=1}^n \alpha_i(x)\varphi(y_i^*). \tag{8.8}$$

It is known that f^* and f_n are retrieved from g^* and g_n, through the decoding, retrieving $f : \mathcal{X} \to \mathcal{Y}$ from $g : \mathcal{X} \to \mathcal{H}$ as

$$f(x) = \operatorname*{arg\,min}_{z\in\mathcal{Y}} \langle \psi(z), g(x) \rangle\,, \tag{8.9}$$

which explains the wording of *plug-in* estimator (Ciliberto et al., 2020). We now introduce a *calibration inequality*, that relates the error between f_n and f^* with surrogate error quantities.

Lemma 57 (Calibration inequality)**.** *When $\mathcal{Y}$ is finite, and the labels are a deterministic function of the input,* i.e.*, when $\mu^*|_x$ is a Dirac for all $x \in \operatorname{supp}\nu_{\mathcal{X}}$, for any weighting scheme such that $\sum_{i=1}^n |\alpha_i(x)| \leq 1$ for all $x \in \operatorname{supp}\nu_{\mathcal{X}}$,*

$$\begin{aligned}\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq\ & 4c_\psi \left\|g_n^* - g_n\right\|_{L^1} \\ & + 8c_\psi c_\varphi\, \mathbb{P}_X\left(\left\|g_n^*(X) - g^*(X)\right\| > \delta\right),\end{aligned} \tag{8.10}$$

with $c_\psi = \sup_{z\in\mathcal{Y}} \|\psi(z)\|$, $c_\varphi = \sup_{z\in\mathcal{Y}} \|\varphi(y)\|$, and δ a parameter that depend on the geometry of ℓ and its decomposition through φ.

This lemma, proven in Appendix 8.A.1, separates a part reading in $\left\|g_n - g_n^*\right\|$, due to the *disambiguation error* between $(\hat{y}_i)$ and (y_i^*) together with the *stability* of the learning algorithm when substituting $(\hat{y}_i)$ for (y_i^*), and a part in $\left\|g_n^* - g^*\right\|$ due to the *consistency* of the fully supervised learning algorithm. The expression of the first part relates to Theorem 7 in Ciliberto et al. (2020) while the second part relates to Theorem 6 in Cabannes et al. (2021c).

8.4.1 Classical learnability assumptions

In the following, we suppose that the weights α are given by nearest neighbors, that $\mathcal{X}$ is a compact metric space endowed with a distance d, that $\mathcal{Y}$ is finite and that ℓ is proper in the sense that it strictly positive except on the diagonal of $\mathcal{Y}\times\mathcal{Y}$ diagonal where it is zero. We now review classical assumptions to prove consistency. First, assume that $\nu_{\mathcal{X}}$ is regular in the following sense.

Assumption 14 ($\nu_{\mathcal{X}}$ well-behaved)**.** *Assume that $\nu_{\mathcal{X}}$ is such that there exists $h_1, c_\mu, q > 0$ satisfying, with $\mathcal{B}$ designing balls in $\mathcal{X}$,*

$$\forall\, x \in \operatorname{supp}\nu_{\mathcal{X}},\ \forall\, r < h_1, \qquad \nu_{\mathcal{X}}(\mathcal{B}(x, r)) > c_\mu r^q.$$

Assumption 14 is useful to make sure that neighbors in $\mathcal{D}_n$ are closed with respect to the distance d, it is usually derived by assuming that $\mathcal{X}$ is a subset of $\mathbb{R}^q$; that $\nu_{\mathcal{X}}$ has a density p against the Lebesgue measure λ with *minimal mass* $p_{\min}$ in the sense that for any $x \in \operatorname{supp} \nu_{\mathcal{X}}$, $p(x) > p_{\min}$; and that $\operatorname{supp} \nu_{\mathcal{X}}$ has regular boundary in the sense that $\lambda(\mathcal{B}(x,r) \cap \operatorname{supp} \nu_{\mathcal{X}}) \geq c\lambda(\mathcal{B}(x,r)$ for any $x \in \operatorname{supp} \nu_{\mathcal{X}}$ and $r < h$ (*e.g.*, Audibert and Tsybakov, 2007).

We now switch to a classical assumption in partial labeling, allowing for population disambiguation.

Assumption 15 (Non ambiguity, Cour et al. (2011))**.** *Assume the existence of $\eta \in [0,1)$, such that for any $x \in \operatorname{supp} \nu_{\mathcal{X}}$, there exists $y_x \in \mathcal{Y}$, such that $\mathbb{P}_\nu(y_x \in S \mid X = x) = 1$, and*

$$\forall\, z \neq y_x, \quad \mathbb{P}_\nu(z \in S \mid X = x) \leq \eta.$$

Assumption 15 states that when given the full distribution ν, there is one, and only one, label that is coherent with every observable sets for a given input. It is a classical assumption in literature about the learnability of the partial labeling problem (*e.g.*, Liu and Dietterich, 2014). When ℓ is proper, this implies that $\mu^*|_x = \delta_{y_x}$, and $f^*(x) = y_x$.

Finally, we assume that g^* is regular. As we are considering local averaging method, we will use Lipschitz-continuity, which is classical in such a setting.[2]

Assumption 16 (Regularity of g^*)**.** *Assume that there exists $c_g > 0$, such that for any $x, x' \in \mathcal{X}$, we have*

$$\|g^*(x) - g^*(x')\|_{\mathcal{H}} \leq c_g d(x, x').$$

It should be noted that regularity of g^*, Assumption 16, together with determinism of $\mu^*|_x$ inherited from Assumption 15 implies that classes $\mathcal{X}_y = \{x \mid f^*(x) = y\}$ are separated in $\mathcal{X}$, in the sense that there exists $h_2 > 0$, such that for any $y, y' \in \mathcal{Y}$ and $(x, x') \in \mathcal{X}_y \times \mathcal{X}_{y'}$, $d(x, x') > h_2$, which is a classical assumption to derive consistency of semi-supervised learning algorithm (*e.g.*, Rigollet, 2007). We detailed those implications in Appendix 8.A.2.

8.4.2 Exponential convergence rates

We are now ready to state our convergence result. We introduce $h = \min(h_1, h_2)$ and $p = c_\mu h^q$, so that for any $x \in \operatorname{supp} \nu_{\mathcal{X}}$, $\nu_{\mathcal{X}}(\mathcal{B}(x,h)) > p$.

Theorem 15 (Exponential convergence rates)**.** *When the weights α are given by nearest neighbors, under Assumptions 14, 15 and 16, the excess of risk in* (8.6) *is bounded by*

$$\begin{aligned}\mathcal{E}(f_n) \leq\ & 8c_\psi c_\varphi (n+1) \exp\left(-\frac{np}{16}\right) \\ & + 8c_\psi c_\varphi m \exp\left(-k\,|\log(\eta)|\right), \end{aligned} \tag{8.11}$$

as soon as $k < np/4$, with $m = |\mathcal{Y}|$. By taking $k_n = k_0 n$, for $k_0 < p/4$, this implies exponential convergence rates $\mathcal{E}(f_n) = O(n\exp(-n))$.

Sketch for Theorem 15. In essence, based on Lemma 57, Theorem 15 can be understood as two folds.

- A fully supervised error between g_n^* and g^*. This error can be controlled in $\exp(-np)$ as the non-ambiguity assumption implies a hard Tsybakov margin condition, a setting in which *the fully supervised estimate g_n^* is known to converge to the population solution g^* with such rates* (Cabannes et al., 2021c).
- A weakly disambiguation error, that is exponential too, since, based on Assumption 15, disambiguating between $z \in \mathcal{Y}$ and y_x from k sets S sampled accordingly to $\nu|_x$ can be done in η^k, and disambiguating between all $z \neq y_x$ and y_x in $m\eta^k = m\exp(-k\,|\log(\eta)|)$.

Appendix 8.A.3 provides details. □

Theorem 15 states that under a non-ambiguity assumption and a regularity assumption implying no-density separation, one can expect exponential convergence rates of f_n learned with weakly supervised data to f^* the solution of the fully supervised learning problem, measured with excess of fully supervised risk. Because of exponential convergence rates, we could expect polynomial convergence rates for a broader class of problems that are approximated by problems satisfying assumptions of Theorem 15. *The derived rates in $n\exp(-n)$ should be compared with rates in $n^{-1/2}$ and $n^{-1/4}$*, respectively derived, *under the same assumptions*, by Cour et al. (2011); Cabannes et al. (2020b).

[2] Its generalization through Hölder-continuity would work too.

8.4.3 Discussion on assumptions

While we have retaken classical assumptions from literature, those assumptions are quite strong, which allows us, by understanding their strength, to derive exponential convergence rates. Assumptions 14 and 16 are classical in the nearest neighbor literature with full supervision. If we were using (reproducing) kernel methods to define the weighting scheme α, those assumptions would be mainly replaced with "g^* belonging to the RKHS". Assumption 15 is the strongest assumption in our view, that we will now discuss.

How to check it in practice ? First, for Assumption 15 to hold, the labels have to be a deterministic function of the inputs. In other words, a zero error is achievable. Finally, Assumption 15 is related to dataset collection. If dealing with images, weak supervision could take the form of some information on shape, color, or texture, etc., Assumption 15 supposes that the weak information potentially given on a specific image x allows retrieving the unique label y of the image (*e.g.*, a "pig" could be recognized from its shape and its color). This is a reasonable assumption, if, for a given x, we ask at random a data labeler to provide us information on shape, color, or texture, etc. However, it will not be the case, if for some reasons (*e.g.* the dataset is built from several weakly annotated datasets), in some regions of the input space, we only get shape information, and in other regions, we only get color information. In particular, it is not verified for semi-supervised learning when the support of the unlabeled data distribution is not the same as the support of the labeled input data distribution.

How to relax it and what results to expect? Previous works used Assumption 15 to derive a calibration inequality between the infimum loss to the original loss (*e.g.*, see Proposition 2 by Cabannes et al., 2020b). In contrast, we relate the surrogate and original problem through a refined calibration inequality (8.10). This technical progress allows us to derive exponential convergence rates similarly to the work of Cabannes et al. (2021c). Importantly, in comparison with previous work, our calibration inequality Lemma 57 can easily be extended without the determinism assumption provided by Assumption 15. Essentially, in our work, Assumption 15 is used to simplify the study of $(\hat{y}_i)_{i\leq n}$ given by the disambiguation algorithm (8.4), and therefore the study of the disambiguation error in (8.10). The study of $(\hat{y}_i)_{i\leq n}$ without Assumption 15 would require other tools than the one presented in this paper. It could be studied in the realm of graphical model and message passing algorithm, or with Wasserstein distance and topological considerations on measures. With much milder forms of Assumption 15, we expect the rates to degrade smoothly with respect to a parameter defining the hardness of the problem, similarly to the works of Audibert and Tsybakov (2007); Cabannes et al. (2021c).

8.5 Optimization considerations

In this section, we focus on implementations to solve (8.4). We explain why disambiguation objectives, such as (8.2) are intrinsically non-convex and express a heuristic strategy to solve (8.4) besides non-convexity in classical well-behaved instances of partial labeling. Note that we do not study implementations to solve (8.5) as this study has already been done by Nowak-Vila et al. (2019). We end this section by considering a practical example to make derivations more concrete.

8.5.1 Non-convexity of disambiguation objectives

For readability, suppose that $\mathcal{X}$ is a singleton, justifying to remove the dependency on the input in the following. Consider $\nu \in \Delta_{2^\mathcal{Y}}$ a distribution modelling weak supervision. While the domain $\{\mu \in \Delta_\mathcal{Y} \mid \mu \vdash \nu\}$ is convex, a disambiguation objective $\mathcal{E} : \Delta_\mathcal{Y} \to \mathbb{R}$ defining $\mu^* \in \arg\min_{\mu\vdash\nu} \mathcal{E}(\mu)$, similarly to (8.2), that is minimized for deterministic distributions, which correspond to μ a Dirac, *i.e.*, minimized on vertices of its definition domain $\Delta_\mathcal{Y}$, can not be convex. In other terms, any disambiguation objective that pushes toward distributions where targets are deterministic function of the input, as mentioned in Remark 56, can not be convex.

Indeed, smooth disambiguation objectives such as entropy and our piecewise linear loss-based principle (8.2), reading pointwise $\mathcal{E}(\mu) = \inf_{z\in\mathcal{Y}} \mathbb{E}_{Y\sim\mu}[\ell(z, Y)]$, are concave. Similarly, its quadratic variant $\mathcal{E}'(\mu) = \mathbb{E}_{Y,Y'\sim\mu}[\ell(Y, Y')]$, is concave as soon as $(\ell(y, y'))_{y,y'\in\mathcal{Y}}$ is semi-definite negative. We illustrate those considerations on a concrete example with graphical illustration in Appendix 8.C. We should see how this translates on generic implementations to solve the empirical objective (8.4).

8.5.2 Generic implementation for (8.4)

Depending on ℓ and on the type of observed set (s_i), (8.4) might be easy to solve. In the following, however, we will introduce optimization considerations to solve it in a generic structured prediction fashion. To do so, we recall the decomposition of ℓ (8.7) and rewrite (8.4) as

$$(\hat{y}_i)_{i\leq n} \in \operatorname*{arg\,min}_{y_i\in C_n} \inf_{(z_i)\in\mathcal{Y}^n} \sum_{i,j=1}^{n} \alpha_j(x_i)\psi(z_i)^\top \varphi(y_j).$$

Since, given (y_j), the objective is linear in $\psi(z_j)$, the constraint $\psi(z_j) \in \psi(\mathcal{Y})$ can be relaxed with $\zeta_i \in \operatorname{Conv}\psi(\mathcal{Y})$.[3] Similarly, with respect to $\varphi(y_j)$, this objective is the infimum of linear functions, therefore is concave, and the constraint $\varphi(y_j) \in \varphi(s_j)$, could be relaxed with $\xi_i \in \operatorname{Conv}\varphi(s_j)$. Hence, with $\mathcal{H}_0 = \operatorname{Conv}\psi(\mathcal{Y})$ and $\Gamma_n = \prod_{j\leq n} \operatorname{Conv}\varphi(s_j)$, the optimization is cast as

$$(\hat{\xi}_i)_{i\leq n} \in \operatorname*{arg\,min}_{(\xi_i)\in\Gamma_n} \inf_{(\zeta_i)\in\mathcal{H}_0^n} \sum_{i,j=1}^{n} \alpha_j(x_i)\zeta_i^\top \xi_j. \tag{8.12}$$

Because of concavity, $(\hat{\xi}_i)$ will be an extreme point of Γ_n, that could be decoded into $\hat{y}_i = \varphi^{-1}(\hat{\xi}_i)$. However, it should be noted that if only interested in f_n and not in the disambiguation $(\hat{y}_i)$, this decoding can be avoided, since (8.5) can be rewritten as $f_n(x) \in \operatorname{arg\,min}_{z\in\mathcal{Y}} \psi(z)^\top \sum_{i=1}^n \alpha_i(x)\hat{\xi}_i$.

8.5.3 Alternative minimization with good initialization

To solve (8.12), we suggest using an alternative minimization scheme. The output of such a scheme is highly dependent to the variable initialization. In the following, we introduce well-behaved problem, where $(\xi_i)_{i\leq n}$ can be initialized smartly, leading to an efficient implementation to solve (8.12).

Definition 58 (Well-behaved partial labeling problem)**.** *A partial labeling problem (ℓ, ν) is said to be well-behaved if for any $s \in \operatorname{supp}\nu_{2\mathcal{Y}}$, there exists a signed measure μ_s on $\mathcal{Y}$ such that the function from $\mathcal{Y}$ to $\mathbb{R}$ defined as $z \to \int_{\mathcal{Y}} \ell(z,y)\,\mathrm{d}\mu_s(y)$ is minimized for, and only for, $z \in s$.*

We provide a real-world example of a well-behaved problem in Section 8.5.4 as well as a synthetic example with graphical illustration in Appendix 8.C. On those problems, we suggest solving (8.12) by considering the initialization $\xi_i^{(0)} = \mathbb{E}_{Y\sim\mu_{s_i}}[\varphi(Y)]$, and performing alternative minimization of (8.12), until attaining $\xi^{(\infty)}$ as the limit of the alternative minimization scheme (which exists since each step decreases the value of the objective in (8.12) and there is a finite number of candidates for (ξ_i)). It corresponds to a disambiguation guess $\tilde{y}_i = \varphi^{-1}(\xi_i^{(\infty)})$. Then we suggest learning $\hat{f}_n$ from $(x_i, \tilde{y}_i)$ based on (8.5), and existing algorithmic tools for this problem (Nowak-Vila et al., 2019). To assert the well-foundedness of this heuristic, we refer to the following proposition, proven in Appendix 8.A.4.

Proposition 59. *Under the non-ambiguity hypothesis, Assumption 15, the solution of* (8.3) *is characterized by $f^* \in \operatorname{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,S)\sim\nu}\left[\mathbb{E}_{Y\sim\mu_S}[\ell(f(X),Y)]\right]$. Moreover, if the surrogate function $g_n^\circ : \mathcal{X} \to \mathcal{H}$ defined as $g_n^\circ(x) = \sum_{i=1}^n \alpha_i(x)\xi_{s_i}$, with $\xi_s = \mathbb{E}_{Y\sim\mu_s}[\varphi(Y)]$, converges toward $g^\circ(x) = \mathbb{E}_{S\sim\nu|_x}[\xi_S]$ in L^1, f_n° defined through the decoding* (8.9) *converges in risk toward f^*.*

Given that our algorithm scheme is initialized for $\xi_i^{(0)} = \xi_{s_i}$ and $\zeta_i^{(0)} = f_n^\circ(x_i)$ and stopped once having attained $\xi_i^{(\infty)}$ and $\zeta_i^{(\infty)} = \hat{f}_n(x_i)$, $\hat{f}_n$ is arguably better than f_n°, which given consistency result exposed in Proposition 59, is already good enough.

Remark 60 (IQP implementation for (8.4))**.** *Other heuristics to solve* (8.4) *are conceivable. For example, considering $z_i = y_i$ in this equation, we remark that the resulting problem is isomorphic to an integer quadratic program (IQP). Similarly to integer linear programming, this problem can be approached with relaxation of the "integer constraint" to get a real-valued solution, before "thresholding" it to recover an integer solution. This heuristic can be seen as a generalization of the Diffrac algorithm (Bach and Harchaoui, 2007; Joulin et al., 2010). We present it in details in Appendix 8.B.*

[3] The minimization pushes toward extreme points of the definition domain.

Remark 61 (Link with EM, (Dempster et al., 1977))**.** *Arguably, our alternative minimization scheme, optimizing respectively the targets $\xi_i = \varphi(y_i)$ and the function estimates $\zeta_i = \psi(f_n(x_i))$ can be seen as the non-parametric version of the Expectation-Maximization algorithm, popular for parametric model (Dempster et al., 1977).*

8.5.4 Application: ranking with partial ordering

Ranking is a problem consisting, for an input x in an input space $\mathcal{X}$, to learn a total ordering y, belonging to $\mathcal{Y} = \mathfrak{S}_m$, modelling preference over m items. It is usually approached with the Kendall loss $\ell(y, z) = -\varphi(y)^\top \varphi(z)$, with $\varphi(y) = (\text{sign}\,(y(i) - y(j)))_{i,j \leq m} \in \{-1, 1\}^{m^2}$ (Kendall, 1938). Full supervision corresponds, for a given x, to be given a total ordering of the m items. This is usually not an option, but one could expect to be given partial ordering that y should follow (Cao et al., 2007; Hüllermeier et al., 2008; Korba et al., 2018). Formally, this equates to the observation of some, but not all, coordinates $\varphi(y)_i$ of the vector $\varphi(y)$ for some $i \in I \subset [\![1, m]\!]^2$.

In this setting, $s \subset \mathcal{Y}$ is a set of total orderings that match the given partial ordering. It can be represented by a vector $\xi_s \in \mathcal{H}$, that satisfies the partial ordering observation, $(\xi_s)_I = \varphi(y)_I$, and that is agnostic on unobserved coordinates, $(\xi_s)_{^c I} = 0$. This vector satisfies that $z \to \psi(z)^\top \xi_s$ is minimized for, and only for, $z \in s$. Hence, it constitutes a good initialization for the alternative minimization scheme detailed above. We provide details in Appendix 8.A.5, where we also show that ξ_s can be formally translated in a μ_s to match the Definition 58, proving that ranking with partial labeling is a well-behaved problem.

Many real world problems can be formalized as a ranking problem with partial ordering observations. For example, x could be a social network user, and the m items could be posts of her connection that the network would like to order on her feed accordingly to her preferences. One might be told that the user x prefer posts from her close rather than from her distant connections, which translates formally as the constraint that for any i corresponding to a post of a close connection and j corresponding to a post of a distant connection, we have $\varphi(y)_{ij} = 1$. Nonetheless, designing non-parametric structured prediction models that scale well when the intrinsic dimension m of the space $\mathcal{Y}$ is very large (such as the number of post on a social network) remains an open problem, that this paper does not tackle.

8.6 Related work

Weakly supervised learning has been approached through parametric and non-parametric methods. Parametric models are usually optimized through maximum likelihood (Heitjan and Rubin, 1991; Jin and Ghahramani, 2002). Hüllermeier (2014) show that this approach, as formalized by Denoeux (2013), equates to disambiguating sets by averaging candidates, which was shown inconsistent by Cabannes et al. (2020b) when data are *not missing at random*. Among non-parametric models, Xu et al. (2004); Bach and Harchaoui (2007) developed an algorithm for clustering, that has been cast for weakly supervised learning problem (Joulin et al., 2010; Alayrac et al., 2016), leading to a disambiguation algorithm similar than ours, yet without consistency results. More recently, half-way between theory and practice, Gong et al. (2018) derived an algorithm geared toward classification, based on a disambiguation objective, incorporating several heuristics, such as class separation, and Laplacian diffusion. Those heuristics could be incorporated formally in our model.

The infimum loss principle has been considered by several authors, among them Cour et al. (2011); Luo and Orabona (2010); Hüllermeier (2014). It was recently analyzed through the prism of structured prediction by Cabannes et al. (2020b), leading to a consistent non-parametric algorithm that will constitute the baseline of our experimental comparison. This principle is interesting as it does not assume knowledge on the corruption process $(S|Y)$ contrarily to the work of Cid-Sueiro et al. (2014) or van Rooyen and Williamson (2017).

The non-ambiguity assumption has been introduced by Cour et al. (2011) and is a classical assumption of learning with partial labeling (Liu and Dietterich, 2014). Assumptions of Lipschitzness and minimal mass are classical assumptions to prove convergence of local averaging method (Audibert and Tsybakov, 2007; Biau and Devroye, 2015). Those assumptions imply class separation in $\mathcal{X}$, which has been leverage in semi-supervised learning, justifying Laplacian regularization (Rigollet, 2007; Zhu et al., 2003).

Note that those assumptions might not hold on raw representation of the data, but with appropriate metrics, which could be learned through unsupervised Duda et al. (2000) or self-supervised learning Doersch

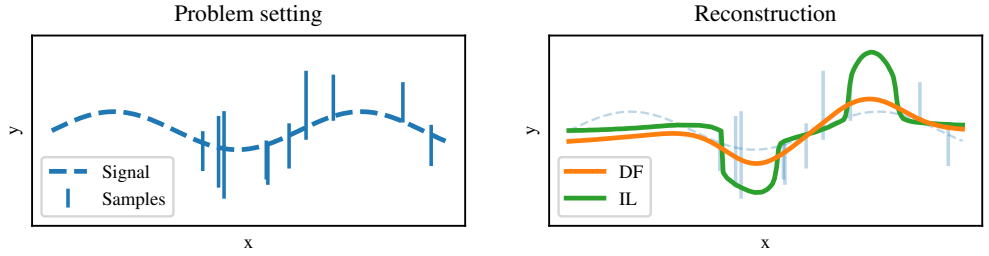

Figure 8.1: Interval regression. See Appendix 8.D for the exact reproducible experimental setup (Left) Setup. The goal is to learn $f^* : \mathcal{X} \to \mathbb{R}$ represented by the dashed line, given samples (x_i, s_i), where (s_i) are intervals represented by the blue segments. (Right) We compare the infimum loss (IL) baseline (8.13) shown in green, with our disambiguation framework (DF), (8.4) and (8.5), shown in orange; with weights α given by kernel ridge regression. (DF) retrieves $\hat{y}_i$ before learning a smooth f_n based on $(x_i, \hat{y}_i)$, while (IL) implicitly retrieves $\hat{y}_i(x)$ differently for each input, leading to irregularity of the consequent estimator of f^*.

and Zisserman (2017). As such, the practitioner might consider weights α given by similarity metrics derived through such techniques, before computing the disambiguation (8.4) and learning f_n from the recollected fully supervised dataset with deep learning.

8.7 Experiments

In this section, we review a baseline, and experiments that showcase the usefulness of our algorithm – which corresponds to (8.4) and (8.5).

Baseline. We consider as a baseline the work of Cabannes et al. (2020b), which is a consistent structured prediction approach to partial labeling through the infimum loss. It is arguably the state-of-the-art of partial labeling approached through structured prediction. It follows the same loss-based variance disambiguation principle, yet in an implicit fashion, leading to the inference algorithm, $f_n : \mathcal{X} \to \mathcal{Y}$,

$$f_n(x) \in \operatorname*{arg\,min}_{z \in \mathcal{Y}} \inf_{(y_i) \in C_n} \sum_{i=1}^{n} \alpha_i(x) \ell(z, y_i). \tag{8.13}$$

Statistically, exponential convergence rates similar to Theorem 15 could be derived. Yet, as we will see, our algorithm outperforms this state-of-the-art baseline.

Disambiguation coherence - Interval regression. The baseline (8.13) implicitly requires disambiguating $(\hat{y}_i(x))$ differently for every $x \in \mathcal{X}$. This is counterintuitive since (y_i^*) does not depend on x. It means that $(\hat{y}_i)$ could be equal to some $(\hat{y}_i^{(0)})$ on a subset $\mathcal{X}_0$ of $\mathcal{X}$, and to another $(\hat{y}_i^{(1)})$ on a disjoint subset $\mathcal{X}_1 \subset \mathcal{X}$, leading to irregularity of f_n between $\mathcal{X}_0$ and $\mathcal{X}_1$. We illustrate this graphically on Figure 8.1. This figure showcases an interval regression problem, which corresponds to the regression setup ($\mathcal{Y} = \mathbb{R}$, $\ell(y, z) = |y - z|^2$) of partial labeling, where one does not observe $y \in \mathbb{R}$ but an interval $s \subset \mathbb{R}$ containing y. Among others this problem appears in physics (Sheppard, 1897) and economy (Tobin, 1958).

Computation attractiveness - Ranking. Computationally, the baseline requires to solve a disambiguation problem, recovering $(\hat{y}_i(x)) \in C_n$ for every $x \in \mathcal{X}$ for which we want to infer $f_n(x)$. This is much more costly, than doing the disambiguation of $(\hat{y}_i) \in C_n$ once, and solving the supervised learning inference problem (8.5), for every $x \in \mathcal{X}$ for which we want to infer $f_n(x)$. To illustrate the computation attractiveness of our algorithm, consider the case of ranking, defined in Section 8.5.4. Fully supervised inference scheme (8.5) corresponds to solving a NP-hard problem, equivalent to the minimum feedback arcset problem (Duchi et al., 2010). While disambiguation approaches with alternative minimization implied by (8.4) and (8.13) require to solve this NP-hard problem for each minimization step. In other terms, the baseline ask to solve multiple NP-hard problem every time one wants to infer f_n given by (8.13) on an input $x \in \mathcal{X}$. Meanwhile, our disambiguation

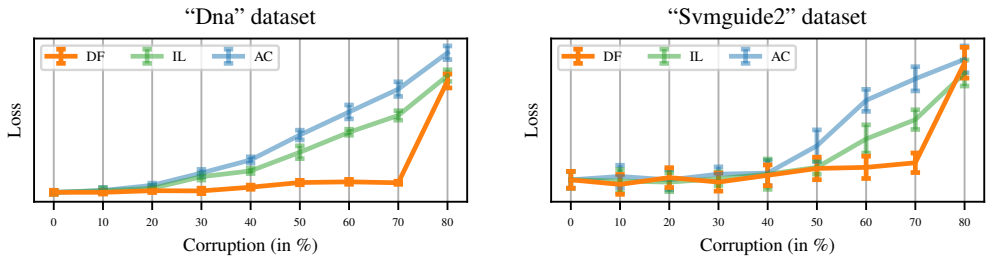

Figure 8.2: Testing errors as function of the supervision corruption on real dataset corresponding to classification with partial labels. We split fully supervised LIBSVM datasets into training and testing dataset. We corrupt training data in order to get partial labels. Corruption is managed through a parameter, represented by the x-axis, that relates to the ambiguity degree η of Assumption 15. For each method (our algorithm (DF), the baseline (IL), and the baseline of the baseline (AC, consisting of averaging candidates y_i in sets S_i)), we consider weights α given by kernel ridge regression with Gaussian kernel, for which we optimized hyperparameters with cross-validation on the training set. We then learn an estimate f_n that we evaluate on the testing set, represented by the y-axis, on which we have full supervision. The figure shows the superiority of our method, that achieves error similar to baseline when full supervision ($x = 0$) or no supervision ($x = 100\%$) is given, but performs better when only in presence of partial supervision. See Appendix 8.D for reproducibility specifications, where we also provide Figure 8.6 showcasing similar empirical results in the case of ranking with partial ordering.

approach asks to solve multiple NP-hard problem upfront to solve (8.4), yet only require to solve one NP-hard problem to infer f_n given by (8.5) on an input $x \in \mathcal{X}$.

Better empirical results - Classification. Finally, we compare our algorithm, our baseline (8.13) and the baseline considered by Cabannes et al. (2020b) on real datasets from the LIBSVM dataset (Chang and Lin, 2011). Those datasets (x_i, y_i) correspond to fully supervised classification problem. In this setup, $\mathcal{Y} = [\![1, m]\!]$ for m a number of classes, and $\ell(y, z) = \mathbf{1}_{y \neq z}$. We "corrupt" labels in order to create a synthetic weak supervision datasets (x_i, s_i). We consider skewed corruption, in the sense that (s_i) is generated by a probability such that $\sum_{z \in \mathcal{Y}} \mathbb{P}_{S_i}(z \in S_i | y_i)$ depends on the value of y_i. This corruption is parametrized by a parameter that related with the ambiguity parameter η of Assumption 15. Results on Figure 8.2 show that, in addition to having a lower computation cost, our algorithm performs better in practice than the state-of-the-art baseline.[4]

Beyond (8.2) **- Semi-supervised learning.** The main limitation of (8.2) is that it is a pointwise principle that decorrelates inputs, in the sense that the optimization of $\mu^*|_x$, for $x \in \mathcal{X}$, only depends on $\nu|_x$ and not on what is happening on $\mathcal{X} \setminus \{x\}$. As such, this principle failed to tackle semi-supervised learning, where $\nu|_x$ is equal to $\mu|_x$ (in the sense that $\pi|_{x,y} = \delta_{\{y\}}$) for $x \in \mathcal{X}_l$ and is equal to $\delta_{\mathcal{Y}}$ for $x \in \mathcal{X}_u := \mathcal{X} \setminus \mathcal{X}_l$. In such a setting, for $x \in \mathcal{X}_u$, $\mu^*|_x$ can be set to any δ_y for $y \in \mathcal{Y}$. Interestingly, in practice, while the baseline suffer the same limitation, for our algorithm, *weighting schemes have a regularization effect*, that contrasts with those considerations. We illustrate it on Figure 8.3.

8.8 Conclusion

In this work, we have introduced a structured prediction algorithm (8.4) and (8.5), to tackle partial labeling. We have derived exponential convergence rates for the nearest neighbors instance of this algorithm under classical learnability assumptions. We provided optimization considerations to implement this algorithm in practice, and have successfully compared it with the state-of-the-art. Several open problems offer prospective follow-up of this works:

- *Semi-supervised learning and beyond.* While we only proved convergence in situation where μ^* of (8.2) is uniquely defined, therefore excluding semi-supervised learning, Figure 8.3 suggests that our algorithm (8.4) could be analyzed in a broader setting than the one considered in this paper. Among

[4]All the code is available online - `https://github.com/VivienCabannes/partial_labelling/`.

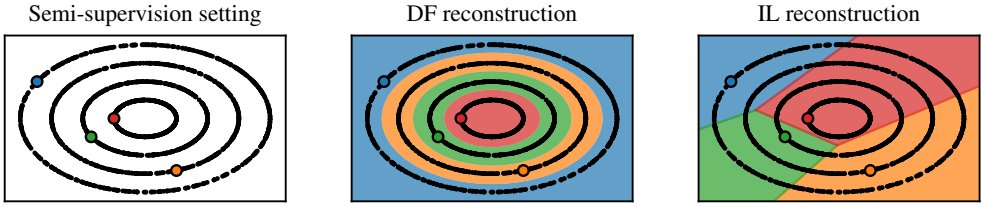

Figure 8.3: Semi-supervised learning, "concentric circle" instance with four classes (red, green, blue, yellow). Reproducibility details provided in Appendix 8.D. (Left) We represent points $x_i \in \mathcal{X} \subset \mathbb{R}^2$, there is many unlabeled points (represented by black dots and corresponding to $S_i = \mathcal{Y}$), and one labeled point for each class (represented in color, corresponding to $S_i = \{y_i\}$). (Middle) Reconstruction $f_n : \mathcal{X} \to \mathcal{Y}$ given by our algorithm (8.4) and (8.5). Our algorithm succeeds to comprehend the concentric circle structure of the input distribution and clusters classes accordingly. (Right) Reconstruction $f_n : \mathcal{X} \to \mathcal{Y}$ given by the baseline (8.13). The baseline performs as if only the four supervised data points where given.

others, we conjecture that the non-ambiguity assumption could be replaced by a cluster assumption (Rigollet, 2007) together with a non-ambiguity assumption cluster-wise in Theorem 15.

- *Hard-coded weak supervision.* Variational principles (8.2) and (8.3) could be extended beyond partial labeling to any type of hard-coded weak supervision, which is when weak supervision can be cast as a set of hard constraint that μ should satisfy, formally written as a set of fully supervised distributions compatible with weak information. Hard-coded weak supervision includes label proportion (Quadrianto et al., 2009; Dulac-Arnold et al., 2019), but excludes supervision of the type "80% of the experts say this nose is broken, and 20% say it is not". Providing a unifying framework for those problems would make an important step in the theoretical foundation of weakly supervised learning.
- *Missing input data.* While weak supervision assumes that only y is partially known, in many applications of machine learning, x is also only partially known, especially when the feature vector x is built from various source of information, leading to missing data. While we only considered a principle to fill missing output information, similar principles could be formalized to fill missing input information.

Appendix

8.A Proofs

Mathematical assumptions. To make formal what should be seen as implicit assumptions heretofore, we consider $\mathcal{X}$ and $\mathcal{Y}$ Polish spaces, $\mathcal{Y}$ compact, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ continuous, $\mathcal{H}$ a separable Hilbert space, φ measurable, and ψ continuous. We also assume that for ν_x-almost every $x \in \mathcal{X}$, and any $\mu \vdash \nu$, that the pushforward measure $\varphi_* \mu|_x$ has a second moment. This is the sufficient setup in order to be able to formally define objects and solutions considered all along the paper.

Notations. Beside standard notations, we use $|\mathcal{Y}|$ to design the cardinality of $\mathcal{Y}$, and $2^{\mathcal{Y}}$ to design the set of subsets of $\mathcal{Y}$. Regarding measures, we use $\mu_{\mathcal{X}}$ and $\mu|_x$ respectively the marginal over $\mathcal{X}$ and the conditional accordingly to x of $\mu \in \Delta_{\mathcal{X}\times\mathcal{Y}}$. We denote by $\mu^{\otimes n}$ the distribution of the random variable $(Z_1, \cdots, Z_n)$, where the Z_i are sampled independently according to μ. For A a Polish space, we consider Δ_A the set of Borel probability measures on this space. For $\varphi : \mathcal{Y} \to \mathcal{H}$ and $S \subset \mathcal{Y}$, we denote by $\varphi(S)$ the set $\{\varphi(y) \,|\, y \in S\}$. For a family of sets (S_i), we denote by $\prod S_i$ the Cartesian product $S_1 \times S_2 \times \cdots$, also defined as the set of points (y_i) such that $y_i \in S_i$ for all index i, and by $\mathcal{Y}^n$ the Cartesian product $\prod_{i \leq n} \mathcal{Y}$. Finally, for E a subset of a vector space E', $\operatorname{Conv} E$ denotes the convex hull of E and $\operatorname{Span}(E)$ its span.

Abuse of notations. For readability' sake, we have abused notations. For a signed measure μ, we denote by $\mathbb{E}_\mu[X]$ the integral $\int x \, \mathrm{d}\mu(x)$, extending this notation usually reserved to probability measure. More importantly, when considering $2^{\mathcal{Y}}$, we should actually restrict ourselves to the subspace $\mathcal{S} \subset 2^{\mathcal{Y}}$ of closed subsets of $\mathcal{Y}$, as $\mathcal{S}$ is a Polish space (metrizable by the Hausdorff distance) while $2^{\mathcal{Y}}$ is not always. However, when $\mathcal{Y}$ is finite, those two spaces are equals, $2^{\mathcal{Y}} = \mathcal{S}$.

8.A.1 Proof of Lemma 57

From Lemma 3 in Cabannes et al. (2021c), we pull the calibration inequality

$$\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq 2c_\psi \, \mathbb{E}\left[\mathbf{1}_{\|g_n(X) - g^*(X)\| > d(g^*(X), F)} \, \|g_n(X) - g^*(X)\|\right].$$

Where F is defined as the set of points $\xi \in \operatorname{Conv} \varphi(\mathcal{Y})$ leading to two decodings

$$F = \left\{ \xi \in \operatorname{Conv} \varphi(\mathcal{Y}) \,\middle|\, \left| \operatorname*{arg\,min}_{z \in \mathcal{Y}} \langle \psi(z), \xi \rangle \right| > 1 \right\},$$

and d is defined as the extension of the norm distance to sets, for $\xi \in \mathcal{H}$

$$d(\xi, F) = \inf_{\xi' \in F} \|\xi - \xi'\|_{\mathcal{H}}.$$

Using that $\|g_n(X) - g^*(X)\| \leq \left\|g_n(X) - g_n^*(X)\right\| + \left\|g_n^*(X) - g^*(X)\right\|$ and that, if $a \leq b + c$,

$$\mathbf{1}_{a > \delta} a \leq \mathbf{1}_{b + c > \delta} b + c \leq \mathbf{1}_{2 \sup(b,c) > \delta} 2 \sup b, c = 2 \sup_{e \in b, c} \mathbf{1}_{e > \delta} e \leq 2\mathbf{1}_{b > \delta} b + 2\mathbf{1}_{c > \delta} c.$$

We get the refined inequality

$$\begin{aligned}
&\mathcal{R}(f_n) - \mathcal{R}(g^*) \\
&\quad \leq 4c_\psi \, \mathbb{E}\left[\mathbf{1}_{2\|g_n(X) - g_n^*(X)\| > d(g^*(X), F)} \left\|g_n(X) - g_n^*(X)\right\| + \mathbf{1}_{2\|g_n^*(X) - g^*(X)\| > d(g^*(X), F)} \left\|g_n^*(X) - g^*(X)\right\|\right].
\end{aligned}$$

The first term is bounded by

$$\mathbb{E}\left[\mathbf{1}_{2\|g_n(X)-g_n^*(X)\|>d(g^*(X),F)}\left\|g_n(X)-g_n^*(X)\right\|\right] \leq \left\|g_n-g_n^*\right\|_{L^1}.$$

While for the second term, we proceed with

$$\begin{aligned}\mathbb{E}&\left[\mathbf{1}_{2\|g_n^*(X)-g^*(X)\|>d(g^*(X),F)}\left\|g_n^*(X)-g^*(X)\right\|\right]\\ &\leq \left\|g_n^*-g^*\right\|_{L^\infty}\mathbb{P}_X\left(2\left\|g_n^*(X)-g^*(X)\right\|>\inf_{x\in\operatorname{supp}\nu_{\mathcal{X}}}d(g^*(X),F)\right).\end{aligned}$$

When weights are positive and sum to one, both $g_n^*(X)$ and $g^*(X)$ are averaging of $\varphi(y)$ for $y\in\mathcal{Y}$, therefore

$$\left\|g_n^*-g^*\right\|_{L^\infty}\leq 2c_\varphi.$$

The same is true when $\sum_{i\leq n}|\alpha_i(x)|\leq 1$. Finally, when the labels are a deterministic function of the input, $g^*(X)=\varphi(f^*(X))$, and $d(g^*(X),F)\leq\sup_{y\in\mathcal{Y}}d(\varphi(y),F)$. Defining $\delta:=\sup_{y\in\mathcal{Y}}d(\varphi(y),F)/2$, and adding everything together leads to Lemma 57.

8.A.2 Implication of Assumptions 15 and 16

Assume that Assumption 15 holds, consider $x\in\operatorname{supp}\nu_{\mathcal{X}}$, let us show that $f^*(x)=y_x$ and $\mu^*|_x=\delta_{y_x}$. First of all, notice that $\bigcap_{S;S\in\operatorname{supp}\nu|_x}=\{y_x\}$; that $\delta_{y_x}\vdash\nu|_x$, as it corresponds to $\pi|_{x,S}=\delta_{y_x}\in\Delta_S$, for all S in the support of $\nu|_x$; and that, because ℓ is well-behaved,

$$\inf_{z\in\mathcal{Y}}\ell(z,y_x)=\ell(y_x,y_x)=0.$$

This infimum is only achieved for $z=y_x$, hence if we prove that $\mu^*|_x=\delta_{y_x}$, we directly have that $f^*(x)=y_x$. Finally, suppose that $\mu|_x\vdash\nu|_x$ charges $y\neq y_x$. Because y does not belong to all sets charged by $\nu|_x$, $\mu|_x$ should charge another $y'\in\mathcal{Y}$, and therefore

$$\inf_{z\in\mathcal{Y}}\mathbb{E}_{Y\sim\mu|_x}[\ell(z,y)]\geq\inf_{z\in\mathcal{Y}}\mu|_x(y)\ell(z,y)+\mu|_x(y')\ell(z,y')>0.$$

Which shows that $\mu^*|_x=\delta_{y_x}$. We deduce that $g^*(x)=y_x$.

Now suppose that Assumption 16 holds too, and consider two $x,x'\in\operatorname{supp}\nu_{\mathcal{X}}$ belonging to two different classes $f(x)=y$ and $f(x')=y'$. We have that $g^*(x)=\varphi(y)$ and $g^*(x')=\varphi(y')$, therefore,

$$d(x,x')\geq c^{-1}\left\|\varphi(y)-\varphi(y')\right\|_{\mathcal{H}}.$$

Define $h_2=\inf_{y\neq y'}c^{-1}\left\|\varphi(y)-\varphi(y')\right\|_{\mathcal{H}}$. Let us now show that $h_2>0$. When $\mathcal{Y}$ is finite, this infimum is a minimum, therefore, $h_2=0$, only if there exists a $y\neq y'$, such that $\varphi(y)=\varphi(y')$, which would implies that $\ell(\cdot,y)=\ell(\cdot,y')$ and therefore $\ell(y,y')=\ell(y,y)$ which is impossible when ℓ is proper.

8.A.3 Proof of Theorem 15

Reusing Lemma 57, we have

$$\mathcal{E}(f_n)\leq 4c_\psi\,\mathbb{E}_{\mathcal{D}_n,X}\left[\left\|g_n^*(X)-g_n(X)\right\|_{\mathcal{H}}\right]+8c_\psi c_\varphi\,\mathbb{E}_{\mathcal{D}_n,X}\left[\mathbf{1}_{\|g_n^*(X)-g^*(X)\|>\delta}\right].$$

We will first prove that

$$\mathbb{E}_{\mathcal{D}_n}\left[\mathbf{1}_{\|g_n^*(X)-g^*(X)\|>\delta}\right]\leq\exp\left(-\frac{np}{8}\right)$$

as long as $k<np/2$. The error between g^* and g_n relates to classical supervised learning of g^* from samples $(X_i,Y_i)\sim\mu^*$. We invite the reader who would like more insights on this fully supervised part of the proof to refer to the several monographs written on local averaging methods and, in particular, nearest neighbors, such as Biau and Devroye (2015). Because of class separation, we know that if k points fall at distance at most h of $x\in\operatorname{supp}\nu_{\mathcal{X}}$, $g_n^*(x)=k^{-1}\sum_{i;X_i\in\mathcal{N}(x)}\varphi(y_i)=\varphi(y_x)=g^*(x)$, where $\mathcal{N}(x)$ designs the k-nearest neighbors of x in (X_i). Because the probability of falling at distance h of x for each X_i is lower bounded by p, we have that

$$\mathbb{P}_{\mathcal{D}_n}(g_n^*(x)\neq g^*(x))\leq\mathbb{P}(\text{Bernoulli}(n,p)<k).$$

This can be upper bound by $\exp(-np/8)$ as soon as $k < np/2$, based on Chernoff multiplicative bound (see Biau and Devroye, 2015, for a reference), meaning

$$\mathbb{E}_{\mathcal{D}_n,X}\left[\mathbf{1}_{\|g_n^*(X)-g^*(X)\|\geq\delta}\right] \leq \exp(-np/8).$$

For the disambiguation part in $\left\|g_n - g_n^*\right\|_{L^1}$, we distinguish two types of datasets, the ones where for any input X_i its k-neighbors at are distance at least h, ensuring that disambiguation can be done by clusters, and datasets that does not verify this property. Consider the event

$$\mathbb{D} = \left\{(X_i)_{i\leq n} \,\middle|\, \sup_i d(X_i, X_{(k)}(X_i)) < h\right\}$$

where $X_{(k)}(x)$ design the k-th nearest neighbor of x in $(X_i)_{i\leq n}$. We proceed with

$$\mathbb{E}_{\mathcal{D}_n,X}\left[\left\|g_n^*(X) - g_n(X)\right\|_{\mathcal{H}}\right] \leq \sup_{X\in\mathcal{X}}\left\|g_n^* - g_n\right\|_\infty \mathbb{P}_{\mathcal{D}_n}((X_i)\notin\mathbb{D}) + \mathbb{E}_{\mathcal{D}_n,X}\left[\left\|g_n^*(X) - g_n(X)\right\|_{\mathcal{H}} \,\middle|\, (X_i)\in\mathbb{D}\right],$$

Which is based on $E[Z] = \mathbb{P}(Z\in A)\,\mathbb{E}[Z|A] + \mathbb{P}(Z\notin A)\,\mathbb{E}[Z|^c A]$. For the term corresponding to bad datasets, we can bound the disambiguation error with the maximum error. Similarly to the derivation for Lemma 57, because $g_n^*(x)$ and $g_n^*(X)$, are averaging of $\varphi(y)$, we have that

$$\sup_{x\in\operatorname{supp}\nu_{\mathcal{X}}}\left\|g_n(x) - g_n^*(x)\right\| \leq 2c_\varphi.$$

Indeed, we allow ourselves to pay the worst error on those datasets as their probability is really small, which can be proved based on the following derivation.

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_n}((X_i)_{i\leq n}\notin\mathbb{D}) &= \mathbb{P}_{(X_i)}(\sup_i d(X_i, X_{(k)}(X_i)) \geq h) = \mathbb{P}_{(X_i)}\left(\cup_{i\leq n}\left\{d(X_i, X_{(k)}(X_i)) \geq h\right\}\right)\\ &\leq \sum_{i=1}^n \mathbb{P}_{(X_i)}\left(d(X_i, X_{(k)}(X_i)) \geq h\right) = n\,\mathbb{P}_{X,\mathcal{D}_{n-1}}\left(d(X, X_{(k)}(X)) \geq h\right).\end{aligned}$$

This last probability has already been work out when dealing with the fully supervised part, and was bounded as

$$\mathbb{P}_{X,\mathcal{D}_{n-1}}\left(d(X, X_{(k)}(X)) \geq h\right) \leq \exp\left(-(n-1)p/8\right).$$

as long as $k < (n-1)p/2$. Finally, we have

$$\sup_{X\in\mathcal{X}}\left\|g_n^* - g_n\right\|_\infty \mathbb{P}_{\mathcal{D}_n}((X_i)_{i\leq n}\notin\mathbb{D}) \leq 2c_\varphi n\exp\left(-(n-1)p/8\right).$$

For the expectation term, corresponding to datasets $\mathcal{D}_n\in\mathbb{D}$ that cluster data accordingly to classes, we have to make sure that $\hat{y}_i = y_i^*$ is the only acceptable solution of (8.4), which is true as soon as the intersection of S_j, for x_j the neighbors of x_i, only contained y_i^*. To work out the disambiguation algorithm, notice that

$$\begin{aligned}\left\|g_n - g_n^*\right\|_{L^1} &= \int_{\mathcal{X}}\left\|\sum_{i=1}^n \alpha_i(x)\varphi(\hat{y}_i) - \varphi(y_i^*)\right\| \mathrm{d}\nu_{\mathcal{X}}(x) \leq \int_{\mathcal{X}} k^{-1}\sum_{i=1}^n \mathbf{1}_{X_i\in\mathcal{N}(x)}\left\|\varphi(\hat{y}_i) - \varphi(y_i^*)\right\| \mathrm{d}\nu_{\mathcal{X}}(x)\\ &= k^{-1}\sum_{i=1}^n \mathbb{P}_X\left(X_i\in\mathcal{N}(X)\right)\left\|\varphi(\hat{y}_i) - \varphi(y_i^*)\right\| \leq 2c_\varphi k^{-1}\sum_{i=1}^n \mathbb{P}_X\left(X_i\in\mathcal{N}(X)\right)\mathbf{1}_{\varphi(\hat{y}_i)\neq\varphi(y_i^*)}.\end{aligned}$$

Finally we have, after proper conditioning, considering the variability in S_i while fixing X_i first,

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_n,X}\left[\left\|g_n^*(X) - g_n(X)\right\|_{\mathcal{H}} \,\middle|\, (X_i)\in\mathbb{D}\right]\\ &\quad= 2c_\varphi k^{-1}\,\mathbb{E}_{(X_i)}\left[\sum_{i=1}^n \mathbb{P}_X\left(X_i\in\mathcal{N}(X)\right)\mathbb{E}_{(S_i)}\left[\mathbf{1}_{\varphi(\hat{y}_i)\neq\varphi(y_i^*)} \,\middle|\, (X_i)\right] \,\middle|\, (X_i)\in\mathbb{D}\right]\\ &\quad= 2c_\varphi k^{-1}\,\mathbb{E}_{(X_i),X}\left[\sum_{i=1}^n \mathbf{1}_{X_i\in\mathcal{N}(X)}\,\mathbb{P}_{(S_i)}\left(\varphi(\hat{y}_i)\neq\varphi(y_i^*) \,\middle|\, (X_i)\right) \,\middle|\, (X_i)\in\mathbb{D}\right].\end{aligned}$$

We design $\mathbb{D}$ so that the k-th nearest neighbor of any input X_i is at distance at most h of X_i, meaning the because of class separation, $y_{x_i} \in S_j$ for any $X_j \in \mathcal{N}(X_i)$. This mean that outputting $(\hat{y}_i) = (y_i^*)$ and $z_j = y_j$, will lead to an optimal error in (8.4). Now suppose that there is another solution for (8.4) such that $\hat{y}_i \neq y_i^*$, it should also achieve an optimal error, therefore it should verify $z_j = \hat{y}_j$ for all j as well as $\hat{y}_j = \hat{y}_i$ for all j such that X_j is one of the k nearest neighbors of X_i. This implies that $\hat{y}_i \in \cap_{j;X_j \in \mathcal{N}(X_i)} S_j$, which happen with probability

$$\mathbb{P}_{(S_j)_{j;X_j \in \mathcal{N}(X_i)}}(\exists z \neq y_i, z \in \cap_j S_j) \leq m\, \mathbb{P}_{S_j}(z \in S_j)^k \leq m\eta^k = m \exp(-k\,|\log(\eta)|).$$

With $m = |\mathcal{Y}|$ the number of elements in $\mathcal{Y}$. We deduce that

$$\mathbb{P}_{(S_i)}\left(\varphi(\hat{y}_i) \neq \varphi(y_i^*) \,\middle|\, (X_i)\right) \leq m \exp(-k\,|\log(\eta)|).$$

And because $\sum_{i=1}^n \mathbf{1}_{X_i \in \mathcal{N}(X)} = k$, we conclude that

$$\mathbb{E}_{\mathcal{D}_n, X}\left[\left\|g_n^*(X) - g_n(X)\right\|_{\mathcal{H}} \,\middle|\, (X_i) \in \mathbb{D}\right] \leq 2c_\varphi m \exp(-k\,|\log(\eta)|).$$

Finally, adding everything together we get

$$\mathcal{E}(f_n) \leq 8c_\varphi c_\psi \exp\left(-\frac{np}{8}\right) + 8c_\varphi c_\psi n \exp\left(-\frac{(n-1)p}{8}\right) + 8c_\varphi c_\psi m \exp\left(-k\,|\log(\eta)|\right).$$

as long as $k < (n-1)p/2$, which implies Theorem 15 as long as $n \geq 2$.

Remark 62 (Other approaches)**.** *While we have proceeded with analysis based on local averaging methods, other paths could be explored to prove convergence results of the algorithm provided* (8.4) *and* (8.5)*. For example, one could prove Wasserstein convergence of $\sum_{i=1}^n \delta_{(x_i, \hat{y}_i)}$ toward $\sum_{i=1}^n \delta_{(x_i, \hat{y}_i^*)}$, together with some continuity of the learning algorithm as a function of those distributions.*[5] *This analysis could be understood as tripartite:*

- *A disambiguation error, comparing $\hat{y}_i$ to y_i^*.*
- *A stability / robustness measure of the algorithm to learn f_n from data when substituting y_i^* by $\hat{y}_i$.*
- *A consistency result regarding f_n^* learned with (x_i, y_i^*).*

Our analysis followed a similar path, yet with the first two parts tackled jointly.

8.A.4 Proof of Proposition 59

Under the non-ambiguity hypothesis (Assumption 15), the solution of (8.3) is characterized pointwise by $f^*(x) = y_x$ for all $x \in \operatorname{supp} \nu_{\mathcal{X}}$. Similarly, under Assumption 15, we have the characterization $f^*(x) \in \cap_{S \in \operatorname{supp} \nu|_x} S$. With the notation of Definition 58, since $f^*(x)$ minimizes $z \to \mathbb{E}_{Y \sim \mu_S}[\ell(z, Y)]$ for all $S \in \operatorname{supp} \nu|_x$, it also minimizes $z \to \mathbb{E}_{S \sim \nu|_x} \mathbb{E}_{Y \sim \mu_S}[\ell(z, Y)]$.

For the second part of the proposition, we use the structured prediction framework of Ciliberto et al. (2020). Define the signed measure μ° defined as $\mu^\circ_{\mathcal{X}} := \nu_{\mathcal{X}}$ and $\mu^\circ|_x := \mathbb{E}_{S \sim \nu|_x} \mathbb{E}_{Y \sim \mu_S}[\delta_Y]$, and $f^\circ : \mathcal{X} \to \mathcal{Y}$ the solution $f^\circ \in \arg\min_{f:\mathcal{X} \to \mathcal{Y}} \mathbb{E}_{(X,Y) \sim \mu^\circ}[\ell(f(X), Y)] = \arg\min_{f:\mathcal{X} \to \mathcal{Y}} \mathbb{E}_{(X,Y) \sim \nu}\left[\mathbb{E}_{Y \sim \mu_S}[\ell(f(X), Y)]\right]$. The first part of the proposition tells us that $f^\circ = f^*$ under Assumption 15. The framework of Ciliberto et al. (2020), tells us that f° is obtained after decoding (8.9) of $g^\circ : \mathcal{X} \to \mathcal{H}$, and that if g_n° converges to g° with the L^1 norm, f_n° converges to f° in terms of the μ°-risk. Under Assumption 15 and mild hypothesis on μ°, it is possible to prove that convergence in terms of the μ°-risk implies convergence in terms of the μ-risk (for example through calibration inequality similar to Proposition 2 of Cabannes et al. (2020b)).

8.A.5 Ranking with partial ordering is a well-behaved problem

Here, we discuss building directly ξ_S to initialize our alternative minimization scheme or considering μ_S given by the definition of well-behaved problem (Definition 58). Since the existence of μ_S implying ξ_S defined as $\mathbb{E}_{Y \sim \mu_S}[\varphi(Y)]$, we will only study when ξ_S can be cast as a μ_S.

[5] The Wasserstein metric is useful to think in terms of distributions, which is natural when considering partial supervision that can be cast as a set of admissible fully supervised distributions. This approach has been successfully followed by Perchet and Quincampoix (2015) to deal with partial monitoring in games.

In ranking, we have that $\psi = -\varphi$, which corresponds to "correlation losses". In this setting, we have that $\mathrm{Span}(\varphi(\mathcal{Y})) = \mathrm{Span}(\psi(\mathcal{Y}))$. More generally, looking at a "minimal" representation of ℓ, one can always assume the equality of those spans, as what happens on the orthogonal of the intersection of those spans, does not modify the scalar product $\varphi(y)^\top \psi(z)$. Similarly, ξ_S can be restricted to $\mathrm{Span}(\psi(\mathcal{Y}))$, and therefore $\mathrm{Span}(\varphi(\mathcal{Y}))$, which exactly the image by $\mu \to \mathbb{E}_{Y\sim\mu}[\varphi(Y)]$ of the set of signed measures, showing the existence of a μ_S matching Definition 58.

8.B IQP implementation

In this section, we introduce an IQP implementation to solve for (8.4). We first mention that our alternative minimization scheme is not restricted to well-behaved problems, before motivating the introduction of the IQP algorithm in two different ways, and finally describing its implementation.

8.B.1 Initialization of alternative minimization for non-well-behaved problem

Before describing the IQP implementation to solve (8.12), we would like to stress that, even for non-well-behaved partial labeling problems, it is possible to search for smart ways to initialize variables of the alternative minimization scheme. For example, one could look at $z_i^{(0)} \in \cap_{j:x_j\in\mathcal{N}_{k_i}} S_j$, where $\mathcal{N}_k$ designs the k nearest neighbors of x_i in $(x_j)_{j\leq n}$, and k_i is chosen such that this intersection is a singleton.

8.B.2 Link with Diffrac and empirical risk minimization

Our IQP algorithm is similar to an existing disambiguation algorithm known as the Diffrac algorithm (Bach and Harchaoui, 2007; Joulin et al., 2010).[6] This algorithm was derived by implicitly following empirical risk minimization of (8.2). This approach leads to algorithms written as

$$(y_i) \in \operatorname*{arg\,min}_{(y_i)\in C_n} \inf_{f\in\mathcal{F}} \sum_{i=1}^n \ell(f(x_i), y_i) + \lambda\Omega(f),$$

for $\mathcal{F}$ a space of functions, and $\Omega : \mathcal{F} \to \mathbb{R}_+$ a measure of complexity. Under some conditions, it is possible to simplify the dependency in f (*e.g.*, Xu et al., 2004; Bach and Harchaoui, 2007). For example, if $\ell(y, z)$ can be written as $\|\varphi(y) - \varphi(z)\|^2$ for a mapping $\varphi : \mathcal{X} \to \mathcal{Y}$, *e.g.* the Kendall loss detailed in Section 8.5.4,[7] and the search of $\varphi(f) : \mathcal{X} \to \varphi(\mathcal{Y})$ is relaxed as a $g : \mathcal{X} \to \mathcal{H}$. With Ω and $\mathcal{F}$ linked with kernel regression on the surrogate functional space $\mathcal{X} \to \mathcal{H}$, it is possible to solve the minimization with respect to g as $g(x_i) = \sum_{j=1}^n \alpha_j(x_i)\varphi(y_i)$, with α given by kernel ridge regression (Ciliberto et al., 2016), and to obtain a disambiguation algorithm written as

$$\operatorname*{arg\,min}_{y_i\in S_i} \sum_{i=1}^n \Big\| \sum_{j=1}^n \alpha_j(x_i)\varphi(y_j) - \varphi(y_i)\Big\|^2.$$

This IQP is a special case of the one we will detail. As such, our IQP is a generalization of the Diffrac algorithm, and this paper provides, to our knowledge, *the first consistency result for Diffrac*.

8.B.3 Link with another determinism measure

While we have considered the measure of determinism given by (8.2), we could have considered its quadratic variant

$$\mu^\star \in \operatorname*{arg\,min}_{\mu\vdash\nu} \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{X\sim\nu_\mathcal{X}} \left[\mathbb{E}_{Y,Y'\sim\mu|_x}\left[\ell(Y, Y')\right]\right].$$

This corresponds to the right drawing of Figure 8.4. We could arguably translate it experimentally as

$$(\hat{y}_i) \in \operatorname*{arg\,min}_{(y_i)\in C_n} \sum_{i,j=1}^n \alpha_i(x_j)\ell(y_i, y_j), \tag{8.14}$$

[6] The Diffrac algorithm was first introduced for clustering, which is a classical approach to unsupervised learning. In practice, it consists of changing the constraint set $C_n = \prod S_i$ by a set of the type $C_n = \arg\max_{(y_i)\in\mathcal{Y}^n} \sum_{i,j=1}^n \mathbf{1}_{y_i\neq y_j}$ in (8.4) and (8.14), meaning that (y_i) should be disambiguated into different classes.

[7] Since $\|\varphi(y)\|$ is constant.

and still derive Theorem 15 when substituting (8.4) by (8.14). When the loss is a correlation loss $\ell(y,z) = -\varphi(y)^\top \varphi(z)$. This leads to the quadratic problem

$$(\hat{y}_i) \in \operatorname*{arg\,min}_{(y_i)\in C_n} -\sum_{i,j=1}^{n} \alpha_i(x_j)\varphi(y_i)^\top \varphi(y_j).$$

8.B.4 IQP implementation

In order to make our implementation possible for any symmetric loss $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$, on a finite space $\mathcal{Y}$, we introduce the following decomposition.

Proposition 63 (Quadratic decomposition)**.** *When $\mathcal{Y}$ is finite, any proper symmetric loss ℓ admits a decomposition with two mappings $\varphi : \mathcal{Y} \to \mathbb{R}^m$, $\psi : \mathcal{Y} \to \mathbb{R}^m$, for an $m \in \mathbb{N}$ and a $c \in \mathbb{R}$, reading*

$$\forall\, y,z \in \mathcal{Y}, \quad \ell(y,z) = \psi(y)^\top \psi(z) - \varphi(y)^\top \varphi(z) \qquad \textit{with} \qquad \|\varphi(y)\| = \|\psi(y)\| = c \tag{8.15}$$

Proof. Consider $\mathcal{Y} = y_1, \cdot, y_m$ and $L = (\ell(y_i,y_j))_{i,j\le m} \in \mathbb{R}^{m\times m}$. L is a symmetric matrix, diagonalizable as $L = \sum_{i=1}^m \lambda_i u_i \otimes u_i$, with (u_i) an orthonormal basis of $\mathbb{R}^m$, and $\lambda_i \in \mathbb{R}$ its eigenvalues. We have, with (e_i) the Cartesian basis of $\mathbb{R}^m$,

$$\ell(y_j,y_k) = L_{jk} = \left\langle e_j, Le_k \right\rangle = \sum_{i=1}^m (\lambda_i)_+ \left\langle e_j, u_i \right\rangle \langle e_k, u_i\rangle - \sum_{i=1}^m (\lambda_i)_- \left\langle e_j, u_i \right\rangle \langle e_k, u_i\rangle .$$

We build the decomposition

$$\tilde{\psi}(y_k) = \left(\sqrt{(\lambda_i)_+}\,\langle e_k, u_i\rangle\right)_{i\le m}, \qquad \text{and} \qquad \tilde{\varphi}(y_k) = \left(\sqrt{(\lambda_i)_-}\,\langle e_k, u_i\rangle\right)_{i\le m}.$$

It satisfies $\ell(y_j,y_k) = \tilde{\psi}(y)^\top\tilde{\psi}(z) - \tilde{\varphi}(y)^\top\tilde{\varphi}(z)$. We only need to show that we can consider φ of constant norm. For this, first consider $C = \max_i |\lambda_i|$, we have $\left\|\tilde{\psi}(y_k)\right\|^2 = \sum_{i=1}^m (\lambda_i)_+ \langle u_i, e_k\rangle^2 \le C\sum_{i=1}^m \langle u_i, e_k\rangle^2 = C\,\|e_k\|^2 = C$ The last equalities being due to the fact that (u_i) is orthonormal. Now, introduce the correction vector $\xi : \mathcal{Y} \to \mathbb{R}^m$, $\xi(y_i) = \sqrt{C - \left\|\tilde{\psi}(y)\right\|^2}e_i$. And consider $\varphi = \binom{\tilde{\varphi}}{\xi}$, $\psi = \binom{\tilde{\psi}}{\xi}$. By construction, ψ is of constant norm being equal to C and that $\ell(y,z) = \psi(y)^T\psi(z) - \varphi(y)^T\varphi(z)$. Finally, because $\ell(y,z) = 0$, we also have φ of constant norm. □

Using the decomposition (8.15), (8.14) reads, with $\mathbf{y} = (y_i)$

$$\hat{\mathbf{y}} \in \operatorname*{arg\,min}_{\mathbf{y}\in C_n} \sum_{i=1}^n \alpha_i(x_j)\psi(y_i)\psi(y_j) - \sum_{i=1}^n \alpha_i(x_j)\varphi(y_i)\varphi(y_j).$$

By defining the matrix $A = (\alpha_i(x_j))_{ij\le n} \in \mathbb{R}^{n\times n}$, $\Psi(\mathbf{y}) = (\psi(y_i))_{i\le n} \in \mathbb{R}^{n\times m}$ and $\Phi(\mathbf{y}) = (\varphi(y_i))_{i\le n} \in \mathbb{R}^{n\times m}$, we cast it as

$$\hat{\mathbf{y}} \in \operatorname*{arg\,min}_{\mathbf{y}\in C_n} \operatorname{Tr}\left(A\Psi(\mathbf{y})\Psi(\mathbf{y})^\top\right) - \operatorname{Tr}\left(A\Phi(\mathbf{y})\Phi(\mathbf{y})^\top\right).$$

Objective convexification. As $\alpha_i(x_j)$ is a measure of similarity between x_i and x_j, A is usually symmetric positive definite, making this objective convex in Ψ and concave in Φ. However, recalling (8.15), we have $\operatorname{Tr}\Phi\Phi^\top = \operatorname{Tr}\Psi\Psi^\top = nc$, therefore considering the spectral norm of A, we convexify the objective as

$$\hat{\mathbf{y}} \in \operatorname*{arg\,min}_{\mathbf{y}\in C_n} \operatorname{Tr}\left((\|A\|_* I + A)\Psi(\mathbf{y})\Psi(\mathbf{y})^\top\right) + \operatorname{Tr}\left((\|A\|_* I - A)\Phi(\mathbf{y})\Phi(\mathbf{y})^\top\right).$$

Considering

$$B = \begin{pmatrix} \|A\|_* I + A & 0 \\ 0 & \|A\|_* I - A \end{pmatrix} \qquad \text{and} \quad \Xi(\mathbf{y}) = \begin{pmatrix}\Psi(\mathbf{y})\\ \Phi(\mathbf{y})\end{pmatrix},$$

simplifies this objective as

$$\hat{\mathbf{y}} \in \operatorname*{arg\,min}_{\mathbf{y}\in C_n} \operatorname{Tr}\left(B\Xi(\mathbf{y})\Xi(\mathbf{y})^\top\right).$$

When parametrized by $\xi = \Xi(\mathbf{y})$, this is an optimization problem with a convex quadratic objective and "integer-like" constraint $\xi \in \Xi(C_n)$, identifying to an integer quadratic program (IQP).

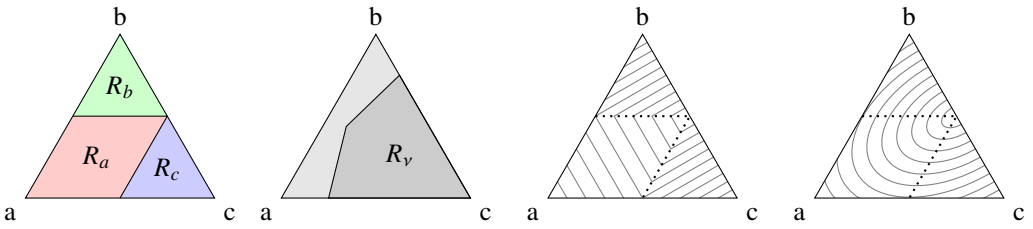

Figure 8.4: Exposition of a pointwise problem in the simplex $\Delta_{\mathcal{Y}}$, with $\mathcal{Y} = \{a, b, c\}$ and a proper symmetric loss defined by $\ell(a, b) = \ell(a, c) = \ell(b, c)/2$. (Left) Representation of the decision regions $R_z = \left\{\mu \in \Delta_{\mathcal{Y}} \,\middle|\, z \in \arg\min_{z' \in \mathcal{Y}} \mathbb{E}_{Y\sim\mu}[\ell(z, y)]\right\}$ for $z \in \mathcal{Y}$. (Middle Left) Representation of $R_\nu = \left\{\mu \in \Delta_{\mathcal{Y}} \,\middle|\, \mu \vdash \nu\right\}$ for $\nu = (5\delta_{\{a,b,c\}} + \delta_{\{c\}} + \delta_{\{a,c\}} + \delta_{\{b,c\}})/8$. (Middle Right) Level curves of the piecewise function $\Delta_{\mathcal{Y}} \to \mathbb{R}; \mu \to \min_{z\in\mathcal{Y}} \mathbb{E}_{Y\sim\mu}[\ell(z, Y)]$ corresponding to (8.2). (Right) Level curves of the quadratic function $\Delta_{\mathcal{Y}} \to \mathbb{R}; \mu \to \mathbb{E}_{Y,Y'\sim\mu}[\ell(Y, Y')]$. Our disambiguation (8.2) corresponds to minimizing the concave function represented in the middle right drawing on the convex domain represented in the middle left drawing.

Relaxation. IQP are known to be NP-hard, several tools exist in literature and optimization libraries implementing them. The most classical approach consists in relaxing the integer constraint $\xi \in \Xi(C_n)$ into the convex constraint $\xi \in \mathrm{Conv}(\Xi(C_n))$, solving the resulting convex quadratic program, and projecting back the solution toward an extreme of the convex set. Arguably, our alternative minimization approach is a better grounded heuristic to solve our specific disambiguation problem.

8.C Example with graphical illustrations

To ease the understanding of the disambiguation principle (8.2), we provide a toy example with a graphical illustration, Figure 8.4. Since (8.2) decorrelates inputs, we will consider $\mathcal{X}$ to be a singleton, in order to remove the dependency to $\mathcal{X}$. In the following, we consider $\mathcal{Y} = \{a, b, c\}$, with the loss given by

$$L = (\ell(y, z))_{y,z\in\mathcal{Y}} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix}.$$

This problem can be represented on a triangle through the embedding of probability measures reading $\xi : \Delta_{\mathcal{Y}} \to \mathbb{R}^3; \mu \to \mu(a)e_1 + \mu(b)e_2 + \mu(c)e_3$, and onto the triangle $\left\{z \in \mathbb{R}^3_+ \,\middle|\, z^\top 1 = 1\right\}$. Note that ξ can be extended from any signed measure of total mass normalized to one onto the plane $\left\{z \in \mathbb{R}^3 \,\middle|\, z^\top 1 = 1\right\}$, as well as the drawings Figure 8.4 can be extended onto the affine span of the represented triangles. The objective (8.2) reads pointwise as $\Delta_{\mathcal{Y}} \to \mathbb{R}; \mu \to \min_{i\leq 3} e_i^\top L\xi(\mu)$, while its quadratic version reads $\Delta_{\mathcal{Y}} \to \mathcal{Y}; \mu \to \xi(\mu)^\top L\xi(\mu)$. Note that while L is not definite negative, one can check that the restriction of $\mathbb{R}^3 \to \mathbb{R}; z \to z^\top Lz$ to the definition domain $\left\{z \in \mathbb{R}^3 \,\middle|\, z^\top 1 = 1\right\}$ is concave, as suggested by the right drawing of Figure 8.4.

It should be noted that (ℓ, ν) being a well-behaved partial labeling problem can be understood graphically, as having the intersection of the decision regions $\cap_{z\in S} R_z$ non-empty for any set S in the support of ν. As such, it is easy to see that our toy problem is well-behaved for any distribution ν. Formally, to match Definition 58, we can define $\mu_{\{e\}} = \delta_e$ for $e \in \{a, b, c\}$ and

$$\mu_{\{a,b\}} = .5\delta_a + .5\delta_b, \quad \mu_{\{a,c\}} = .5\delta_b + .5\delta_c, \quad \mu_{\{b,c\}} = \delta_b + \delta_c - \delta_a, \quad \mu_{\{a,b,c\}} = .5\delta_b + .5\delta_c.$$

Graphically $\xi(\mu_{\{a,b\}})$ can be chosen as any points on the horizontal dashed line on the middle right drawing of Figure 8.4 (similarly for $\xi\mu_{\{a,c\}}$), while $\xi(\mu_{\{a,b,c\}})$ has to be chosen has the intersection $.5e_2 + .5e_3$, and while $\xi(\mu_{\{b,c\}})$ has to be chosen outside the simplex on the half-line leaving $.5e_2 + .5e_3$ supported by the perpendicular bisector of $[e_2, e_3]$ and not containing e_1.

8.D Experiments

While our results are much more theoretical than experimental, out of principle, as well as for reproducibility, comparison and usage sake, we detail our experiments.

8.D.1 Interval regression - Figure 8.1

Figure 8.1 corresponds to the regression setup consisting of learning $f^* : [0,1] \to \mathbb{R}; x \to \sin(\omega x)$, with $\omega = 10 \approx 3\pi$. The dataset represented on Figure 8.1 is collected in the following way. We sample $(x_i)_{i\leq n}$ with $n = 10$, uniformly at random on $\mathcal{X} = [0,1]$, after fixing a random seed for reproducibility. We collect $y_i = f(x_i)$. We create (s_i) by sampling u_i uniformly on $[0,1]$, defining $r_i = r - \gamma \log(u_i)$, with $r = 1$ and $\gamma = 3^{-1}$, sampling c_i uniformly at random on $[0, r_i]$, and defining $s_i = y_i + \text{sign}(y_i) \cdot c_i + [-r_i, r_i]$. The corruption is skewed on purpose to showcase disambiguation instability of the baseline (8.13) compared to our method. We solve (8.4) with alternative minimization, initialized by taking $y_i^{(0)}$ at the center of s_i, and stopping the minimization scheme when $\sum_{i\leq n} |y_i^{(t+1)} - y_i^{(t)}| < \varepsilon$ for ε a stopping criterion fixed to 10^{-6}. For $x \in \mathcal{X}$, the inference (8.5) and (8.13) is done through grid search, considering, for $f_n(x)$, 1000 guesses dividing uniformly $[-6,6] \subset \mathcal{Y} = \mathbb{R}$. We consider weights α given by kernel ridge regression with Gaussian kernel, defined as

$$\alpha(x) = (K + n\lambda I)^{-1} K_x \in \mathbb{R}^n, \quad K = (k(x_i, x_j))_{i,j\leq n} \in \mathbb{R}^{n\times n}, \quad K_x = (k(x_i, x))_{i\leq n} \in \mathbb{R}^n,$$

with $k(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$, and λ a regularization parameter, and σ a standard deviation parameter. In our simulation, we fix $\sigma = .1$ based on simple considerations on the data, while we consider $\lambda \in [10^{-1}, 10^{-3}, 10^{-6}]$. The evaluation of the mean square error between f_n and f^*, which is equivalent to evaluating the risk with the regression loss $\ell(y,z) = \|y - z\|^2$, is done by considering 200 points dividing uniformly $\mathcal{X} = [0,1]$ and evaluating f_n and f^* on it. The best hyperparameter λ is chosen by minimizing this error. It leads to $\lambda = 10^{-1}$ for the baseline (8.13), and $\lambda = 10^{-6}$ for our algorithm (8.4) and (8.5). This difference in λ is normal since both methods are not estimating the same surrogate quantities. The fact that λ is smaller for our algorithm is natural as our disambiguation objective (8.4) already has a regularization effect on the solution.[8] Note that we used the same weights α for (8.4) and (8.5), which is suboptimal, but fair to the baseline, as, consequently, both methods have the same number of hyperparameters.

8.D.2 Classification - Figure 8.2

Figure 8.2 corresponds to classification problems, based on real datasets from the LIBSVM datasets repository. At the time of writing, the datasets are available at `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html`. We present results on the "DNA" and "Svmguide2" datasets, that both have 3 classes ($m = 3$), and respectively have 4000 samples with 180 features ($n = 4000$,$d = 180$) and 391 samples with 20 features ($n = 391$, $d = 20$).

In terms of *complexity*, when $\mathcal{Y} = [\![1, m]\!] = \{1, 2, \cdots, m\}$, and weights based on kernel ridge regression with Gaussian kernel as described in the last paragraph the complexity of performing inference for (8.5) and (8.13) can be done in $O(nm)$ in time and $O(n+m)$ in space, where n is the number of training samples (Nowak-Vila et al., 2019; Cabannes et al., 2020b). The disambiguation (8.4) performed with alternative minimization is done in $O(cn^2m)$ in time and in $O(n(n+m))$ in space, with c the number of steps in the alternative minimization scheme. In practice, c is really small, which can be understood since we are minimizing a concave function and each step leads to a guess on the border of the constraint domain.

Based on the dataset (x_i, y_i), we create (s_i) by sampling it accordingly to $\gamma\delta_{\{y_i\}} + 1 - \gamma\delta_{\{y, y_i\}}$, with y the most present labels in the dataset (indeed we choose the two datasets because they were not too big and presenting unequal labels proportion), and $\gamma \in [0,1]$ the corruption parameter represented in percentage on the x-axis of Figure 8.2. This skewed corruption allows distinguishing methods and invalidates the simple approach consisting of averaging candidate (AC) in set to recover y_i from s_i, which works well when data are *missing at random* (Heitjan and Rubin, 1991). We separate (x_i, s_i) in 8 folds, consider $\sigma \in d \cdot [1, .1, .01]$, where d is the dimension of $\mathcal{X}$, and $\lambda \in n^{-1/2} \cdot [1, 10^{-3}, 10^{-6}]$, where n is the number of data. We tested the different hyperparameter setup and reported the best error for each corruption parameter on Figure 8.2. Those errors are measured with the 0-1 loss, computed as averaged over the 8 folds, *i.e.* cross-validated, with standard deviation represented as error bars on the figure. The best hyperparameter generally corresponds to $\sigma = .1$ and $\lambda = 10^{-3}$ when the corruption is small and $\sigma = 1$, $\lambda = 10^{-3}$ when the corruption is big.

[8] Moreover, the analysis in Cabannes et al. (2020b) suggests that the baseline is estimating a surrogate function in $\mathcal{X} \to 2^{\mathbb{R}}$, while our method is estimating a function in $\mathcal{X} \to \mathbb{R}$, which is a much smaller function space, hence needing less regularization. However, those reflections are based on upper bounds, that might be suboptimal, which could invalidate those considerations.

Differences between cross-validated error and testing error were small, and we presented the first one out of simplicity.

In terms of *energy cost*, the experiments were run on a personal laptop that has two processors, each of them running 2.3 billion instructions per second. During experiments, all the data were stored on the random access memory of 8 GB. Experiments were run on Python, extensively relying on the NumPy library (Harris et al., 2020). The heaviest computation is Figure 8.2. Its total runtime, cross-validation included, was around 70 seconds. This paper is the result of experimentation, we evaluate the total cost of our experimentation to be three orders of magnitude higher than the cost of reproducing the final computations presented on Figure 8.1, 8.2 and 8.3. The total computational energy cost is negligible.

8.D.3 Semi-supervised learning - Figure 8.3

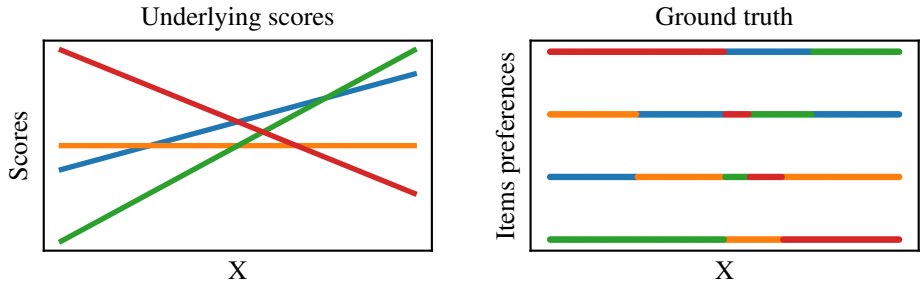

Figure 8.5: Ranking setting. We consider $\mathcal{X}$ an interval of $\mathbb{R}$, and $\mathcal{Y} = \mathfrak{S}_m$ with $m = 4$ on the figure. (Right) To create a ranking dataset, we sample randomly m lines in $\mathbb{R}^2$, embedding a value, or equivalently a score, associated to each item as a function of the input x. (Left) By ordering those lines, we create preferences between items as a function of x. On the figure, when x is small, the "red" item is preferred over the "orange" item, itself preferred over the "blue" item, itself preferred over the "green" item. While when x is big, "green" is preferred over "blue", preferred over "orange", preferred over "red". We create a partial labeling dataset by sampling $(x_i) \in \mathcal{X}^n$, and providing only partial ordering that the (y_i) follow. For example, for a small x, we might only give the partial information that "red" is preferred over "blue".

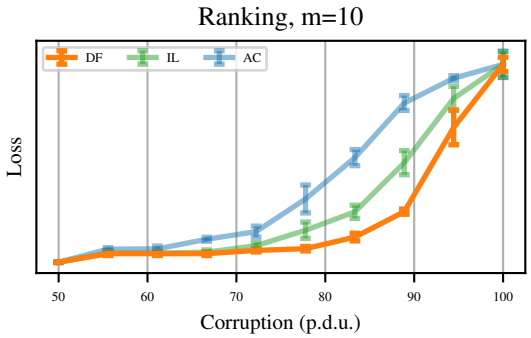

Figure 8.6: Performance of our algorithm for ranking with partial ordering. This figure is similar to Figure 8.2, but is based on the ranking problem illustrated on Figure 8.5. For this figure, we consider $m = 10$, as it is arguably the limit where the LP relaxation provided by Cabannes et al. (2020b) of the NP-hard minimum feedback arcset problem still performs well. The corruption parameter corresponds to the proportion of coordinates lost in the Kendall embedding when creating s_i from y_i. Because the Kendall embedding satisfies transitivity constraints, a corruption smaller than 50% is almost ineffective to remove any information. In this figure, we observe a similar behavior for ranking to the one observed for classification on Figure 8.2, suggesting that those empirical findings are not spurious.

On Figure 8.3, we review a semi-supervised classification problem with $\mathcal{Y} = [\![1, 4]\!]$, $\mathcal{X} = [-4.5, 4.5]^2$, $\mu_{\mathcal{X}}$ only charging $\left\{x = (x_1, x_2) \in \mathbb{R}^2 \,\middle|\, x_1^2 + x_2^2 \in \mathbb{N}^*\right\}$ and the solution $f^* : \mathcal{X} \to \mathcal{Y}$ being defined almost

everywhere as $f^*(x) = x_1^2 + x_2^2$. We collect a dataset (x_i, s_i), by sampling 2000 points θ_i uniformly at random on $[0, 1]$, as well as r_i uniformly at random in $[\![1, 4]\!] = \{1, 2, 3, 4\}$, before building $x_i = r_i \cdot (\cos(2\pi\theta_i), \sin(2\pi\theta_i)) \in \mathcal{X}$, and $s_i = \mathcal{Y}$. We add four labeled points to this dataset $x_{2001} = (-2\sqrt{3}, 2)$ with $s_{2001} = \{4\}$, $x_{2001} = (1, -2\sqrt{2})$ with $s_{2002} = \{3\}$, $x_{2001} = (-\sqrt{3}, -1)$ with $s_{2003} = \{2\}$ and $x_{2001} = (-1, 0)$ with $s_{2004} = \{1\}$. We designed the weights α in (8.4) with k-nearest neighbors, with $k = 20$, and solve this equation with a variant of alternative minimization, leading to the optimal solution $\tilde{y}_i = y_i^*$. In order to be able to compute the baseline (8.13), we design weights α for the inference task based on Nadaraya-Watson estimators with Gaussian kernel, defined as $\alpha_i(x) = \exp\left(\|x - x_i\|^2 / h\right)$, with $h = .08$. We solve the inference task on a grid of $\mathcal{X}$ composed of 2500 points, and artificially recreate the observations to make them neat and reduce the resulting PDF size. Note that it is possible to design weights α that capture the cluster structure of the data, which, in this case, will lead to a nice behavior of the baseline as well as our algorithm. Arguably, this experiment showcases a regularization property of our algorithm (8.4).

8.D.4 Ranking with partial ordering

To conclude this experiment section, we look at ranking with partial ordering. We refer to Section 8.5.4 for a clear description of this instance of partial labeling. We provide to the reader eager to use our method, an implementation of our algorithm, available online at `https://github.com/VivienCabannes/partial_labelling/`. It is based on LP relaxation of the NP-hard minimum feedback arcset problem. This relaxation was proven exact when $m \leq 6$ by Cabannes et al. (2020b). The LP implementation relies on CPLEX (IBM, 2017). As complementary experiments, we will not provide much reproducibility details, those details would be really similar to the previous paragraphs, and the curious reader could run our code instead. We present our ranking setup on Figure 8.5 and our results on Figure 8.6.

Chapter 9

Laplacian Regularization

The following is a reproduction of Cabannes et al. (2021a).

As annotations of data can be scarce in large-scale practical problems, leveraging unlabeled examples is one of the most important aspects of machine learning. This is the aim of semi-supervised learning. To benefit from the access to unlabeled data, it is natural to diffuse knowledge of labeled data to unlabeled one. This induces the use of Laplacian regularization. Yet, current implementations of Laplacian regularization suffer from several drawbacks, notably the well-known curse of dimensionality. In this paper, we provide a statistical analysis to overcome those issues, and unveil a large body of spectral filtering methods that exhibit desirable behaviors. They are implemented through (reproducing) kernel methods, for which we provide realistic computational guidelines in order to make our method usable with large amounts of data.

9.1 Introduction

In the last decade, machine learning has been able to tackle amazingly complex tasks, which was mainly allowed by computational power to train large learning models on large annotated datasets. For instance, ImageNet is made of tens of millions of images, which have all been manually annotated by humans (Deng et al., 2009). The greediness in data annotation of such a current learning paradigm is a major limitation. In particular, when annotation of data demands in-depth expertise, relying on techniques that require zillions of labeled data is not viable. This motivates several research streams to overcome the need for annotations, such as self-supervised learning for images or natural language processing (Devlin et al., 2019). Aiming for generality, semi-supervised learning is the most classical one, assuming access to a vast amount of input data, but among which only a scarce percentage is labeled. To leverage the presence of unlabeled data, most semi-supervised techniques assume a form of low-density separation hypothesis, as detailed in the recent review of van Engelen and Hoos (2020), and illustrated by state-of the-art models (Berthelot et al., 2019; Verma et al., 2019). This hypothesis assumes that the function to learn from the data varies smoothly in highly populated regions of the input space, but might vary more strongly in sparsely populated areas, or that the decision frontiers between classes lie in regions with low-density. In such a setting, it is natural to enforce constraints on the variations of the function to learn. While semi-supervised learning is an important learning framework, it has not provided as many exciting realizations as one could have expected. This might be related to the fact that it is classically approached through graph-based Laplacian, a technique that does not scale well with the dimension of the input space (Bengio et al., 2006).

Paper organization. In Section 9.2, we motivate Laplacian regularization, and recall drawbacks of naive implementations. These limitations are overcome in Section 9.3 where we expose a theoretically principled path to derive well-behaved algorithms. More precisely, we unveil a vast class of estimates based on spectral filtering. We turn to implementation in Section 9.4 where we provide realistic guidelines to ensure scalability of the proposed algorithms. Statistical properties of our estimators are stated in Section 9.5.

Contributions. They are two folds. (*i*) Statistically, we explain that Laplacian regularization can be properly leveraged based on functional space considerations, and that those considerations can be turned into concrete implementations thanks to kernel methods. As a result, we provide consistent estimators that exhibit fast

convergence rates under a low density separation hypothesis, and that, in particular, do not suffer from the curse of dimensionality. (*ii*) Computationally, we avoid dealing with large matrices of derivatives by providing a low-rank approximation that allows dealing with $n^\gamma \log(n) \times n^\gamma \log(n)$ matrices, with a parameter $\gamma \in (0, 1]$ depending on the regularity of the problem, instead of $n(d+1) \times n(d+1)$ matrices, thus cutting down to $\mathcal{O}(\log(n)^2 n^{1+2\gamma} d)$ the potential $\mathcal{O}(n^3 d^3)$ training cost.

Related work. Interplay between graph theory and machine learning were proven successful in the 2000s (Smola and Kondor, 2003). The seminal paper of Zhu et al. (2003) introduced graph-Laplacian as a transductive method in the context of semi-supervised learning. A smoothing variant was proposed by (Zhou et al., 2003), which is coherent with the fact that enforcing constraints on labeled points leads to spikes (Alaoui et al., 2016). Interestingly, graph Laplacian do converge to diffusion operators linked with the weighted Laplace Beltrami operator (Hein et al., 2007; García Trillos et al., 2019). However, these local diffusion methods are known to suffer from the curse of dimensionality (Bengio et al., 2006). That is, local averaging methods are intuitive learning methods that have been used for more than half a century (Fix and Hodges, 1951). Yet, those methods do not scale well with the dimension of the input space (Yang, 1999). This is related to the fact that to cover $[0, 1]^d$, we need ε^{-d} balls of radius ε. Interestingly, if the function to learn is m times differentiable with smooth partial derivatives, it is possible to leverage more information from function evaluations and overcome the curse of dimensionality when $m \gtrsim d$. This property is related to covering numbers (a.k.a. capacity) of Sobolev spaces (Kolmogorov and Tikhomirov, 1959), and is leveraged by (reproducing) kernel methods (Steinwart and Christmann, 2008; Caponnetto and De Vito, 2006). The crux of this paper is to apply this fact to Laplacian regularization techniques. Note that derivatives with reproducing kernel methods in machine learning have already been considered in different settings by (Zhou, 2008; Rosasco et al., 2013; Eriksson et al., 2018).

9.2 Laplacian regularization

In this section, we introduce the notations and concepts related to the semi-supervised learning regression problem, noting that most of our results extend to any convex loss beyond least-squares. We motivate and describe Laplacian regularization that will allow us to leverage the low-density separation hypothesis. We explain statistical drawbacks usually linked with Laplacian regularization, and discuss how to circumvent them.

In the following, we denote by $\mathcal{X} = \mathbb{R}^d$ the input space, $\mathcal{Y} = \mathbb{R}$ the output space, and by $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ the joint distribution on $\mathcal{X} \times \mathcal{Y}$. For simplicity, we assume that ρ has compact support. In the following, we denote by $\rho_{\mathcal{X}}$ the marginal of ρ over $\mathcal{X}$, and by $\rho|_x$ the conditional distribution of Y given $X = x$. As usual, for $p \in \mathbb{N}^*$, $L^p(\mathbb{R}^d)$ is the space of functions f such that f^p is integrable. Moreover, we define usual Sobolev spaces: for $s \in \mathbb{N}$, $W^{s,p}(\mathbb{R}^d)$ stands for the space of functions whose weak derivatives of order s-th are in $L^p(\mathbb{R}^d)$. When $p = 2$, they have a Hilbertian structure, and we denote, $H^s(\mathbb{R}^d) = W^{s,2}(\mathbb{R}^d)$ these Hilbertian spaces. Ideally, we would like to retrieve the mapping $g^* : \mathcal{X} \to \mathcal{Y}$ defined as

$$g^* = \operatorname*{arg\,min}_{g \in L^2(\rho_{\mathcal{X}})} \mathbb{E}_{(X,Y)\sim\rho} \left[\|g(X) - Y\|^2 \right] = \operatorname*{arg\,min}_{g \in L^2(\rho_{\mathcal{X}})} \left\| g - g_\rho \right\|^2_{L^2(\rho_{\mathcal{X}})} = g_\rho, \tag{9.1}$$

where $g_\rho : \mathcal{X} \to \mathcal{Y}$ is defined as $g_\rho(x) = \mathbb{E}[Y \mid X = x]$. In semi-supervised learning, we assume that we do not have access to ρ, but we have access to n independent samples $(X_i)_{i\le n} \sim \rho_{\mathcal{X}}^{\otimes n}$, among which we have n_ℓ labels $Y_i \sim \rho|_{X_i}$ for $i \le n_\ell$, with n_ℓ potentially much smaller than n. In other terms, we have n_ℓ supervised pairs $(X_i, Y_i)_{i\le n_\ell}$, and $n - n_\ell$ unsupervised samples $(X_i)_{n_\ell < i \le n}$. While we restrict ourselves to real-valued regression for simplicity, our exposition indeed applies generically to partially supervised learning. In particular, it can be used off-the-shelve to complement the approaches of (Cabannes et al., 2020b, 2021b) as we detailed in Appendix 9.A.

9.2.1 Diffusion operator $\mathcal{L}$

In order to leverage unlabeled data, we will assume that g^* varies smoothly on highly populated regions of $\mathcal{X}$, and might vary highly on low density regions. For example, this is the case when data are clustered in well separated regions of space, and labels are constant on clusters. This is captured by the fact that the Dirichlet

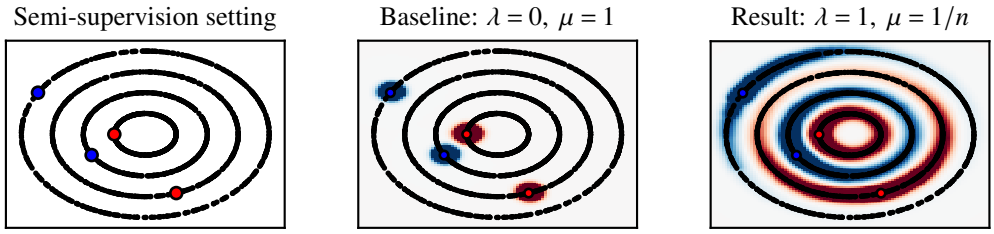

Figure 9.1: Motivating example. (Left) We suppose given $n = 2000$ points in $\mathcal{X} = \mathbb{R}^2$, represented as black dots, spanning 4 concentric circles. Among those points are $n_\ell = 4$ labeled points, with labels being either 1 represented in red, and -1 represented in blue. In this setting, it is natural to assume that g^* should be constant on each circle, which can be encoded as $\|\nabla g^*\| = 0$ on $\operatorname{supp} \rho_{\mathcal{X}}$. (Middle) Kernel ridge regression estimate based on the labeled points with Gaussian kernel of bandwidth $\sigma = .2r$, r being the radius of the innermost circle. (Right) Laplacian regularization reconstruction. The reconstruction is based on approximate empirical risk minimization with $p = n$, which ensures a computational complexity of $O(p^2nd)$, instead of $O(n^3d^3)$ needed to recover the exact empirical risk minimizer (9.5).

energy

$$\int_{\mathcal{X}} \|\nabla g^*(x)\|^2 \rho_{\mathcal{X}}(\mathrm{d}x) = \mathbb{E}_{X\sim\rho_{\mathcal{X}}}\left[\|\nabla g^*(X)\|^2\right] =: \left\|\mathcal{L}^{1/2} g\right\|^2_{L^2(\rho_{\mathcal{X}})}, \tag{9.2}$$

is assumed to be small. Because the quadratic functional (9.2) will play a crucial role in our exposition, we define $\mathcal{L}$ as the self-adjoint operator on $L^2(\rho_{\mathcal{X}})$, extending the operator on $H^1(\rho_{\mathcal{X}})$ representing this functional. Under mild assumptions on $\rho_{\mathcal{X}}$, $\mathcal{L}^{-1}$ can be shown to be a compact operator, which we will assume in the following. In essence, we will assume that if we have a lot of unlabeled data and $\left\|\mathcal{L}^{1/2} g\right\|$ can be well approximated for any function g, then we do not need a lot of labeled data to estimate correctly g^*. To illustrate this, at one extreme, if we know that $\left\|\mathcal{L}^{1/2} g^*\right\| = 0$, then g^* is known to be constant on each connected component of $\rho_{\mathcal{X}}$ so that, along with the knowledge of $\rho_{\mathcal{X}}$, only a few labeled points would be sufficient to recover perfectly g^*. We illustrate those considerations on Figure 9.1.

9.2.2 Drawbacks of naive Laplacian regularization

Following the motivations presented previously, it is natural to consider the regularized objective and solution defined, for $\lambda > 0$, as

$$\begin{aligned} g_\lambda &= \operatorname*{arg\,min}_{g\in H^1(\rho_{\mathcal{X}})} \mathbb{E}_{(X,Y)\sim\rho}\left[\|g(X) - Y\|^2\right] + \lambda\, \mathbb{E}_{X\sim\rho_{\mathcal{X}}}\left[\|\nabla g(X)\|^2_{\mathbb{R}^d}\right] \\ &= \operatorname*{arg\,min}_{g\in H^1(\rho_{\mathcal{X}})} \left\|g - g_\rho\right\|^2_{L^2(\rho_{\mathcal{X}})} + \lambda \left\|\mathcal{L}^{1/2} g\right\|^2_{L^2(\rho_{\mathcal{X}})} = (I + \lambda\mathcal{L})^{-1} g_\rho. \end{aligned} \tag{9.3}$$

This regularization has nice properties. In particular, for small λ, it can be seen as a first order approximation of the heat equation solution $e^{-\lambda\mathcal{L}} g_\rho$, which represents the temperature profile at time $t = \lambda$, instantiated with the initial profile g_ρ, and with $\rho_{\mathcal{X}}$ modeling the thermal conductivity. It also has interpretations in terms of random walk and Langevin diffusion (Pillaud-Vivien, 2020a; Klus et al., 2020). In a word, g_λ is the diffusion of g_ρ with respect to the density $\rho_{\mathcal{X}}$, which relates to the idea of diffusing labeled data with respect to the intrinsic geometry of the data, which is the idea captured by (Zhu et al., 2003).

However, from a learning perspective, (9.3) is linked with the prior that g^* belongs to $H^1(\rho_{\mathcal{X}})$, a prior that is not strong enough to overcome the curse of dimensionality as we saw in the related work section. Moreover, assuming we have enough unsupervised data to suppose known $\rho_{\mathcal{X}}$, and therefore $\mathcal{L}$, (9.3) leads to the naive empirical estimate $g_{(\text{naive})} \in \arg\min_{g:\mathcal{X}\to\mathbb{R}} \sum_{i=1}^{n_\ell} \|g(X_i) - Y_i\|^2 + n_\ell\lambda \left\|\mathcal{L}^{1/2} g\right\|^2$. While the definition of $g_{(\text{naive})}$ could seem like a great idea, in fact, such an estimate $g_{(\text{naive})}$ is known to be mostly constant and spiking to interpolate the data (X_i, Y_i) as soon as $d > 2$ (Nadler et al., 2009). This is to be related with the capacity of the space associated with the pseudo-norm $\left\|\mathcal{L}^{1/2} g\right\|$ in L^2. This capacity, related to H^1, is too large for the Laplacian regularization term to constraint $g_{(\text{naive})}$ in a meaningful way. In other terms, we need to regularize with stronger penalties.

9.2.3 Stronger regularization

In this subsection, we discuss techniques to overcome the issues encountered with $g_{(\text{naive})}$. Those techniques are based on functional space constraints or on spectral filtering techniques.

Functional spaces. A solution to overcome the capacity issue of H^1 in L^2 is to constrain the estimate of g^* to belong to a smaller functional space. In the realm of graph Laplacian, (Alaoui et al., 2016) proposed to solve this problem by considering the r-Laplacian regularization reading $\Omega_r = \int_{\mathcal{X}} \|\nabla g(X)\|^r \rho(\mathrm{d}x)$, with $r > d$. In essence, this restricts g to live in $W^{1,r}(\rho_{\mathcal{X}})$ for $r > d$, and allows avoiding spikes associated with $g_{(\text{naive})}$. However, considering high power of the gradient is likely to introduce instability (think that d is the potentially colossal dimension of the input space), and from a learning perspective, the capacity of $W^{1,r}$, which compares to the one of H^2, is still too big. In this paper, we will rather keep the diffusion operator $\mathcal{L}$, and add a second penalty to reduce the space in which we look for the solution. With $\mathcal{G}$ a Hilbert space of functions, we could look for, with $\mu > 0$ a second regularization parameter

$$g_{\lambda,\mu} = \operatorname*{arg\,min}_{g:\mathcal{G}\cap H^1(\rho_{\mathcal{X}})} \left\|g - g_\rho\right\|^2_{L^2(\rho_{\mathcal{X}})} + \lambda \left\|\mathcal{L}^{1/2} g\right\|^2_{L^2(\rho_{\mathcal{X}})} + \lambda\mu \|g\|^2_{\mathcal{G}} . \tag{9.4}$$

This formulation restricts $g_{\lambda,\mu}$ to belong both to $H^1(\rho_{\mathcal{X}})$ (thanks to the term in λ) and $\mathcal{G}$ (thanks to the term in μ). In particular the resulting space $H^1(\rho_{\mathcal{X}}) \cap \mathcal{G}$ to which $g_{\lambda,\mu}$ belongs, has a smaller capacity in L^2 than the one of $\mathcal{G}$ in L^2. In practice, we do not have access to ρ and $\rho_{\mathcal{X}}$ but to $(X_i, Y_i)_{i\le n_\ell}$ and $(X_i)_{i\le n}$, and we might consider the empirical estimator defined through empirical risk minimization

$$g_{n_\ell,n} = \operatorname*{arg\,min}_{g\in\mathcal{G}} n_\ell^{-1} \sum_{i=1}^{n_\ell} \|g(X_i) - Y_i\|^2 + \lambda n^{-1} \sum_{i=1}^{n} \|\nabla g(X_i)\|^2 + \lambda\mu \|g\|^2_{\mathcal{G}} . \tag{9.5}$$

For example, we could consider $\mathcal{G}$ to be the Sobolev space $H^m(\mathrm{d}x)$. Note the difference between $\mathcal{G}$ linked with $\mathrm{d}x$, the Lebesgue measure, that is known, and $\mathcal{L}$ linked with $\rho_{\mathcal{X}}$, the marginal of ρ over $\mathcal{X}$, that is not known. In this setting, the regularization $\|\mathcal{L}^{1/2} g\|^2 + \mu\|g\|^2_{\mathcal{G}}$ reads $\int_{\mathcal{X}} \|Dg(x)\|^2 \rho_{\mathcal{X}}(\mathrm{d}x) + \mu \int_{\mathcal{X}} \sum_{\alpha=0}^{m} \|D^\alpha g(x)\|^2 \mathrm{d}x$. Because of the size of H^m in L^2, this allows for efficient approximation of $g_{\lambda,\mu}$ based on empirical risk minimization. In particular, if $n = +\infty$, we expect the minimizer (9.5) to converge toward $g_{\lambda,\mu}$ at rates in L^2 scaling similarly to $n_\ell^{-m/d}$ in n_ℓ. To complete the picture, depending on a prior on g_ρ, $g_{\lambda,\mu}$ might exhibit good convergence properties toward g_ρ as λ and μ go to zero. This contrasts with the problem encountered with $g_{(\text{naive})}$. Those considerations are exactly what reproducing kernel Hilbert space will provide, additionally with a computationally friendly framework to perform the estimation. Note that quantities similar to $g_{\lambda,\mu}$ were considered in (Zhou, 2008; Rosasco et al., 2013).

Spectral filtering. Without looking for higher power-norm, (Nadler et al., 2009) proposed to overcome the capacity issue by considering approximation of the operator $\mathcal{L}$ based on the graph-based technique provided by (Belkin and Niyogi, 2003; Coifman and Lafon, 2006) and to reduce the search of g_{n_ℓ} on the space spanned by the first few eigenvectors of the Laplacian. In particular, on Figure 9.1, g^* could be searched in the null space of $\mathcal{L}$, that is, among functions that are constant on each connected component of $\operatorname{supp}\rho_{\mathcal{X}}$. This technique exhibits two parts, the "unsupervised" estimation of $\mathcal{L}$ that will depend on the total number of data n, and the "supervised" search for g_ρ on the first few eigenvectors of $\mathcal{L}$ that will depend on the number of labels n_ℓ. While, at first sight, this technique seems to be completely different from Tikhonov regularization (9.4), it can be cast, along with gradient descent, into the same *spectral filtering* framework (Lin et al., 2020). This point of view enables the use of a wide range of techniques offered by spectral manipulations on the diffusion operator $\mathcal{L}$.

This paper is motivated by the fact that current well-grounded semi-supervised learning techniques are implemented based on graph-based Laplacian, which is a local averaging method that does not leverage smartly functional capacity. In particular, as recalled earlier, graph-based Laplacian is known to suffer from the curse of dimensionality, in the sense that the convergence of the empirical estimator $\widehat{\mathcal{L}}$ toward the $\mathcal{L}$ exhibits a rate of convergence of order $\mathcal{O}(n^{-1/d})$ with d the dimension of the input space $\mathcal{X}$ (Hein et al., 2007). In this work, we will bypass this curse of dimensionality by looking for g in a smooth universal reproducing kernel Hilbert space, which will lead to efficient empirical estimates.

9.3 Spectral filtering with kernel Laplacian

In this section, we approach Laplacian regularization from a functional analysis perspective. We first introduce kernel methods and derivatives in reproducing kernel Hilbert space (RKHS). We then translate the considerations provided in Section 9.2.3 in the realm of kernel methods.

9.3.1 Kernel methods and derivatives evaluation maps

In this subsection, we introduce kernel methods (see (Aronszajn, 1950; Scholkopf and Smola, 2001; Steinwart and Christmann, 2008) for more details). Consider $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ a reproducing kernel Hilbert space, that is a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$ such that the evaluation functionals $L_x : \mathcal{H} \to \mathbb{R}; g \to g(x)$ are continuous linear forms for any $x \in \mathcal{X}$. Such forms can be represented by $k_x \in \mathcal{H}$ such that, for any $g \in \mathcal{H}$, $L_x(g) = \langle k_x, g\rangle_{\mathcal{H}}$. A reproducing kernel Hilbert space can alternatively be defined from a symmetric positive semi-definite kernel $k : \mathcal{X} \to \mathcal{X} \to \mathbb{R}$, that is a function such that for any $n \in \mathbb{N}$ and $(x_i)_{i\leq n} \in \mathcal{X}^n$ the matrix $(k(x_i, x_j))_{i,j}$ is symmetric positive semi-definite, by building $(k_x)_{x\in\mathcal{X}}$ such that $k(x, x') = \langle k_x, k_{x'}\rangle_{\mathcal{H}}$. From a learning perspective, it is useful to use the evaluation maps to rewrite $\mathcal{H} = \{g_\theta : x \to \langle k_x, \theta\rangle_{\mathcal{H}} \,|\, \theta \in \mathcal{H}\}$. As such, kernel methods can be seen as "linear models" with features k_x, allowing to parametrize large spaces of functions (Micchelli et al., 2006). In the following, we will differentiate θ seen as an element of $\mathcal{H}$ and g_θ seen as its embedding in L^2. To make this distinction formal, we define the embedding $S : (\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}}) \hookrightarrow (L^2(\rho_{\mathcal{X}}), \langle\cdot,\cdot\rangle_{L2}); \theta \to g_\theta$, as well as its adjoint $S^\star : L^2(\rho_{\mathcal{X}}) \to \mathcal{H}$.

Given a linear parametric model of functions $g_\theta(x) = \langle\theta, k_x\rangle_{\mathcal{H}}$, it is possible to compute derivatives of g_θ based on derivatives of the feature vector – think of $\mathcal{H} = \mathbb{R}^p$ and of $k_x = \varphi(x)$ as a feature vector with $\varphi : \mathbb{R}^d \to \mathbb{R}^p$. For $\alpha \in \mathbb{N}^d$, with $|\alpha| = \sum_{i\leq d} \alpha_i$, we have the following equality of partial derivatives, when k is $2|\alpha|$ times differentiable,

$$D^\alpha g_\theta(x) = \langle\theta, D^\alpha k_x\rangle\,, \qquad \text{where} \qquad D^\alpha = \frac{\partial^{|\alpha|}}{(\partial x_1)^{\alpha_1}(\partial x_2)^{\alpha_2}\cdots(\partial x_d)^{\alpha_d}}.$$

Here and $D^\alpha k_x$ has to be understood as the partial derivative of the mapping of $x \in \mathcal{X}$ to $k_x \in \mathcal{H}$, which can be shown to belong to $\mathcal{H}$ (Zhou, 2008). In the following, we assume that k is twice differentiable with continuous derivatives, and will make an extensive use of derivatives of the form $\partial_i k_x = \partial k_x/\partial x_i$ for $i \leq d$ and $x \in \mathcal{X}$. Note that, as well as we can describe completely the Hilbertian geometry of the space $\mathrm{Span}\{k_x \,|\, x \in \mathcal{X}\}$ through $k(x, x') = \langle k_x, k'_x\rangle$, for $x, x' \in \mathcal{X}$, we can describe the Hilbertian geometry of $\mathrm{Span}\{k_x \,|\, x \in \mathcal{X}\} + \mathrm{Span}\{\partial_i k_x \,|\, x \in \mathcal{X}\}$, through

$$\partial_{1,i}k(x, x') = \langle\partial_i k_x, k_{x'}\rangle_{\mathcal{H}}\,, \qquad \text{and} \qquad \partial_{1,i}\partial_{2,j}k(x, x') = \left\langle\partial_i k_x, \partial_j k_{x'}\right\rangle_{\mathcal{H}}\,,$$

where $\partial_{1,i}$ denotes the partial derivative with respect to the i-th coordinates of the first variable. This echoes the so-called "representer theorems".

Example 16 (Gaussian kernel)**.** *A classical kernel is the Gaussian kernel, also known as radial basis function, defined for $\sigma > 0$ as the following k, and satisfying, for $i \neq j$, the following equalities,*

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right), \qquad \partial_{1,i}\partial_{2,j}k(x, y) = -\frac{(x_i - y_i)(x_j - y_j)}{\sigma^4}k(x, y),$$
$$\partial_{1,i}k(x, y) = -\frac{(x_i - y_i)}{\sigma^2}k(x, y), \qquad \partial_{1,i}\partial_{2,i}k(x, y) = \left(\frac{1}{\sigma^2} - \frac{(x_i - y_i)^2}{\sigma^4}\right)k(x, y),$$

where x_i designs the i-th coordinates of the vector $x \in \mathcal{X} = \mathbb{R}^d$.

9.3.2 Tikhonov, spectral filtering and dimensionality reduction

Given the kernel k, its associated RKHS $\mathcal{H}$ and S the embedding of $\mathcal{H}$ in L^2, we rewrite (9.4) under its "parametrized" version

$$g_{\lambda,\mu} = S \operatorname*{arg\,min}_{\theta\in\mathcal{H}} \left\{\left\|S\theta - g_\rho\right\|^2_{L^2(\rho_{\mathcal{X}})} + \lambda\left\|\mathcal{L}^{1/2}S\theta\right\|^2_{L^2(\rho_{\mathcal{X}})} + \lambda\mu\,\|\theta\|^2_{\mathcal{H}}\right\}. \tag{9.4}$$

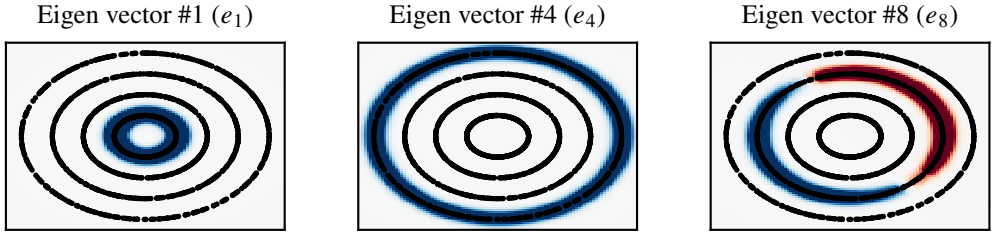

Figure 9.2: Few of the first generalized eigenvectors of $(\hat{\Sigma}; \hat{L} + \mu I)$ (with $\mu = 1/n$). The first four eigenvectors correspond to constant functions on each circle, as shown with e_1 and e_4. The few eigenvectors after correspond to second harmonics localized on a single circle as shown with e_8.

Do not hesitate to refer to Table 9.1 to keep track of notations. In the following, we will use that $\left\|\mathcal{L}^{1/2} S\theta\right\|^2_{L^2(\rho_{\mathcal{X}})} + \mu \|\theta\|^2_{\mathcal{H}} = \left\|(S^\star \mathcal{L} S + \mu I)^{1/2}\theta\right\|^2_{\mathcal{H}}$. This equality explains why we consider $\mu\lambda$ instead of μ in the last term. In the RKHS setting, the study of (9.4) unveils the three operators Σ, L, and I on $\mathcal{H}$, (indeed $g_{\lambda,\mu} = S \arg\min_{\theta\in\mathcal{H}} \left\{\theta^\star(\Sigma + \lambda L + \lambda\mu)\theta - 2\theta^\star S^\star g_\rho\right\}$) where I is the identity, and, as we detail in Appendix 9.C,

$$\Sigma = S^\star S = \mathbb{E}_{X\sim\rho_{\mathcal{X}}}\left[k_X \otimes k_X\right], \qquad \text{and} \qquad L = S^\star \mathcal{L} S = \mathbb{E}_{X\sim\rho_{\mathcal{X}}}\left[\sum_{i=1}^{d} \partial_j k_X \otimes \partial_j k_X\right]. \tag{9.6}$$

Regularization and spectral filtering have been well-studied in the inverse-problem literature. In particular, the regularization (9.4) is known to be linked with the generalized singular value decomposition of $[\Sigma; L + \mu I]$ (see, *e.g.*, Edelman and Wang (2020)), which is linked to the generalized eigenvalue decomposition of $(\Sigma, L + \mu I)$ (Golub and Loan, 2013). We derive the following characterization of (9.4), whose proof is reported in Appendix 9.D.

Proposition 64. *Let $(\lambda_{i,\mu})_{i\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$, $(\theta_{i,\mu})_{i\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ be the generalized eigenvalue decomposition of the pair $(\Sigma, L + \mu I)$, that is $(\theta_{i,\mu})$ generating $\mathcal{H}$ and such that for any $i, j \in \mathbb{N}$, $\Sigma\theta_{i,\mu} = \lambda_{i,\mu}(L + \mu I)\theta_{i,\mu}$, and $\left\langle\theta_{i,\mu}, (L + \mu I)\theta_{j,\mu}\right\rangle = \mathbf{1}_{i=j}$. (9.4) can be rewritten as*

$$g_{\lambda,\mu} = \left(\sum_{i\in\mathbb{N}} \psi(\lambda_{i,\mu}) S\theta_{i,\mu} \otimes S\theta_{i,\mu}\right) g_\rho = \sum_{i\in\mathbb{N}} \psi(\lambda_{i,\mu}) \left\langle S^\star g_\rho, \theta_{i,\mu}\right\rangle S\theta_{i,\mu}, \tag{9.7}$$

with $\psi : \mathbb{R}_+ \to \mathbb{R}; x \to (x + \lambda)^{-1}$. (9.7) should be seen as a specific instance of spectral filtering based on a filter function $\psi : \mathbb{R}_+ \to \mathbb{R}$.

Interestingly, the generalized eigenvalue decomposition of the pair $(\Sigma, L + \mu I)$ was already considered by Pillaud-Vivien (2020a) to estimate the first eigenvalue of the Laplacian. Moreover, Pillaud-Vivien (2020b) suggests leveraging this decomposition for dimensionality reduction based on the first eigenvectors of the Laplacian. As well as (9.4) contrasts with graph-based semi-supervised learning techniques, this dimensionality reduction technique contrasts with methods based on graph Laplacian provided by (Belkin and Niyogi, 2003; Coifman and Lafon, 2006). Remarkably, the semi-supervised learning algorithm that consists in using the unsupervised data to perform dimensionality reduction based on the Laplacian eigenvalue decomposition, before solving a small linear regression problem on the small resulting space, can be seen as a specific instance of spectral filtering, based on regularization by thresholding/cutting-off eigenvalue, which corresponds to $\psi : x \to x^{-1}\mathbf{1}_{x>\lambda}$ for a given threshold $\lambda > 0$ in (9.7).

9.4 Implementation

In this section, we discuss how to practically implement estimates for (9.7) based on empirical data $(X_i, Y_i)_{i\leq n_\ell}$ and $(X_i)_{n_\ell < i \leq n}$. We first review how we can approximate the integral operators of (9.6) based on data. We then discuss how to implement our methods practically on a computer. We end this section by considering approximations that allow cutting down high computational costs associated with kernel methods involving derivatives.

Algorithm 1: Empirical estimates based on spectral filtering.

Data: $(X_i, Y_i)_{i\leq n_\ell}$, $(X_i)_{n_\ell<i\leq n}$, a kernel k, a filter ψ, a regularizer μ
Result: $\hat{g}_p$ through $c \in \mathbb{R}^p$ defining $\hat{g}_p(x) = \sum_{i=1}^n c_i k(x, X_i) = k_x^\star T_a c$
Compute $S_n T_a = (k(X_i, X_j))_{i\leq n, j\leq p} \in \mathbb{R}^{n\times p}$ in $\mathcal{O}(pn)$
Compute $Z_n T_a = (\partial_{1,j} k(X_l, X_i))_{(j\leq d, l\leq n), i\leq p} \in \mathbb{R}^{nd\times p}$ in $\mathcal{O}(pnd)$
Build $T_a^\star \hat{\Sigma} T_a = n^{-1}(S_n T_a)^\top (S_n T_a)$ in $\mathcal{O}(p^2 n)$
Build $T_a^\star \hat{L} T_a = n^{-1}(Z_n T_a)^\top (Z_n T_a)$ in $\mathcal{O}(p^2 nd)$[a]
Build $T_a^\star T_a = (k(X_i, X_j))_{i,j\leq p} \in \mathbb{R}^{p\times p}$ in $\mathcal{O}(1)$ as a partial copy of $S_n T_a$
Get $(\lambda_{i,\mu}, u_{i,\mu})_{i\leq n}$ the generalized eigenelements of $(T_a^\star \hat{\Sigma} T_a, T_a^\star(\hat{L} + \mu I)T_a)$ in $\mathcal{O}(p^3)$
Get $b = T_a^\star \hat{\theta} = (n_\ell^{-1} \sum_{i=1}^{n_\ell} Y_i k(X_i, X_j))_{j\leq p} \in \mathbb{R}^p$ in $\mathcal{O}(pn_\ell)$
Return $c = \sum_{i=1}^n \psi(\lambda_i) u_i u_i^\top b \in \mathbb{R}^p$ in $\mathcal{O}(p^3)$.

[a] Building this matrix can be avoided by using the generalized singular value decomposition rather than the generalized eigenvector decomposition. Implemented with Lapack, such a procedure will also require $O(p^2 nd)$ floating point operations, but with a smaller constant in the big O (Golub and Loan, 2013).

9.4.1 Integral operators' approximation

The classical empirical risk minimization in (9.5) can be understood as the plugging of the approximate distributions $\hat{\rho} = n_\ell^{-1} \sum_{i=1}^{n_\ell} \delta_{X_i} \otimes \delta_{Y_i}$ and $\hat{\rho}_{\mathcal{X}} = n^{-1} \sum_{i=1}^n \delta_{X_i}$ instead of ρ and $\rho_{\mathcal{X}}$ in (9.4). It can also be understood as the same replacement when dealing with integral operators, leading to the three following important quantities to rewrite (9.7),

$$\hat{\Sigma} := n^{-1} \sum_{i=1}^n k_{X_i} \otimes k_{X_i},\ \hat{L} := n^{-1} \sum_{i=1}^n \sum_{j=1}^d \partial_j k_{X_i} \otimes \partial_j k_{X_i},\ \hat{\theta} := \widehat{S^\star g_\rho} := n_\ell^{-1} \sum_{i=1}^{n_\ell} Y_i k_{X_i}. \tag{9.8}$$

It should be noted that while considering n in the definition of $\hat{\Sigma}$ is natural from the spectral filtering perspective, to make it formally equivalent with the empirical risk minimization (9.5), it should be replaced by n_ℓ. (9.8) allows rewriting (9.7) without relying on the knowledge of ρ, by considering $(\hat{\lambda}_{i,\mu}, \hat{\theta}_{i,\mu})$ the generalized eigenvalue decomposition of $(\hat{\Sigma}, \hat{L})$ and considering

$$\hat{g} = \sum_{i\in\mathbb{N}} \psi(\hat{\lambda}_{i,\mu}) \left\langle \widehat{S^\star g_\rho}, \hat{\theta}_{i,\mu} \right\rangle S\hat{\theta}_{i,\mu}, \tag{9.9}$$

We present the first eigenvectors (after plunging them in L^2 through S) of the generalized eigenvalue decomposition of $(\hat{\Sigma}, \hat{L} + \mu I)$ on Figure 9.2. The first eigenvectors recover the null space of $\mathcal{L}$. This explains clearly the behavior on the right of Figure 9.1.

9.4.2 Matrix representation and approximation of operators

Currently, we are dealing with operators $(\hat{\Sigma}, \hat{L})$ and vectors (*e.g.*, $\hat{\theta}$) in the Hilbert space $\mathcal{H}$. It is natural to wonder how to represent this on a computer. The answer is the object of representer theorems (see Theorem 1 of (Zhou, 2008)), and consists in noticing that all the objects introduced are actually defined in, or operate on, $\mathcal{H}_n + \mathcal{H}_{n,\partial} \subset \mathcal{H}$, with $\mathcal{H}_n = \text{Span}\left\{k_{X_i} \,\middle|\, i \leq n\right\}$ and $\mathcal{H}_{n,\partial} = \text{Span}\left\{\partial_j k_{X_i} \,\middle|\, i \leq n, j \leq d\right\}$. This subspace of $\mathcal{H}$ is of dimension at most $n(d+1)$ and if $T : \mathbb{R}^p \to \mathcal{H}_n + \mathcal{H}_{n,\partial}$ (with $p \leq n(d+1)$) parametrizes $\mathcal{H}_n + \mathcal{H}_{n,\partial}$, our problem can be cast in $\mathbb{R}^p$ by considering the $p \times p$ matrices $T^\star \hat{\Sigma} T$ and $T^\star(\hat{L} + \mu I)T$ instead of the operators $\hat{\Sigma}$ and $\hat{L} + \mu I$. The canonical representation consists in taking $p = n(d+1)$ and considering for $c \in \mathbb{R}^{n(d+1)}$, the mapping $T_c c = \sum_{i=1}^n c_{i0} k_{X_i} + \sum_{j=1}^d c_{ij} \partial_j k_{X_i}$ (Zhou, 2008; Rosasco et al., 2013).

This exact implementation implies dealing and finding the generalized eigenvalue decomposition of $p \times p$ matrices with $p = n(d+1)$, which leads to computational costs in $\mathcal{O}(n^3 d^3)$, which can be prohibitive. Two solutions are known to cut down prohibitive computational costs of kernel methods. Both methods consist in looking for a space that can be parametrized by $\mathbb{R}^p$ for a small p and that approximates well the space $\mathcal{H}_n + \mathcal{H}_{n,\partial} \subset \mathcal{H}$. The first solution is provided by random features (Rahimi and Recht, 2007). It consists in approximating $\mathcal{H}$ with a space of small dimension $p \in \mathbb{N}$, linked with an explicit representation $\varphi : \mathcal{X} \to \mathbb{R}^p$ that approximate $k(x, x') \simeq k_\varphi(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathbb{R}^p}$. In theory, it substitutes the kernel k by k_φ. In practice, all computations can be done with the explicit feature φ.

Approximate solution. The second solution, which we are going to use in this work, consists in approximating $\mathcal{H}_n + \mathcal{H}_{n,\partial}$ by $\mathcal{H}_p = \text{Span}\left\{k_{X_i}\right\}_{i\leq p}$ for $p \leq n$. This method echoes the celebrated Nyström method (Williams and Seeger, 2000), as well as the Rayleigh–Ritz method for Sturm–Liouville problems. In essence, (Rudi et al., 2015) shows that, when considering subsampling based on leverage score, $p = n^\gamma \log(n)$, with $\gamma \in (0, 1]$ linked to the "size" of the RKHS and the regularity of the solution, is a good enough approximation, in the sense that it only downgrades the sample complexity by a constant factor. In theory, we know that the space $\mathcal{H}_p$ will converge to $\mathcal{H} = \text{Closure Span}\,\{k_x\}_{x\in\text{supp}\,\rho_\mathcal{X}}$ as p goes to infinity. In practice, it means considering the approximation mapping $T_a : \mathbb{R}^p \to \mathcal{H}; c \to \sum_{i=1}^p c_i k_{X_i}$, and dealing with the $p \times p$ matrices $T_a^\star \Sigma T_a$ and $T_a^\star L T_a$. It should be noted that the computation of $T_a^\star L T_a$ requires to multiply a $p \times nd$ matrix by its transpose. Overall, training this method can be done with $\mathcal{O}(p^2 nd)$ basic operations, and inference with this method can be done in $\mathcal{O}(p)$. The saving cost of this approximate method is huge: without compromising the precision of our estimator, we went from $O(n^3 d^3)$ run time complexities to $O(\log(n)^2 n^{1+2\gamma} d)$ computations, with γ possibly very small. Similarly, the memory cost went from $O(n^2 d^2)$ down to $O(nd + n^{2\gamma})$.[1]

9.5 Statistical analysis

In this section, we are interested in quantifying the risk of the learned mapping $\hat{g}$. We study it through the generalization bound, which consists in obtaining a bound on the averaged excess risk $\mathbb{E}_{\text{data}} \|\hat{g} - g_\rho\|_{L^2}^2$. In particular, we want to answer the following points.

1. How, and under which assumptions, Laplacian regularization boosts learning?
2. How the excess of risk relates to the number of labeled and unlabeled data?

In terms of priors, we want to leverage a low-density separation hypothesis. In particular, we can suppose that when diffusing g_ρ with $e^{-t\mathcal{L}}$ we stay close to g_ρ, or that g_ρ is supported on a finite dimensional space of functions on which $\left\|\mathcal{L}^{1/2} g\right\|$ (which measures the variation of g) is small. Both those assumptions can be made formal by assuming the g_ρ is supported by the first eigenvectors of the diffusion operator $\mathcal{L}$.

Assumption 17 (Source condition)**.** *g_ρ is supported on a finite dimensional space that is left stable by the diffusion operator $\mathcal{L}$. In other terms, if $(e_i) \in (L^2)^\mathbb{N}$ are the eigenvectors of $\mathcal{L}$, there exists $r \in \mathbb{N}$, such that $g_\rho \in \text{Span}\,\{e_i\}_{i\leq r}$.*

We will also assume that the diffusion operator $\mathcal{L}$ can be well approximated by the RKHS associated with k. In practice, under mild assumptions, *c.f.* Appendix 9.C, the eigenvectors of the Laplacian are known to be regular, in particular to belong to H^m for $m \in \mathbb{N}$ bigger than d. As such, many classical kernels would verify the following assumption.

Assumption 18 (Approximation condition)**.** *The eigenvectors (e_i) of $\mathcal{L}$ belong to the RKHS $\mathcal{H}$.*

We add one technical assumptions regarding the eigenvalue decay of the operator Σ compared to the operator L, with $\preceq$ denoting the Löwner order (*i.e.*, for A and B symmetric, $A \preceq B$ if $B - A$ is positive semi-definite).

Assumption 19 (Eigenvalue decay)**.** *There exists $a \in [0, 1]$ and $c > 0$ such that $L \preceq c\Sigma^\alpha$.*

Note that, in our setting, L is compact and bounded and Assumption 19 is always satisfied with $a = 0$. For translation-invariant kernels, such as Gaussian or Laplace kernels, based on considerations linking eigenvalue decay of operators with functional space capacities (Steinwart and Christmann, 2008), under mild assumptions, we can take $a > 1 - 2/d$. We discuss all assumptions in more detail in Appendix 9.C.

To study the consistency of our algorithms, we can reuse the extensive literature on kernel ridge regression (Caponnetto and De Vito, 2006; Lin et al., 2020).

This literature body provides an extensive picture on convergence rates relying on various filters and assumptions of capacity, a.k.a. effective dimension, and source conditions. Our setting is slightly different and showcases two specificities: (*i*) the eigenelements $(\lambda_{i,\mu}, \theta_{i,\mu})$ are dependent of μ; (*ii*) the low-rank approximation in Algorithm 1 is specific to settings with derivatives. We end our exposition with the following convergence result, proven in Appendix 9.E. Note that the dependency of p in n can be improved based on subsampling techniques that leverage expressiveness of the different (k_{X_i}) (Rudi et al., 2015).

[1] Our code is available online at `https://github.com/VivienCabannes/partial_labelling`.

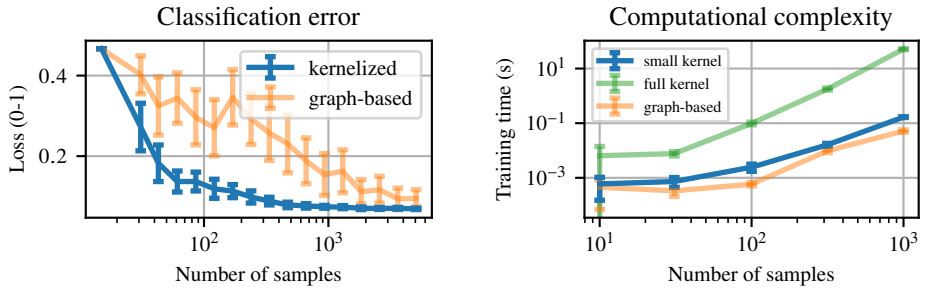

Figure 9.3: (Left) Comparison between our kernelized Laplacian method (Tikhonov regularization version with $\lambda = 1$, $\mu_n = n^{-1}$, $p = 50$) and graph-based Laplacian based on the same Gaussian kernel with bandwidth $\sigma_n = n^{-\frac{1}{d+4}} \log(n)$ as suggested by graph-based theoretical results (Hein et al., 2007). We report classification error as a function of the number of samples n. The error is averaged over 50 trials, with error bars representing standard deviations. We fixed the ratio n_ℓ/n to one tenth, and generated the data according to two Gaussian in dimension $d = 10$ with unit variance and whose centers are at distance $\delta = 3$ of each other (similar to the setting of (Castelli and Cover, 1995; Lelarge and Miolane, 2019)). Our method discovers the structure of the data much faster than graph-based Laplacian (to get a 20% error we need 40 points, while graph-based needs 700). (Right) Time to perform training with graph-based Laplacian in orange, with Algorithm 1 in blue (with the specification of the left figure), and with the naive representation in $\mathbb{R}^{n(d+1)}$ of the empirical minimizer (9.5) in green. When dealing with 1000 points, our algorithm, as well as graph-based Laplacian, can be computed in about one tenth of a second on a 2 GHz processor, while the naive kernel implementation requires 10 seconds. We show in Appendix 9.B that this cut in costs is not associated with a loss in performance.

Moreover, universal consistency results could also be provided when the RKHS is dense in H^1, as well as convergence rates for other filters and laxer assumptions which we discuss in Appendix 9.E (in particular, the source condition can be relaxed by considering the biggest $q \in (0, 1]$ such that $g \in \operatorname{im} \mathcal{L}^q$).

Theorem 16 (Convergence rates)**.** *Under Assumptions 17, 18 and 19, for $n_\ell, n \in \mathbb{N}$, when considering the spectral filtering Algorithm 1 with $\psi_\lambda : x \to (x + \lambda)^{-1}$, there exists a constant C independent of n, n_ℓ, λ, μ and p such that the estimate $\hat{g}_p$ defined in Algorithm 1 verifies*

$$\mathbb{E}_{\mathcal{D}_n}\left[\left\|\hat{g}_p - g_\rho\right\|_{L^2}^2\right] \leq C\left(\lambda^2 + \lambda\mu + \frac{\sigma_\ell^2 n_\ell^{-1} + n_\ell^{-2} + n^{-1}}{\lambda\mu} + \frac{\log(p)}{p} + \lambda\frac{\log(p)^a}{p^a}\right), \tag{9.10}$$

with σ_ℓ^2 is a variance parameter that relates to the variance of the variable $Y(I + \lambda\mathcal{L})^{-1}\delta_X$, inheriting its randomness from $(X, Y) \sim \rho$. In particular, when the ratio $r = n_\ell/n$ is fixed, with the regularization scheme $\lambda_n = \lambda_0 n^{-1/4}$, $\mu_n = \mu_0 n^{-1/4}$, for any $\lambda_0 > 0$ and $\mu_0 > 0$, and the subsampling scheme $p_n = p_0 n^s \log(n)$ for any $p_0 > 0$ and with $s = \max(1/2, 1/4a)$, there exists a constant C' independent of n and n_ℓ such that the excess of risk verifies

$$\mathbb{E}_{\mathcal{D}_n}\left[\left\|\hat{g}_p - g_\rho\right\|_{L^2}^2\right] \leq C'(n^{-1/2} + \sigma_\ell^2 n_\ell^{-1/2}). \tag{9.11}$$

Theorem 16 answers the two questions asked at the beginning of this section. In particular, it characterizes the dependency of the need for labeled data to a variance parameter linked with the diffusion of observations (X_i, Y_i) based on the density $\rho_{\mathcal{X}}$ through the operator $\mathcal{L}$. Intuitively, if the index set $I \subset \{1, 2, \cdots, n\}$ of data $(X_i)_{i\in I}$ we labeled does not change the profile of the diffusion solution $\hat{g}$, then we do not need that much labeled data – as this is the case on Figure 9.1.

Finally, Theorem 16 is remarkable in that it exhibits no dependency to the dimension of $\mathcal{X}$ in the power of n and n_ℓ. This contrasts with graph-based Laplacian methods that do not scale well with the input space dimensionality (Bengio et al., 2006; Hein et al., 2007). Indeed, Figure 9.3 shows the superiority of our method over graph-based Laplacian in dimension $d = 10$, with a mixture of Gaussian. We provide details as well as additional experiments in Appendix 9.B.

9.6 Conclusion

Diffusing information or enforcing regularity through penalties on derivatives are natural ideas to tackle many machine learning problems. Those ideas can be captured informally with graph-based techniques and finite

element differences, or captured more formally with the diffusion operator we introduced in this work. This formalization allowed us to shed light on Laplacian regularization techniques based on statistical learning considerations. In order to make our method usable in practice, we provided strong computational guidelines to cut down prohibitive costs associated with a naive implementation of our methods. In particular, we were able to develop computationally efficient semi-supervised techniques that do not suffer from the curse of dimensionality.

This work paves the way to many extensions beyond semi-supervised learning. For example, in Appendix 9.A, we describe its usefulness to the partial supervised learning problem, where minimizing the Dirichlet energy provide a learning principle, in order to bypass the restrictive non-ambiguity assumption usually made in this setup (Cour et al., 2011; Liu and Dietterich, 2014; Cabannes et al., 2020b, 2021b). Moreover, in the context of active learning, retaking the strategy of Karzand and Nowak (2020), this energy provides a computationally-effective, theoretically-grounded, data-dependent score to select the next point to query. As such, follow-ups would be of interest to see how this introductory theoretical paper makes its way into the world of concrete applications.

Appendix

9.A Extension to partially supervised learning

In this section, we first show how our work can be extended to generic semi-supervised learning problems, beyond real-valued regression. This first extension is based on the least-square surrogate introduced by Ciliberto et al. (2020) for structured prediction problems. We later show how our work can be extended to generic partially-supervised learning. This second extension is based on the work of Cabannes et al. (2020b).

9.A.1 Structured prediction and least-square surrogate

Until now, we have considered the least-square problem with $Y \in \mathbb{R}$. Indeed, our work can be extended easily to a wide class of learning problems. Consider $\mathcal{Y}$ an output space, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a loss function, and keep $\mathcal{X} \subset \mathbb{R}^d$ and $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$. Suppose that we want to retrieve

$$f^* = \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f), \qquad \text{with} \qquad \mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\rho}\left[\ell(f(X), Y)\right]. \tag{9.12}$$

Ciliberto et al. (2020) showed that as soon as ℓ can be decomposed through two mappings $\varphi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ and $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ with $\mathcal{H}_\mathcal{Y}$ a Hilbert space as $\ell(y, z) = \langle \varphi(y), \psi(z)\rangle_{\mathcal{H}_\mathcal{Y}}$, it is possible to leverage the least-square regression by considering the surrogate problem

$$g^* \in \operatorname*{arg\,min}_{g:\mathcal{X}\to\mathcal{H}_\mathcal{Y}} \mathbb{E}_{(X,Y)\sim\rho}\left[\|g(X) - \varphi(Y)\|^2_{\mathcal{H}_\mathcal{Y}}\right]. \tag{9.13}$$

This surrogate problem relates to the original one through the decoding d that relates a surrogate estimate $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ to an estimate of the original problem $f : \mathcal{X} \to \mathcal{Y}$ as $f = d(g)$ defined through, for $x \in \operatorname{supp}\rho_\mathcal{X}$,

$$f(x) = \operatorname*{arg\,min}_{z\in\mathcal{Y}} \langle \psi(z), g(x)\rangle_{\mathcal{H}_\mathcal{Y}}. \tag{9.14}$$

In the real-valued regression case, presented previously, our estimates for g_n can all be written as $g_n(x) = \sum_{i=1}^{n_\ell} \beta_i(x) Y_i$, where $\beta_i(x)$ is a function of the $(X_i)_{i\le n}$, involving the kernel k and its derivatives. Those estimates can be cast to vector-valued regression by considering coordinates-wise regression,[2] which leads to $g_n(x) = \sum_{i=1}^{n_\ell} \beta_i(x)\varphi(Y_i)$, and to the original estimates, for any $x \in \operatorname{supp}\rho_\mathcal{X}$,

$$f_n(x) \in \operatorname*{arg\,min}_{z\in\mathcal{Z}} \sum_{i=1}^{n_\ell} \beta_i(x)\ell(z, Y_i). \tag{9.15}$$

The behavior of f_n being independent of the decomposition (φ, ψ) of ℓ was referred to as the loss trick. In particular, Ciliberto et al. (2020) showed that convergence rates derived between $\|g_n - g^*\|_{L^2}$ does not change if we consider $g : \mathcal{X} \to \mathbb{R}$ or $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ and that those rates can be cast directly as convergence rates between $\mathcal{R}(f_n)$ and $\mathcal{R}(f^*)$ with $f_n = d(g_n)$ defined by (9.14). Moreover, when $\mathcal{Y}$ is a discrete output space, it is possible to get much better generalization bound on $\mathcal{R}(f_n) - \mathcal{R}^*$by introducing geometrical considerations regarding g^* and decision frontier between classes (Cabannes et al., 2021c).

[2] To parametrize functions g from $\mathcal{X}$ to $\mathcal{H}_\mathcal{Y}$, we can parametrize independently each coordinates $\langle g, e_i\rangle_{\mathcal{H}_\mathcal{Y}}$, for (e_i) a basis of $\mathcal{H}_\mathcal{Y}$, by the space $\mathcal{G}$ – note that it is possible to generalize real-valued kernel to parametrize coordinates in a joint fashion (Caponnetto and De Vito, 2006). The coordinate-wise parametrization corresponds to the tensorization $\mathcal{H}' = \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}$ and to the parametric space $\mathcal{G}' = \{x \to \Theta k_x \,|\, \Theta \in \mathcal{H}'\}$ of functions from $\mathcal{X}$ to $\mathcal{H}_\mathcal{Y}$. $\mathcal{G}'$ naturally inherits the Hilbertian structure of $\mathcal{H}'$, itself inherited from the structure of $\mathcal{H}$ and $\mathcal{H}_\mathcal{Y}$.

Example 17 (Binary classification)**.** *This framework aims at generalizing well known surrogate considerations in the case of the binary classification. Binary classification corresponds to* $\mathcal{Y} = \{-1, 1\}$*,* ℓ *the* $0-1$ *loss. In this setting,* $\mathcal{H}_\mathcal{Y} = \mathbb{R}$*,* $\varphi : \mathcal{Y} \to \mathbb{R}; y \to y$*, and* $\psi = -\varphi$*. This definition verifies* $\ell(y, z) = .5 - .5\varphi(y)^\top \varphi(z) \simeq \varphi(y)^\top \psi(z)$*. This corresponds to the usual least-square surrogate, which is* $\mathcal{R}_S(g) = \mathbb{E}[\|g(X) - Y\|^2]$*,* $g(x) = \mathbb{E}[Y \mid X = x]$ *and* $f = \operatorname{sign} g$*.*

Beyond least-squares. Considering a least-square surrogate assumes that retrieving g^* (9.13) is the way to solve the original problem (9.12) and that the low-density separation hypothesis can be expressed as Assumption 17 being verified by g^*. We would like to point out that the low-density separation could be expressed under a much weaker form, which is that there exists g such that $f^* = d(g)$ (9.14) and g verifies Assumption 17. In particular, the cluster assumption (Rigollet, 2007) could be understood as assuming that $g = \varphi(f^*)$, the trivial embedding of f^* in $\mathcal{H}_\mathcal{Y}$, is constant on clusters, which means that g belongs to the kernel of the Laplacian operator $\mathcal{L}$. Yet, $g^* : x \to \mathbb{E}[\varphi(Y)|X = x]$, which depends on the labeling noise, could be really non-smooth, even under the cluster assumption. Those considerations are related to an open problem in machine learning, which is that we do not know what is the best statistical way (and the best surrogate problem) to solve the fully supervised binary classification problem (see *e.g.* Zhang and Agarwal, 2020). However, many points introduced in the work could be retaken with other surrogate, could it be SVM (which leads to $g^* = \varphi(f^*)$, with g^* minimizing the Hinge loss), softmax regression (used in deep learning) or others.

9.A.2 Partially supervised learning

Partial supervision is a popular instance of weak supervision, which generalizes semi-supervised learning. It has been known under the name of partial labeling (Cour et al., 2011), superset learning (Liu and Dietterich, 2014), as well as learning with partial label (Grandvalet, 2002), with partial annotation (Lou and Hamprecht, 2012), with candidate labeling set (Luo and Orabona, 2010) or with multiple label (Jin and Ghahramani, 2002). It encompasses many problems such as "classification with partial labels" (Nguyen and Caruana, 2008; Cour et al., 2011), "multilabeling with missing labels" (Yu et al., 2014), "ranking with partial ordering" (Hüllermeier et al., 2008), "regression with censored data" (Tobin, 1958), "segmentation with pixel annotation" (Verbeek and Triggs, 2008; Papandreou et al., 2015), as well as instances of "action retrieval", especially on instructional videos (Alayrac et al., 2016; Miech et al., 2019). It consists, for a given input x, in not observing its label $y \in \mathcal{Y}$, but observing a set of potential labels $s \in 2^\mathcal{Y}$ that contains the labels ($y \in s$). Typically, if $\mathcal{Y}$ is the space $\mathfrak{S}_m$ of orderings between m items (*e.g.* movies on a streaming website), for a given input x (*e.g.* some feature vectors characterizing a user) s might be specified by a partial ordering that the true label y should satisfy (*e.g.* the user prefers romantic movies over action movies).

In this setting, it is natural to create consensus between the different sets giving information on $(y|x)$, which has been formalized mathematically by the infimum loss $(z, s) \in \mathcal{Y} \times 2^\mathcal{Y} \to \inf_{y \in s} \ell(z, y) \in \mathbb{R}$ for $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a specified loss on the underlying fully supervised learning problem. This leads, for $\tau \in \Delta_{\mathcal{X} \times 2^\mathcal{Y}}$ encoding the distribution generating samples (X, S), to the formulation $f^* \in \mathcal{F} = \arg\min_{f : \mathcal{X} \to \mathcal{Y}} \mathbb{E}_{(X,S)\sim\tau} \left[\inf_{Y \in S} \ell(f(X), Y)\right]$. To study this problem, a non-ambiguity assumption is usually made (Cour et al., 2011; Luo and Orabona, 2010; Liu and Dietterich, 2014; Cabannes et al., 2020b, 2021b). This is a very strong assumption to ensure that $\mathcal{F}$ is, in essence, a singleton. Highly adequate to this setting, the Laplacian regularization allows relaxing this assumption, assuming that $\mathcal{F}$ can be big, but that we can discriminate between function in $\mathcal{F}$ by looking for the smoothest one in the sense defined by the Laplacian penalty. Moreover, the loss trick (9.15) allows endowing, in an off-the-shelf fashion, the recent work of Cabannes et al. (2020b, 2021b) on the partial supervised learning problem with our considerations on Laplacian regularization.

9.B Experiments

9.B.1 Low-rank approximation

Cutting computation cost thanks to low-rank approximation, as we did by going from the naive exact empirical risk minimizer $\hat{g}$ (9.5) to the smart implementation $\hat{g}_p$ Algorithm 1, is associated with a trade-off between computational versus statistical performance. This trade-off can be studied theoretically thanks to Theorem

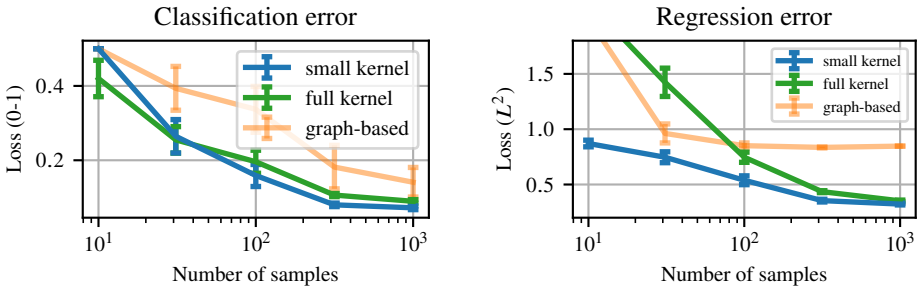

Figure 9.4: Cut in computation costs are not associated with a loss in performance. The estimate $\hat{g}_p$ Algorithm 1 (in blue), based on low-rank approximation that cut computation cuts performs as well as the exact computation of $\hat{g}$ (9.5). (Left) Classification error in the setting of Figure 9.3. (Right) Regression error in the same setting. The fact that the error of the graph-based method stalls around one, is due to the amplitude of the estimate being very small, which is coherent with behaviors described in (Nadler et al., 2009).

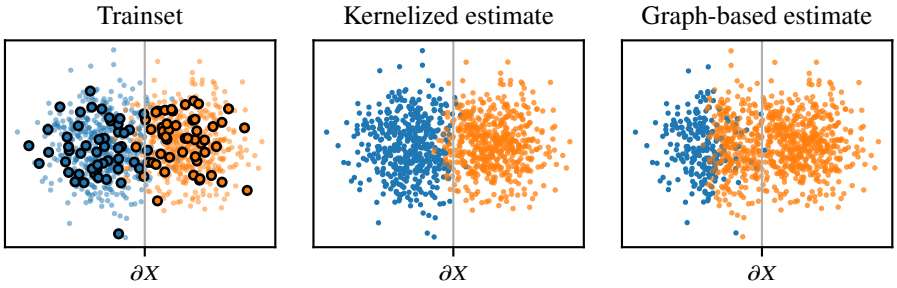

Figure 9.5: Setting of Figure 9.3 with $n = 1000$. (Left) Training set. We represent a cut of $\mathcal{X} \subset \mathbb{R}^d$ according to the two first coordinates $\{(x_1, x_2) \,|\, (x_1, x_2, \cdots, x_d) \in \mathcal{X}\}$. We have two Gaussian distributions with unit variance, one centered at $x = (0, 0, \cdots 0)$ and the other one centered at $x = (3, 0, \cdots, 0)$. One of the Gaussian distributions is associated with the blue class, the other one with the orange class. We consider $n = 1000$ unlabeled points, represented by small points, colored according to their classes, and $n_\ell = 100$ labeled points, represented in color with black edges. (Middle) Reconstruction with our kernelized Laplacian methods. Our method uncovers correctly the structure of the problem, and allows making a quite optimal reconstruction. The optimal decision frontier is illustrated by the gray line ∂X. (Right) Reconstruction with graph-Laplacian. The graph-Laplacian diffuses information too far away from what it should, leading to many incorrect guesses.

16, which shows that under mild assumptions, considering $p = n^{1/2}\log(n)$ does not lead to any loss in performance, in the sense that the convergence rates in n, the number of samples, are only changed by a constant factor. We show on Figure 9.4 that in the setting of Figure 9.3, our low-rank approximation is not associated with a loss in performance. Actually low-rank approximation can even be beneficial as it tends to lower the risk for overfitting as discussed by Rudi et al. (2015).

9.B.2 Comparison with graph-based Laplacian

One the main goal of this paper is to make people drop graph-based Laplacian methods and adopt our "kernelized" technique. As such, we would like to discuss in more detail our comparison with graph-based Laplacian. In particular, we will discuss how and why we choose the hyperparameters and the setting of Figure 9.3.

The setting of Figure 9.3 is the one of Figure 9.5, we considered two Gaussian with unit variance and whose centers are at distance $\delta = 3$ of each other. We chose Gaussian distributions as it is a well-understood setting. We chose $\delta = 3$ so that there is a mild overlap between the two distributions. For the bandwidth parameter, we considered $\sigma_n = Cn^{-\frac{1}{d+4}}\log(n)$ as this is known to be the optimal bandwidth for graph Laplacian (Hein et al., 2007). We chose $C = 1$ as this leads to σ_n of the order of δ. We chose $\lambda = 1$ to enforce Laplacian regularization and $\mu_n = 1/n$, as this is a classical regularization parameter in RKHS. Furthermore, we did not cross-validate parameters in order to be fair with graph-Laplacian that do not have

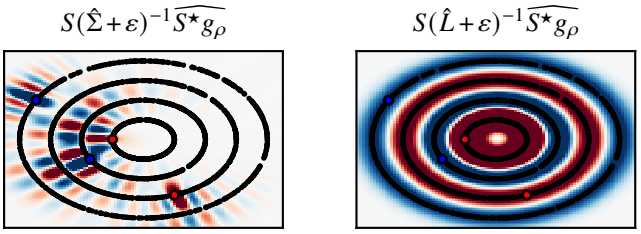

Figure 9.6: Usefulness of Laplacian regularization. We illustrate the reconstruction based on our spectral filtering techniques based on the sole use of the covariance matrix Σ on the left, and on the sole use of the Laplacian matrix L on the right. We see that the covariance matrix does not capture the geometry of the problem, which contrasts with the use of Laplacian regularization.

as many parameters as our kernel method. We compute the error in a transductive setting, retaking the exact problem and algorithm of Zhu et al. (2003). We choose $d = 10$, as we know that this is a good dimension parameter in order to illustrate the curse of dimensionality phenomenon without needing too much data.[3]

9.B.3 Usefulness of Laplacian regularization

It is natural to ask about the relevance of Laplacian regularization. To give convergence results, we have used Assumptions 17 and 18, which imply that g^* belongs to the RKHS $\mathcal{H}$, and we got convergence rates in $n_l^{1/2}$, which is not better than the rates we could get with pure kernel ridge regression. In particular, our algorithm can be split between an unsupervised part that learn the penalty $\left\|\mathcal{L}^{1/2}g\right\|^2_{L^2(\rho_\mathcal{X})}$ and a supervised part, that solve the problem of estimating g_λ from few labels (X_i, Y_i) given the penalty associated to $\mathcal{L}$. But the same method can be used for pure kernel ridge regression: unsupervised data could be leveraged to learn the covariance matrix Σ (9.6), and supervised data could be used to get $\widehat{S^\star g_\rho}$ to converge toward $S^\star g_\rho$. The same analysis would yield the same type of convergence rates. Yet the parameter σ_ℓ appearing in Theorem 16 would not be linked with the variance of $Y(I + \lambda\mathcal{L} + \lambda\mu K^{-1})^{-1}\delta_X$ but with the variance of $Y(I + \mu K^{-1})^{-1}\delta_X$. This is a key fact, the geometry of the covariance operator Σ is not supposed to be that relevant to the problem, while the one of L is. We illustrate this fact on Figure 9.6.

9.C Central operators

The paper makes an intensive use of operators. This section aims at providing details and intuitions on those operators, in order to help the reader. In particular, we discuss Assumptions 17 and 18, and we prove the equality in (9.6).

9.C.1 The diffusion operator

In this subsection, we extend on the diffusion operator, and recall its basic properties.

The diffusion operator is a well-known operator in the realm of partial differential equations. Let us assume that $\rho_\mathcal{X}$ admits a smooth density $\rho_\mathcal{X}(dx) = p(x)dx$, say $p \in \mathcal{C}^2(\mathbb{R}^d)$ that cancels outside a domain $\Omega \subset \mathbb{R}^d$. Then the diffusion operator $\mathcal{L}$ can be explicitly written, for g twice differentiable, as

$$\mathcal{L}g(x) = -\Delta g(x) + \frac{1}{p(x)} \langle \nabla p(x), \nabla g(x) \rangle .$$

[3]Note that our consistency result Theorem 16 describes a convergence regime that applies to a vast class of problems. Such a regime usually takes place after a certain number of data (depending on the value of the constant C). Before entering this regime, describing the error of our algorithm would require more precise analysis specific to each problem instance, eventually involving tools from random matrix theory.

Table 9.1: Notations

Symbol	Description
$(X_i)_{i\leq n}$	n samples of input data
$(Y_i)_{i\leq n_l}$	n_l labels
ρ	Distribution of (X, Y)
g_ρ	Function to learn (9.1)
λ, μ	Regularization parameters
$g_\lambda, g_{\lambda,\mu}$	Biased estimates (9.3, 9.4)
$\hat{g}$	Empirical estimate (9.5)
$\hat{g}_p$	Empirical estimate with low-rank approximation (Algo. 1)
$\mathcal{H}$	Reproducing kernel Hilbert space
k	Reproducing kernel
S	Embedding of $\mathcal{H}$ in L^2
$S^\star$	Adjoint of S, operating from L^2 to $\mathcal{H}$
$\Sigma = S^\star S$	Covariance operator on $\mathcal{H}$
$K = SS^\star$	Equivalent of Σ on L^2
$\mathcal{L}$	Diffusion operator (a.k.a. Laplacian)
$L = S^\star \mathcal{L} S$	Restriction of the diffusion operator to $\mathcal{H}$
g	Generic element in L^2
θ	Generic element in $\mathcal{H}$
λ_i	Generic eigenvalue
e_i	Generic eigenvector in L^2

This follows from the fact that for f once and g twice differentiable, using Stokes theorem,

$$\begin{aligned}\langle f, \mathcal{L}g\rangle_{L^2(\rho_\mathcal{X})} &= \langle \nabla f, \nabla g\rangle_{L^2(\rho_\mathcal{X})} = \langle \nabla f, p\nabla g\rangle_{L^2(\mathrm{d}x)} \\ &= \int_\mathcal{X} \operatorname{div}(fp\nabla g)\,\mathrm{d}x - \langle f, \operatorname{div}(p\nabla g)\rangle_{L^2(\mathrm{d}x)} = -\langle f, \operatorname{div}(p\nabla g)\rangle_{L^2(\mathrm{d}x)} \\ &= -\left\langle f, p^{-1}\operatorname{div}(p\nabla g)\right\rangle_{L^2(\rho_\mathcal{X})} = -\left\langle f, \operatorname{div}\nabla g + p^{-1}(\nabla p).\nabla g\right\rangle_{L^2(\rho_\mathcal{X})}.\end{aligned}$$

Note that when the distribution is uniform on Ω, the diffusion operator is exactly the opposite of the usual Laplacian operator Δ. As for the Laplacian case, it can be shown that under mild assumption on p, whose smoothness properties directly translates to the smoothness properties of the boundary of Ω, that the diffusion operator $\mathcal{L}$ has a compact resolvent (that is, for $\lambda \notin \operatorname{spec}(\mathcal{L})$, $(\mathcal{L}+\lambda I)^{-1}$ is compact). This is a standard result implied by a standard version of the famous Rellich-Kondrachov compactness embedding theorem: $H^2(\Omega)$ is compactly injected in $L^2(\Omega)$ whenever Ω is a bounded open with C^2-boundaries.

In such a setting, we can consider the eigenvalue decomposition of $\mathcal{L}^{-1}$, that is, $(\lambda_i, e_i) \in (\mathbb{R}_+ \times L^2)^\mathbb{N}$, with $(e_i)_{i\in\mathbb{N}}$ an orthonormal basis of L^2 and $(e_i)_{i\leq \dim\ker\mathcal{L}}$ generating the null space of $\mathcal{L}$, with the convention $\lambda_i = M$ for $i \leq \dim\ker\mathcal{L}$, with M an abstraction representing $+\infty$, and (λ_i) decreasing toward zero afterwards. This decomposition reads

$$\mathcal{L}^{-1} = \sum_{i\in\mathbb{N}} \lambda_i e_i \otimes e_i. \tag{9.16}$$

Note that the fact that all the (λ_i) are positive, is due to the fact that $\mathcal{L}^{-1}$ is the inverse of a positive self-adjoint operator. As a consequence, the diffusion operator has discrete spectrum, and can be written as

$$\mathcal{L} = \sum_{i\in\mathbb{N}} \lambda_i^{-1} e_i \otimes e_i. \tag{9.17}$$

In such a setting, the kernel-free Tikhonov regularization (9.3) reads

$$g_\lambda = \sum_{i\in\mathbb{N}} \psi(\lambda_i) \left\langle g_\rho, \lambda_i^{1/2} e_i\right\rangle_{L^2} \lambda_i^{1/2} e_i, \tag{9.18}$$

with $\psi : x \to (x+\lambda)^{-1}$, and the convention $M\psi(M) = 1$.

9.C.2 Regularity of the eigenvectors of the diffusion operator

In this subsection, we extend on the regularity assumed in Assumption 18.

Introducing the kernel k and its associated RKHS $\mathcal{H}$ is useful when the eigenvectors of $\mathcal{L}$ can be well approximated by functions in $\mathcal{H}$. In applications, people tend to go for kernels that are translation-invariant, which implied that the RKHS $\mathcal{H}$ is made of smooth functions, could it be analytical functions (for the Gaussian kernel) or functions in H^m (for Sobolev kernels). As a consequence, we should investigate the regularity of those eigenvectors. Indeed, if ρ derives from a Gibbs potential, that is $\rho(\mathrm{d}x) = e^{-V(x)}\,\mathrm{d}x$, the eigenvectors of $\mathcal{L}$ can be shown to inherit from the smoothness of the potential V (Pillaud-Vivien, 2020b). For example if V belongs to H^m, and $H^m \subset \mathcal{H}$, we expect (e_i) to belong to $\mathcal{H}$, thus verifying Assumption 18.

Counter-example and beyond Assumption 18. Note that if ρ has several connected components of non-empty interiors, the null space of $\mathcal{L}$ is made of functions that are constants on each connected component of $\operatorname{supp}\rho_{\mathcal{X}}$. Those functions are not analytic. In such a setting, the Gaussian kernel is not sufficient for Assumption 18 to hold, and one should favor kernel associated with richer functional space such as the Laplace kernel or the neural tangent kernel (Chen and Xu, 2021). However, as illustrated by Figure 9.2, e_i not belonging to $\mathcal{H}$ does not mean that e_i can not be well approximated by $\mathcal{H}$. Indeed, it is well known that the approximation power of $\mathcal{H}$ for e_i can be measure in the biggest power p such that $e_i \in \operatorname{im} K^p$ (Caponnetto and De Vito, 2006), where $K = SS^*$. Assumption 18 corresponds to $p = 1/2$, but it should be seen as a specific instance of more generic approximation conditions.

Handling constants in RKHS. Finally, note also that many RKHS do not contain constant functions, and therefore might not contain the constant function e_0 (although we are only looking for equality in the support of $\rho_{\mathcal{X}}$), however this specific point with e_0 can easily be circumvented either by assuming that g_ρ has zero mean, either by centering the covariance matrices Σ and $\hat{\Sigma}$ (Pillaud-Vivien, 2020b). This relates with the usual technique for SVM consisting in adding an unpenalized bias (Steinwart and Christmann, 2008).

9.C.3 Low-density separation

In this subsection, we discuss how Assumption 17 relates to the idea of low-density separation.

Low-variation intuition. The low-density separation supposes that the variations of g^* take place in region with low-density, so that $\left\|\mathcal{L}^{1/2} g^*\right\| / \|g^*\|$ is small. As such, using Courant-Fischer principle, Assumption 17 can be reformulated as g^* belonging to the space

$$\operatorname{Span}\{e_i\}_{i\le r} = \operatorname*{arg\,min}_{\substack{\mathcal{F}\subset L^2;\\ \dim\mathcal{F}=r}} \max_{g\in\mathcal{F}} \frac{\left\|\mathcal{L}^{-1/2} g\right\|_{L^2}^2}{\|g\|_{L^2}^2}.$$

In other terms, Assumption 17 can be restated as g^* belonging to a finite dimensional space that minimizes a measure of variation given by the Dirichlet energy.

To tell the story differently, suppose that we are in a classification setting, *i.e.* $Y \in \{-1, 1\}$, and that the $\operatorname{supp}\rho_{\mathcal{X}}$ is connected. Then we know that the null space of $\mathcal{L}$ is made of constant functions. Then the first eigenvector e_2 of $\mathcal{L}$ is a function that is orthogonal to constants. Hence, e_2 is a function that changes its sign and that is "balanced" in the sense that $\mathbb{E}[e_2] = 0$ – *i.e.* if $e_2(x) = \mathbb{E}_\mu[Y \mid X = x]$ for some measure μ, we have $\mathbb{E}_\mu[Y] = 0$, meaning that classes are "balanced". Moreover, in order to minimize $\left\|\mathcal{L}^{1/2} e_2\right\|$, the variations of e_2 should take place in low-density regions of $\mathcal{X}$.

Diffusion intuition. Finally, as $\mathcal{L}$ is a diffusion operator, we also have an interpretation of Assumption 17 in terms of diffusion. Consider (λ_i, e_i) the eigenelements of (9.17). The diffusion of g_ρ according the density $\rho_{\mathcal{X}}$ can be written as, for $t \in \mathbb{R}$,

$$g_t = e^{-t\mathcal{L}} g_\rho = \sum_{i\in\mathbb{N}} e^{-t\lambda_i^{-1}} \left\langle g_\rho, e_i \right\rangle e_i.$$

This diffusion will cut off the high frequencies of g_ρ that corresponds to $\langle g_\rho, e_i \rangle$ for big i, and big λ_i^{-1}. Indeed, the difference between the diffusion and the original g_ρ can be measured as

$$\left\| g_t - g_\rho \right\|_{L^2}^2 = \sum_{i \in \mathbb{N}} (e^{-t\lambda_i^{-1}} - 1)^2 \langle g_\rho, e_i \rangle^2 = \sum_{i \in \mathbb{N}} t^2 \lambda_i^{-2} \langle g_\rho, e_i \rangle^2 + o(t^2 \lambda_i^{-2}).$$

Hence, assuming that g_ρ is supported on a few of the first eigenvectors of $\mathcal{L}$, can be rephrased as saying that the diffusion of g_ρ does not modify it too much.

The variance σ_ℓ. Theorem 16 shows that the need for labels depends on the variance parameter σ_ℓ^2. It is natural to wonder how this parameter relates to the low-density hypothesis. As we discussed, this parameter is linked to the variance of $Z = Y(I + \lambda\mathcal{L})^{-1}\delta_X$. We can separate the variability of this variable due to X and the variability due to Y

$$Z = Z_X + Z_Y, \qquad \text{with} \qquad Z_X = (I + \lambda\mathcal{L})^{-1} g_\rho(X)\delta_X, \quad Z_Y = (I + \lambda\mathcal{L})^{-1}(Y - \mathbb{E}[Y \mid X])\delta_X.$$

As such we see that this variance depends on the structure of the density $\rho_\mathcal{X}$ with the variance of $(I+\lambda\mathcal{L})^{-1}\delta_X$, and the labeling noise with the variance of $(Y \mid X)$. The low-density separation does not tell us anything about the level of noise in Y or the diffusion structure linked with $\rho_\mathcal{X}$, but additional hypotheses could be made to characterize those.

9.C.4 Kernel operators

In this subsection, we define formally the operators S and Σ.

We now turn toward operators linked with the Hilbert space $\mathcal{H}$. Recall that for $k : \mathcal{X} \to \mathcal{X} \to \mathbb{R}$ a kernel, $\mathcal{H}$ is defined the closure of the span of the elements k_x under the scalar product $\langle k_x, k_{x'} \rangle = k(x, x')$. In particular, $\|k_x\|_\mathcal{H}^2 = k(x, x)$. $\mathcal{H}$ parametrizes a vast class of functions in $\mathbb{R}^\mathcal{X}$ through the mapping

$$\begin{array}{rccc} S: & \mathcal{H} & \to & \mathbb{R}^\mathcal{X} \\ & \theta & \to & (\langle k_x, \theta \rangle)_{x \in \mathcal{X}}. \end{array}$$

Under mild assumptions, S maps $\mathcal{H}$ to a space of functions belonging to L^2.

Proposition 65. *When $x \to k(x, x)$ belongs to $L^1(\rho_\mathcal{X})$, S is a continuous mapping from $\mathcal{H}$ to $L^2(\rho_\mathcal{X})$. This is particularly the case when $\rho_\mathcal{X}$ has compact support and k is continuous.*

Proof. Consider $\theta \in \mathcal{H}$, we have

$$\begin{aligned} \|S\theta\|_{L^2}^2 &= \int_\mathcal{X} \langle k_x, \theta \rangle^2 \rho_\mathcal{X}(\mathrm{d}x) \leq \int_\mathcal{X} \langle k_x, \theta \rangle_\mathcal{H}^2 \rho_\mathcal{X}(\mathrm{d}x) \leq \int_\mathcal{X} \|k_x\|_\mathcal{H}^2 \|\theta\|_\mathcal{H}^2 \rho_\mathcal{X}(\mathrm{d}x) \\ &= \|\theta\|_\mathcal{H}^2 \int_\mathcal{X} k(x, x) \rho_\mathcal{X}(\mathrm{d}x) = \|\theta\|_\mathcal{H}^2 \|x \to k(x, x)\|_{L^1}. \end{aligned}$$

Moreover, when $\rho_\mathcal{X}$ has compact support and k is continuous, k is bounded on the support of $\rho_\mathcal{X}$ therefore $x \to k(x, x)$ belongs to L^1. □

As a continuous operator from the Hilbert space $\mathcal{H}$ to the Hilbert space L^2, S is naturally associated with many linear structures: in particular its adjoint $S^\star$, but also the self-adjoint operators $K = SS^\star$ and $\Sigma = S^\star S$.

Proposition 66. *The adjoint of S is defined as*

$$\begin{array}{rccc} S^\star: & L^2 & \to & \mathcal{H} \\ & g & \to & \int_\mathcal{X} g(x) k_x \rho_\mathcal{X}(\mathrm{d}x) = \mathbb{E}_{X \sim \rho_\mathcal{X}}[g(X) k_X]. \end{array}$$

To S is associated the kernel self-adjoint operator on L^2

$$\begin{array}{rccc} K := SS^\star: & L^2 & \to & L^2 \\ & g & \to & (x \to \int_\mathcal{X} k(x, x') g(x') \rho_\mathcal{X}(\mathrm{d}x')), \end{array}$$

as well as the (non-centered) covariance on $\mathcal{H}$, $\Sigma := S^\star S = \mathbb{E}_{X \sim \rho_\mathcal{X}}[k_X \otimes k_X]$.

Proof. We shall prove the equality defining those operators. Consider $\theta \in \mathcal{H}$ and $g \in L^2$, we have

$$\left\langle S^\star g, \theta \right\rangle_{\mathcal{H}} = \langle g, S\theta \rangle_{L^2} = \mathbb{E}_{X\sim\rho_{\mathcal{X}}}[g(X)\,\langle k_X, \theta\rangle_{\mathcal{H}}] = \left\langle \mathbb{E}_{X\sim\rho_{\mathcal{X}}}[g(X)k_X], \theta \right\rangle_{\mathcal{H}}.$$

We also have, for $x \in \mathcal{X}$,

$$(SS^\star g)(x) = \left\langle k_x, \mathbb{E}_{X\sim\rho_{\mathcal{X}}}[g(X)k_X] \right\rangle_{\mathcal{H}} = \mathbb{E}_{X\sim\rho_{\mathcal{X}}}[g(X)\,\langle k_x, k_X\rangle_{\mathcal{H}}] = \mathbb{E}_{X\sim\rho_{\mathcal{X}}}[g(X)k(X,x)].$$

Finally, we have

$$S^\star S\theta = \mathbb{E}_{X\sim\rho_{\mathcal{X}}}[S\theta(X)k_X] = \mathbb{E}_{X\sim\rho_{\mathcal{X}}}[\langle\theta, k_X\rangle_{\mathcal{H}}\, k_X] = \mathbb{E}_{X\sim\rho_{\mathcal{X}}}[k_X \otimes k_X]\theta.$$

This provides the last of all the equalities stated above. □

The functional space $\mathcal{H}$. In the main text, we have written everything in terms of θ, highlighting the parametric nature of kernel methods. This made it easier to dissociate the norm on functions derived from $\mathcal{H}$ and the one derived from L^2 or H^1. In literature, people tend to keep everything in terms of functions $g_\theta = S\theta$ without even mentioning the dependency in θ. Such a setting consists in considering directly the RKHS $\mathcal{H}$ whose scalar product is defined for $g, g' \in (\ker K)^\perp$ by $\langle g, g'\rangle_{\mathcal{H}} = \left\langle g, K^{-1}g' \right\rangle_{L^2}$.

9.C.5 Derivative operators

In this subsection, we extend on derivatives in RKHS, and we formally define the operator L.

As well as evaluation maps can be represented in $\mathcal{H}$, under mild assumptions, derivative evaluation maps can benefit from such a property. Indeed, for $g_\theta = S\theta$, $x \in \mathcal{X}$ and $u \in \mathcal{B}_{\mathcal{X}}(0,1)$ a unit vector, we have

$$\partial_u g_\theta(x) = \lim_{t\to 0} \frac{g_\theta(x+tu) - g_\theta(x)}{t} = \lim_{t\to 0} \frac{\langle\theta, k_{x+tu}\rangle_{\mathcal{H}} - \langle g, k_x\rangle_{\mathcal{H}}}{t} = \lim_{t\to 0} \left\langle \theta, \frac{k_{x+tu} - k_x}{t} \right\rangle_{\mathcal{H}}$$

As a linear combination of elements in $\mathcal{H}$, the difference quotient evaluation map $t^{-1}(k_{x+tu} - k_x)$ belongs to $\mathcal{H}$ and has a norm

$$\left\| \frac{k_{x+tu} - k_x}{t} \right\|_{\mathcal{H}}^2 = \frac{k(x+tu, x+tu) - 2k(x+tu, x) + k(x,x)}{t^2}.$$

In order for the limit when t goes to zero to belong to $\mathcal{H}$, we see the importance of k to be twice differentiable. This limit $\partial_u k_x$, whose existence is proven formally by Zhou (2008), provides a derivative evaluation map in the sense that

$$\partial_u g_\theta(x) = \langle\theta, \partial_u k_x\rangle_{\mathcal{H}}.$$

From this equality, we derive that $\partial_{1i}k(x,x') = \langle k_{x'}, \partial_i k_x\rangle$, and recursively that $\left\langle \partial_i k_x, \partial_j k_{x'} \right\rangle = \partial_{1i}\partial_{2j}k(x,x')$.

Similarly to the operator S, we can introduce the operators Z_i for $i \in [1,d]$, defined as

$$\begin{array}{rccc} Z_i: & \mathcal{H} & \to & \mathbb{R}^{\mathcal{X}} \\ & \theta & \to & (\langle\partial_i k_x, \theta\rangle_{\mathcal{H}})_{i\le d} \end{array}.$$

Once again, under mild assumptions, $\operatorname{im} Z_i$ inherit from a Hilbertian structure.

Proposition 67. *When $x \to \partial_{1i}\partial_{2i}k(x,x)$ belongs to $L^1(\rho_{\mathcal{X}})$, Z_i is a continuous mapping from $\mathcal{H}$ to $L^2(\rho_{\mathcal{X}})$. This is particularly the case when $\rho_{\mathcal{X}}$ has compact support and k is twice differentiable with continuous derivatives.*

Proof. Consider $\theta \in \mathcal{H}$, similarly to before, we have

$$\|Z\theta\|_{L^2}^2 = \int_{\mathcal{X}} \langle\partial_i k_x, \theta\rangle^2\, \rho_{\mathcal{X}}(\mathrm{d}x) \le \|\theta\|_{\mathcal{H}}^2 \int_{\mathcal{X}} \|\partial_i k_x\|_{\mathcal{H}}^2\, \rho_{\mathcal{X}}(\mathrm{d}x) = \|\theta\|_{\mathcal{H}}^2 \left\| x \to \partial_{1,i}\partial_{2,i}k(x,x) \right\|_{L^1}.$$

Moreover, when $\rho_{\mathcal{X}}$ has compact support and $\partial_{1,i}\partial_{2,i}k$ is continuous, $\partial_{1,i}\partial_{2,i}k$ is bounded on the support of $\rho_{\mathcal{X}}$ therefore $x \to \partial_{1,i}\partial_{2,i}k$ belongs to L^1. □

Among the linear operators that can be built from Z_i, in the theoretical part of this paper, we are mainly interested in $Z_i^\star Z_i$. In the empirical part however, we might be interested in $Z_i Z_j^\star$ as well as $Z_i S^\star$ as it appears in Algorithm 1 (where the notation Z_n there has to be understood as the empirical version of $Z = [Z_1; \cdots; Z_d]$).

Proposition 68. *The Dirichlet energy on $\mathcal{H}$ can be represented through the operator*

$$S^\star \mathcal{L} S = \sum_{i=1}^d Z_i^\star Z_i = \sum_{i=1}^d \mathbb{E}_{X\sim\rho_\mathcal{X}}[\partial_i k_X \otimes \partial_i k_X].$$

Proof. Let $\theta \in \mathcal{H}$ and $g_\theta = S\theta$, we have

$$\begin{aligned}\langle g_\theta, \mathcal{L} g_\theta\rangle_{L^2} = \left\langle \theta, S^\star \mathcal{L} S\theta\right\rangle_\mathcal{H} &= \mathbb{E}_{X\sim\rho_\mathcal{X}}\left[\|\nabla g_\theta(X)\|^2\right] = \sum_{i=1}^d \mathbb{E}_{X\sim\rho_\mathcal{X}}\left[(\partial_i g_\theta(X))^2\right] \\ &= \sum_{i=1}^d \mathbb{E}_{X\sim\rho_\mathcal{X}}\left[\langle \partial_i k_X, \theta\rangle_\mathcal{H}^2\right] = \sum_{i=1}^d \|Z_i\theta\|_{L^2}^2 = \left\langle \theta, \sum_{i=1}^d Z_i^\star Z_i \theta\right\rangle_\mathcal{H} \\ &= \sum_{i=1}^d \mathbb{E}_{X\sim\rho_\mathcal{X}}\left[\langle \theta, (\partial_i k_X \otimes \partial_i k_X)\theta\rangle_\mathcal{H}\right] = \left\langle \theta, \sum_{i=1}^d \mathbb{E}_{X\sim\rho_\mathcal{X}}\left[\partial_i k_X \otimes \partial_i k_X\right]\theta\right\rangle_\mathcal{H}.\end{aligned}$$

Since the three operators are self-adjoint, and they all represent the same quadratic form, they are equals. □

9.C.6 Relation between Σ and L

In this subsection, we discuss the relation between Σ and L and show that we can expect the existence of $a \in (1-2/d, 1]$ and $c > 0$ such that $L \preceq c\Sigma^a$.

Informal capacity considerations. We want to compare Σ and L, as $L \preceq c\Sigma^a$ with the biggest a possible. This depends on how fast the eigenvalues are decreasing, which is linked to the entropy numbers of those two compact operators. Those entropy numbers are linked with the capacity of the functional spaces $\left\{g \in L^2 \,\middle|\, \left\|K^{-1/2}g\right\|_{L^2} < \infty\right\}$ and $\left\{g \in L^2 \,\middle|\, \left\|K^{-1/2}\mathcal{L}^{-1/2}g\right\|_{L^2} < \infty\right\}$. The first space is the reproducing kernel Hilbert space linked with k, the second space is, roughly speaking, a space of functions whose integral belongs to the first space. As such, if the first space is $\mathcal{H}^m$, the second is $\mathcal{H}^{m-1}$, and we can consider $a = (m-1)/m$. Because we are considering kernels, we have $m > d/2$ (this to make sure that the evaluation functionals $L_X : f \to f(x)$ are continuous), so that $a > 1 - 2/d$. Without trying to make those "algebraic" considerations more formal, we will give an example on the torus.

Translation-invariant kernel and Fourier transform. Consider $L^2([0,1]^d, \mathrm{d}x)$ the space of periodic functions in dimension d, square integrable against the Lebesgue measure on $[0,1]^d$. For simplicity, we will suppose that $\rho_\mathcal{X}$ is the Lebesgue measure on $[0,1]^d$. Consider a translation invariant kernel

$$k(x,y) = q(x-y) \qquad \text{for } q : \mathbb{R}^d \to \mathbb{R} \text{ that is one periodic.}$$

In this setting, the operator K, operating on L^2, is the convolution by q, that is

$$K : \begin{array}{ccc} L^2 & \to & L^2 \\ g & \to & q * g \end{array}, \qquad \text{hence} \qquad \widehat{Kg} = \hat{q}\hat{g}.$$

Where we have used the fact that convolutions can be represented by a product in the Fourier domain. Note that, from Böchner theorem, we know that k being positive definite implies that the Fourier transform of q exists and is not negative. Let us define the Fourier coefficient and the inverse Fourier transform as

$$\forall\, \omega \in \mathbb{Z}^d, \quad \hat{g}(\omega) = \int_{[0,1]^d} g(x) e^{-2i\pi\omega^\top x}\,\mathrm{d}x, \quad \text{and} \quad \forall\, x \in [0,1]^d, \quad g(x) = \sum_{\omega\in\mathbb{Z}^d} e^{2i\pi\omega^\top x}\hat{g}(\omega).$$

K being a convolution operator, it is diagonalizable with eigenelements $(\hat{q}(\omega), x \to e^{2i\pi\omega^\top x})_{\omega\in\mathbb{Z}^d}$. From this, we can make explicit many of our abstract operators. First of all, using Perceval's theorem,

$$\|g\|_\mathcal{H}^2 = \left\langle g, K^{-1}g\right\rangle_{L^2} = \sum_{\omega\in\mathbb{Z}^d} \frac{|\hat{g}(\omega)|^2}{\hat{q}(\omega)}.$$

Hence, we can parametrize $\mathcal{H}$ with $(\theta_\omega)_{\omega\in\mathbb{Z}^d} \in \mathbb{C}^{\mathbb{Z}^d}$ and the ℓ^2-metric, where $\theta_\omega = \hat{g}(\omega)/\sqrt{\hat{q}(\omega)}$ and

$$(S\theta)(x) = g_\theta(x) = \sum_{\omega\in\mathbb{Z}^d} e^{2i\pi\omega^\top x}\sqrt{\hat{q}(\omega)}\theta_\omega.$$

Note that this is not the usual parametrization of $\mathcal{H}$ by elements $\theta \in \mathcal{H}$ as $(\mathbb{C}^{\mathbb{Z}^d}, \ell^2)$ is not a space of functions. However, such a parametrization of $\mathcal{H}$ does not change any of the precedent algebraic considerations on the operators S, Σ, K, and L.

Diffusion operator and Fourier transform. As well as convolution operators are well represented in the Fourier domain, derivation operators are. Indeed, when g is regular, we have

$$\left\|\mathcal{L}^{1/2}g\right\|_{L^2}^2 = \|\nabla g\|_{L^2}^2 = \sum_{j=1}^d \left\|\partial_j g\right\|_{L^2}^2 = \sum_{j=1}^d \sum_{\omega\in\mathbb{Z}^d} \omega_j^2 \left|\hat{g}(\omega)\right|^2.$$

As a consequence, using the expression of $S\theta$, we have

$$\Sigma\theta = \sum_{\omega\in\mathbb{Z}^d} \hat{q}(\omega)\theta_\omega, \qquad \text{while} \qquad L\theta = \sum_{\omega\in\mathbb{Z}^d} \|\omega\|_2^2\, \hat{q}(\omega)\theta_\omega, \quad \text{where} \quad \|\omega\|_2^2 = \sum_{j=1}^d \omega_i^2.$$

In this setting, the eigenelements of Σ are $(\hat{q}(\omega), \delta_\omega)_{\omega\in\mathbb{Z}^d}$ while the one of L are $(\|\omega\|_2^2\, \hat{q}(\omega), \delta_\omega)_{\omega\in\mathbb{Z}^d}$.

Eigenvalue decay comparison. Hence, having $L \preceq c\Sigma^a$ is equivalent to having $\|\omega\|_2^2\, \hat{q}(\omega) \leq c\hat{q}(\omega)^a$. Now suppose that the decay of $\hat{q}$ is governed by

$$c_1(1+\sigma^{-1}\|\omega\|_2^2)^{-m} \leq \hat{q}(\omega) \leq c_2(1+\sigma^{-1}\|\omega\|_2^2)^{-m},$$

for two constants $c_1, c_2 > 0$. In particular, this is verified for Matérn kernels, corresponding to the fractional Sobolev space H^m, and for the Laplace kernel with $m = (d+1)/2$, which reads $k(x,y) = \exp(-\sigma^{-1}\|x-y\|)$. The Gaussian kernel could be seen as $m = +\infty$ as it has exponential decay. With such a decay we have, assuming without restrictions that we are in one dimension

$$\omega^2\hat{q}(\omega) \leq c_2\omega^2(1+\sigma^{-1}\omega^2)^{-m} \leq c_2\sigma(1+\sigma^{-1}\omega^2)^{-(m-1)} \leq c_1^{\frac{m}{m-1}} c_2\sigma\hat{q}(\omega)^{\frac{m-1}{m}}.$$

In other terms, we can consider $c = c_1^{\frac{m}{m-1}} c_2\sigma$ and $a = (m-1)/m$. Assuming that q is square-integrable, so is $\hat{q}$, which implies that $2m > d$. As a consequence, we do have $a > 1 - 2/d$. Note that this reasoning could be extended to the case where $\rho_\mathcal{X}$ has a density against the Lebesgue measure, that is bounded above and below away from zero.

9.D Spectral decomposition

In this section, we recall facts on spectral regularization, before proving Proposition 64 and extending it to the case $\mu = 0$.

9.D.1 Generalized singular value with matrices

Generalized singular value decomposition. Let $A \in \mathbb{R}^{m1\times n}$ and $B \in \mathbb{R}^{m_2\times n}$ be two matrices. There exists $U \in \mathbb{R}^{m_1\times m_1}$, $V \in \mathbb{R}^{m_2\times m_2}$ two orthogonal matrices, $c \in \mathbb{R}^{m_1\times r}$ and $s \in \mathbb{R}^{m_2\times r}$ two 1-diagonal matrices such that $c^\top c + s^\top s = I_r$, and $H \in \mathbb{R}^{n\times r}$ a non-singular matrix such that

$$A = UcH^{-1}, \qquad B = VsH^{-1}.$$

To be more precise c is such that only entries $c_{ii} = \cos(\theta_i)$ for $i < \min(r, m_2)$ are non-zeros and s such that only entries $s_{m_1-i,r-i} = \sin(\theta_{r-i})$ for $i < \min(r, m_1)$ are non-zeros, with $\theta_i \in [-\pi/2, \pi/2]$ some angle. Here, c stands for cosine, s for sinus and r for rank.

Link with generalized eigenvalue problem. As well as the singular value of A is linked with the eigenvalue of $A^\top A$, the generalized singular value decomposition of $[A; B]$ is linked with the generalized eigenvalue problem linked with $(A^\top A, B^\top B)$. Indeed, we have

$$A^\top A = H^{-\top} c^\top c H^{-1}, \qquad B^\top B = H^{-\top} s^\top s H^{-1}.$$

In particular, with (e_i) the canonical basis of $\mathbb{R}^r$, and h_i the i-th column of H, we get

$$H^\top A^\top A h_i = \cos(\theta_i)^2 e_i = \tan(\theta_i)^{-2} \sin(\theta_i)^2 e_i = \tan(\theta_i)^{-2} H^\top B^\top B h_i.$$

From which we deduce that, since $\operatorname{im} A \cup \operatorname{im} B \subset \operatorname{im} H^\top$,

$$A^\top A h_i = \tan(\theta_i)^{-2} B^\top B h_i, \qquad h_j^\top B^\top B h_i = \sin(\theta_i)^2 \mathbf{1}_{i=j}.$$

So if we denote by $f_i = |\sin(\theta_i)|^{-1} h_i$ and $\lambda_i = \tan(\theta_i)^{-2}$, assuming $\lambda_i \neq 0$ for all $i \leq r$ (which corresponds to $\ker B \subset \ker A$), $(\lambda_i)_{i \leq r}$, $(f_i)_{i \leq r}$ provide the generalized eigenvalue decomposition of $(A^\top A, B^\top B)$ in the sense that

$$A^\top A f_i = \lambda_i B^\top B f_i, \qquad f_j^\top B^\top B f_i = \mathbf{1}_{i=j}, \qquad f_j^\top A^\top A f_i = \lambda_i \mathbf{1}_{i=j}.$$

9.D.2 Tikhonov regularization

Define the Tikhonov regularization

$$x_\lambda = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \|Ax - b\|^2 + \lambda \|Bx\|^2 .$$

When this problem is well-defined, the solution is defined as

$$x_\lambda = (A^\top A + \lambda B^\top B)^\dagger A^\top b.$$

With the generalized singular value decomposition of A and B, we have

$$A^\top A + \lambda B^\top B = H^{-\top} \gamma_\lambda H^{-1}, \quad \text{with} \quad \gamma_\lambda = c^\top c + \lambda s^\top s.$$

Using the fact that $A^\top b = H^{-\top} c^\top U^\top b$, we get

$$x_\lambda = H \gamma_\lambda^{-1} c^\top U^\top b = \left(\sum_{i=1}^r \frac{\cos(\theta_i)}{\cos(\theta_i)^2 + \lambda \sin(\theta_i)^2} h_i \otimes u_i \right) b.$$

Now, we would like to replace c_{ii}, s_{ii}, h_i and u_i with quantities that depend on λ_i, f_i and A. To do so recall that $AH = Uc$, therefore $\cos(\theta_i) u_i = A h_i$, and recall that $h_i = \sin(\theta_i) f_i$ and $\lambda_i = \cos(\theta_i)^2 / \sin(\theta_i)^2$. Inputting this equality in the last expression of x_λ we get

$$x_\lambda = \left(\sum_{i=1}^r \frac{\sin(\theta_i)^2}{\cos(\theta_i)^2 + \lambda \sin(\theta_i)^2} f_i \otimes A f_i \right) b = \left(\sum_{i=1}^r \frac{1}{\lambda_i + \lambda} f_i \otimes A f_i \right) b.$$

Finally,

$$b_\lambda = A x_\lambda = \sum_{i=1}^r \psi(\lambda_i) \langle A f_i, b \rangle A f_i, \qquad \text{where} \qquad \psi(x) = \frac{1}{x + \lambda}.$$

9.D.3 Extension to operators

To end the proof of Proposition 64, we should prove that we can apply the generalized eigenvalue decomposition to operators. We will only prove that it is possible for $(\Sigma, L + \mu)$ based on simple considerations.

Proposition 69. *When k is continuous and* $\operatorname{supp} \rho_{\mathcal{X}}$ *is bounded, Σ is a compact operator.*

Proof. We have $\Sigma = \mathbb{E}[k_X \otimes k_X]$ and $\|k_x\| = k(x, x)$. Since k is continuous and $\operatorname{supp} \rho_{\mathcal{X}}$ is compact, for $x \in \operatorname{supp} \rho_{\mathcal{X}}$, $k(x, x)$ is bounded. Hence, Σ is a trace class and compact operator. □

Proposition 70. *When k is twice differentiable with continuous derivative, and* $\operatorname{supp}\rho_{\mathcal{X}}$ *is compact, L is a compact operator. As a consequence, L has a compact spectrum, and has a pseudo-inverse that we will denote, with a slight abuse of notation, by L^{-1}.*

Proof. The proof is similar to the one showing that Σ is compact, based on the fact that $L = \sum_{i=1}^{d} \mathbb{E}[\partial_i k_X \otimes \partial_i k_X]$, and $\|\partial_i k_X\|^2 = \partial_{1i}\partial_{2i} k(x,x)$. □

Proposition 71. *When Σ is compact, for all $\mu > 0$, $(L+\mu)^{-1/2}\Sigma(L+\mu)^{-1/2}$ is a compact operator.*

Proof. The proof is straightforward

$$\operatorname{Tr}((L+\mu)^{-1/2}\Sigma(L+\mu)^{-1/2}) = \operatorname{Tr}(\Sigma(L+\mu)^{-1}) \leq \left\|(L+\mu)^{-1}\right\|_{\text{op}} \operatorname{Tr}(\Sigma) \leq \mu^{-1}\operatorname{Tr}(\Sigma) < +\infty.$$

Therefore, the operator is trace class, hence compact. □

Proposition 72. *For any $\mu > 0$, the generalized eigenvalue decomposition of $(\Sigma, L+\mu)$ as defined in Proposition 64 exists.*

Proof. Using the spectral theorem, since $(L+\mu)^{-1/2}\Sigma(L+\mu)^{-1/2}$ is positive self-adjoint compact operator, there exists (ξ_i) a basis of $\mathcal{H}$ and $(\lambda_i) \in \mathbb{R}_+$ a decreasing sequence (note that $\ker(L+\mu) = \ker\Sigma = \{0\}$), such that

$$(L+\mu)^{-1/2}\Sigma(L+\mu)^{-1/2} = \sum_{i\in\mathbb{N}} \lambda_i \xi_i \otimes \xi_i.$$

Taking $\theta_i = (L+\mu)^{-1/2}\xi_i$, we get $\Sigma\theta_i = \lambda_i L\theta_i$. Because (ξ_i) generates $\mathcal{H}$, and $(L+\mu)^{-1/2}$ is bijective (since L is compact, $(L+\mu)^{-1}$ is coercive), $((L+\mu)^{-1/2}\xi_i)$ generates $\mathcal{H}$. □

Proposition 64 follows from prior discussion on Tikhonov regularization extended to infinite summations.

9.D.4 The case $\mu = 0$

When $\mu = 0$, (9.7) should be seen as the rewriting of (9.18) based on the RKHS $\mathcal{G} = \operatorname{im} S$. This can only be done when the eigenvectors of $\mathcal{L}$ appearing in (9.17) belongs to $\mathcal{G} = \operatorname{im} S = \operatorname{im} K^{1/2}$, which is exactly what Assumption 18 provides. In such a setting, we can find $(\theta_i) \in \mathcal{H}^{\mathbb{N}}$ to write $\lambda_i^{1/2} e_i = S\theta_i$ as soon as $\lambda_i \neq 0$ (write $Me_i = S\theta_i$ for M an abstraction representing $+\infty$ when $\lambda_i = 0$, handling the potential fact that $\ker B \not\subset \ker A$), we get $\theta_i S^* S\theta_j = \lambda_i \mathbf{1}_{i=j}$, and $L\theta_i = \lambda_i^{-1}\Sigma\theta_i$, and we can extend Proposition 64 to the case $\mu = 0$, with

$$g_\lambda = \sum_{i\in\mathbb{N}} \psi(\lambda_i) \left\langle S^\star g_\rho, \theta_i \right\rangle S\theta_i, \tag{9.19}$$

where we handle the null space of $\mathcal{L}$ with the equality $M\psi(M) = 1$, verified by M our abstraction representing $+\infty$, so that $\psi(M)\left\langle S^\star g_\rho, \theta_i\right\rangle S\theta_i = \left\langle g_\rho, e_i\right\rangle e_i$ as soon as $\lambda_i = 0$.

Beyond Assumption 18. Assumption 18 could be made generic by considering the biggest $(p_i) \in \mathbb{R}_+^{\mathbb{N}}$ such that $K^{-p_i} e_i$ belongs to L^2, and rewriting (9.19) under the form $g_\lambda = \sum_{i\in\mathbb{N}} \psi(\lambda_i)\left\langle (S_0 K^{p_i})^\star g_\rho, \theta_i\right\rangle S_0 K^{p_i}\theta_i$, with $\theta_i = \lambda_i^{-1/2} S_0^{-1} K^{-p_i}\theta_i$ and $S_0 = K^{-1/2}S$ the isomorphism between $\mathcal{H}$ and L^2 (assuming that S is dense in L^2). Such an assumption would describe all situations from no assumption ($p_i = 0$ for all i), Assumption 18 ($p_i = 1/2$ for all i) to even more optimistic assumptions ($p_i \geq 1$ for all i).

9.E Consistency analysis

This section is devoted to the proof of Theorem 16. The proof is based on (9.7) and (9.19), and splits the error of $\left\|g_\rho - \hat{g}_p\right\|_{L^2}^2$ into several components linked with how well we approximate $S^\star g_\rho$, and how well we approximate the eigenvalue decomposition (λ_i, θ_i) of (Σ, L).

9.E.1 Sketch and understanding of the proof

In this subsection, we explain how the proofs work for consistency theorems such as Theorem 16.

Let us define the mapping $G : \mathcal{H} \times \mathcal{C} \to L^2$ with $\mathcal{C}$ the set of pairs of self-adjoint operators on $\mathcal{H}$ that admit a generalized eigenvalue decomposition, as

$$G(\theta, (A, B)) = \sum_{i \in \mathbb{N}} \psi(\lambda_i) \langle \theta, \theta_i \rangle S\theta_i \qquad \text{with} \qquad (\lambda_i, \theta_i) \in \text{GEVD}(A, B). \tag{9.20}$$

$G(\theta, (A, B)) \in L^2$ corresponds to writing $\theta \in \mathcal{H}$ in the basis associated with the generalized eigenvalue decomposition (GEVD) of (A, B).

Proposition 73. *Under Assumptions 17 and 18, and with ψ defined in Theorem 16*

$$g_\lambda = G(S^\star g_\rho, (\Sigma, L)), \qquad \textit{and} \qquad \hat{g}_p = G(\widehat{S^\star g_\rho}, (P\hat{\Sigma}P, P\hat{L}P + \mu P)),$$

with P the projection matrix from $\mathcal{H}$ to Span $\left\{k_{X_i}\right\}_{i \leq p}$.

Proof. This is a direct application of Assumptions 17, 18, (9.19) and Algorithm 1. □

The main point of the proof is to relate g_ρ to $\hat{g}_p$. To do so, we will use several functions in L^2 generated by G. We detail our steps in Table 9.2. Table 9.2 gives a first answer to the two questions asked in the opening of Section 9.5. The number of unlabeled data controls the convergence of the operators $(P\hat{\Sigma}P, P\hat{L}P + \mu P)$ toward $(\Sigma, L + \mu)$. The number of labeled data controls the convergence of the vector $\widehat{S^\star g_\rho}$ toward $S^\star g_\rho$. Priors on the structure of the problem, such as source and approximation conditions, control the convergence of the bias estimate $g_{\lambda,\mu}$ toward g_ρ. Furthermore, a more precise study reveals that the concentration of operators are related to efficient dimension (Caponnetto and De Vito, 2006) and are accelerated by capacity assumptions on the functional space whose norm is $\|g\| = \left\|(\mathcal{L} + K^{-1})^{1/2} g\right\|_{L^2}$, and that the concentration of the vector $\widehat{S^\star g_\rho}$ is accelerated by assumptions on moments of the variable $Y(I + \lambda\mathcal{L})^{-1}\delta_X$ (inheriting randomness from $(X, Y) \sim \rho$).

Table 9.2: Steps in the consistency analysis

Estimate	Vector	Property of convergence	Basis
$\hat{g}_p$	$\widehat{S^\star g_\rho}$		$(P\hat{\Sigma}P, P\hat{L}P + \mu P)$
		Low-rank approx. (Rudi et al., 2015)	$\downarrow$
$\hat{g}$	$\widehat{S^\star g_\rho}$		$(\hat{\Sigma}, \hat{L} + \mu)$
		Operator concentration (Minsker, 2017)	$\downarrow$
g_{n_ℓ}	$\widehat{S^\star g_\rho}$		$(\Sigma, L + \mu)$
		Vector concentration (Yurinskii, 1970)	$\downarrow$
$g_{\lambda,\mu}$	$S^\star g_\rho$		$(\Sigma, L + \mu)$
		Source condition (Lin et al., 2020)	$\downarrow$
g_λ	$S^\star g_\rho$		(Σ, L)
$\downarrow$		Source condition (Caponnetto and De Vito, 2006)	
g_ρ			

Control of biases. We begin our study in a downward fashion regarding Table 9.2. Indeed, for Tikhonov regularization (9.3), we can show that for $q \in [0,1]$,

$$\left\|g_\lambda - g_\rho\right\|_{L^2} \leq \lambda^q \left\|\mathcal{L}^q g_\rho\right\|_{L^2}.$$

Meaning that if we have the source condition $g_\rho \in \operatorname{im} \mathcal{L}^q$ (which is a condition on how fast $(\langle g_\rho, e_i\rangle)_{i\in\mathbb{N}}$ decreases compared to $(\lambda_i)_{i\in\mathbb{N}}$ for (λ_i, e_i) the eigenvalue decomposition of $\mathcal{L}^{-1}$), the rates of convergence of this term when n goes to infinity is controlled by the regularization scheme λ_n^q.

Similarly to the kernel-free bias above, for $(q_i) \in (0,1)^{\mathbb{N}}$, we can have

$$\left\|g_{\lambda,\mu} - g_\lambda\right\|_{L^2}^2 \leq 2\sum_{i\in\mathbb{N}} \lambda^{2q_i}\mu^{2q_i}\left(\frac{\lambda_i}{\lambda+\lambda_i}\right)^2 \left|\langle e_i, g_\rho\rangle\right|^2 \left\|K^{-q_i}e_i\right\|_{L^2}^2.$$

This shows explicitly the usefulness of controlling at the same time how g_ρ is supported on the eigenspaces of $\mathcal{L}$ and how the eigenvectors are well approximated by the RKHS $\mathcal{H}$, which can be read in the value of (q_i) such that all $e_i \in \operatorname{im} K^{q_i}$.

Vector concentration. Let us now switch to concentration of $\widehat{S^\star g_\rho} = n_\ell^{-1}\sum_{i=1}^{n_\ell} Y_i k_{X_i}$ toward $S^\star g_\rho = \mathbb{E}_{(X,Y)\sim\rho}[Yk_X]$, it will allow controlling $\left\|g_{n_\ell} - g_{\lambda,\mu}\right\|_{L^2}^2$ with the notations appearing in Table 9.2. Note how this convergence should be measured in terms of the reconstruction error

$$\left\|\sum_{i\in\mathbb{N}} \psi(\lambda_{i,\mu})\left\langle S^\star g_\rho - \widehat{S^\star g_\rho}, \theta_{i,\mu}\right\rangle S\theta_{i,\mu}\right\|_{L^2}.$$

This error might behave in a much better fashion than the L^2 error between $SS^\star g_\rho$ and $S\widehat{S^\star g_\rho}$. In particular, on Figure 9.1, we can consider $\psi(\lambda_{i,\mu}) = 0$ for $i > 4$, and we might have $\langle Yk_X, \theta_{i,\mu}\rangle = Y\mathbf{1}_{x\in C_i}$, for $i \leq 4$ and $(X,Y) \in \mathcal{X}\times\mathcal{Y}$, where C_i is the i-th innermost circle. In this setting, when all four $(Y \mid X \in C_i)$ are deterministic, we only need one labeled point per circle to clear the reconstruction error. Based on concentration results in Hilbert space, when $|Y|$ is bounded by a constant $c_{\mathcal{Y}}$, and $x \to k(x,x)$ by a constant κ^2, we have, with $\mathcal{D}_{n_\ell} \sim \rho^{\otimes n_\ell}$ the dataset generating the labeled data

$$\mathbb{E}_{\mathcal{D}_{n_\ell}}\left[\left\|g_{n_\ell} - g_{\lambda,\mu}\right\|_{L^2}^2\right] \leq 2\sigma_\ell^2(\mu\lambda n_\ell)^{-1} + \frac{4}{9}c_{\mathcal{Y}}^2\kappa^2(\mu\lambda n_\ell^2)^{-1}.$$

where $\sigma_\ell^2 \leq c_{\mathcal{Y}}^2 \operatorname{Tr}(\Sigma)$ is a variance parameter to relate to the variance of $Y(I+\lambda\mathcal{L})^{-1}\delta_X$ (where the randomness is inherited from $(X,Y) \sim \rho$). The fact that the need for labeled data depends on the variance of (X,Y) after being diffused through $\mathcal{L}$ is coherent with the results obtained by Lelarge and Miolane (2019) in the specific case of a mixture of two Gaussian.

Basis concentration. We are left with the comparison of g_{n_ℓ}, which is the filtering of $\widehat{S^\star g_\rho}$ with the operators $(\Sigma, L+\mu)$, and $\hat{g}_p$, which is the filtering of the same vector with the operators $(P\hat{\Sigma}P, P\hat{L}P + \mu P)$. As the number of samples grows toward infinity, we know that $(P\hat{\Sigma}P, P\hat{L}P + \mu P)$ will converge in operator norm toward $(\Sigma, L+\mu)$. Yet, how to leverage this property to quantify the convergence of $\hat{g}_p$ toward g_{n_ℓ}? Let us write $(\lambda_i, \theta_i) = \operatorname{GEVD}(\Sigma, L+\mu)$, and $(\lambda_i', \theta_i') = \operatorname{GEVD}(P\hat{\Sigma}P, P\hat{L}P+\mu P)$, we have

$$\left\|\hat{g}_p - g_{n_\ell}\right\|_{L^2} = \left\|\sum_{i\in\mathbb{N}} \psi(\lambda_i)\left\langle\theta_i, \hat{\theta}_\rho\right\rangle S\theta_i - \psi(\lambda_i')\left\langle\theta_i', \hat{\theta}_\rho\right\rangle S\theta_i'\right\| \qquad \text{with} \qquad \hat{\theta}_\rho = \widehat{S^\star g_\rho}.$$

The generic study of this quantity requires controlling eigenspaces one by one. Note that we expect convergence of eigenspaces to depend on gaps between eigenvalues. However, when considering Tikhonov regularization, this quantity can be written under a simpler form. In particular, the concentration of operators is controlled, up to few leftovers, through the quantity $\left\|(\Sigma+\lambda L+\mu\lambda)^{-1/2}((\Sigma-\hat{\Sigma}) + \lambda(L-\hat{L}))(\Sigma+\lambda L+\mu\lambda)^{-1/2}\right\|_{\text{op}}$ where $\|\cdot\|_{\text{op}}$ designs the operator norm. In this setting, the low-rank approximation is controlled through $\left\|(\Sigma+\lambda L)^{1/2}(I-P)\right\|_{\text{op}}$, and when $L \leq c\Sigma^\alpha$, this term can be controlled by $\left\|\Sigma^{1/2}(I-P)\right\|_{\text{op}} + \lambda^{1/2}\left\|\Sigma^{1/2}(I-P)\right\|_{\text{op}}^\alpha$ which can be controlled based on the work of Rudi et al. (2015).

9.E.2 Risk decomposition

In this subsection, we decompose the risk appearing in Theorem 16.

Control of biases

We begin by splitting the error $\left\|g_\rho - \hat{g}_p\right\|_{L^2}$ between a bias term due to the regularization parameters and a variance term due to the data. With the notation of Table 9.2,

$$\left\|g_\rho - \hat{g}_p\right\|_{L^2} \leq \left\|g_\rho - g_\lambda\right\|_{L^2} + \left\|g_\lambda - g_{\lambda,\mu}\right\|_{L^2} + \left\|g_{\lambda,\mu} - \hat{g}_p\right\|_{L^2}. \tag{9.21}$$

We will control the first two terms here, and the last term in the following subsections.

Proposition 74 (Bias in λ)**.** *Under Assumption 17*

$$\left\|g_\lambda - g_\rho\right\|_{L^2} \leq \lambda \left\|\mathcal{L} g_\rho\right\|_{L^2}. \tag{9.22}$$

Proof. Based on the definition of $g_\lambda = (I + \lambda\mathcal{L})^{-1} g_\rho$, we have

$$g_\lambda - g_\rho = ((I + \lambda\mathcal{L})^{-1} - I) g_\rho = -\lambda\mathcal{L} g_\rho.$$

Because g_ρ is supported on the first eigenvectors of the Laplacian (Assumption 17), we have $g_\rho = \sum_{i=1}^r \langle g_\rho, e_i \rangle e_i$, with e_i the eigenvector of $\mathcal{L}$ appearing in (9.17), and

$$\left\|\mathcal{L} g_\rho\right\|_{L^2}^2 = \left\| \sum_{i=1}^r \lambda_i^{-1} \langle g_\rho, e_i \rangle e_i \right\|_{L^2}^2 = \sum_{i=1}^r \lambda_i^{-2} \langle g_\rho, e_i \rangle^2 \leq \lambda_r^{-2} \left\|g_\rho\right\|_{L^2}^2 < +\infty.$$

This ends the proof of this proposition. □

Proposition 75 (Bias in μ)**.** *Under Assumptions 17 and 18, we have*

$$\left\|g_{\lambda,\mu} - g_\lambda\right\|_{L^2}^2 \leq \lambda\mu c_a^2 \left\|g_\rho\right\|_{L^2}^2, \qquad \textit{with} \qquad c_a^2 = \sum_{i=1}^r \left\|K^{-1/2} e_i\right\|_{L^2}^2 = \sum_{i=1}^r \|e_i\|_{\mathcal{H}}^2. \tag{9.23}$$

Proof. Before diving into the proof, recall that the RKHS norm penalization can be written as $\|g\|_{\mathcal{H}} = \left\|K^{-1/2} g\right\|_{L^2}$. Using the fact that $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, we have

$$\begin{aligned} g_{\lambda,\mu} - g_\lambda &= ((I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1} - (I + \lambda\mathcal{L})^{-1}) g_\rho \\ &= -(I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1} \lambda\mu K^{-1} (I + \lambda\mathcal{L})^{-1} g_\rho \\ &= -(\lambda\mu)^{1/2} (I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1/2} (I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1/2} \\ &\qquad \cdots \times (\lambda\mu K^{-1})^{1/2} K^{-1/2} (I + \lambda\mathcal{L})^{-1} g_\rho. \end{aligned}$$

As a consequence,

$$\left\|g_{\lambda,\mu} - g_\lambda\right\|_{L^2}^2 \leq \lambda\mu \left\| K^{-1/2} (I + \lambda\mathcal{L})^{-1} g_\rho \right\|_{L^2}^2,$$

where we used the fact that $I + \lambda\mathcal{L} + \mu\lambda K^{-1} \geq I$, so that $\left\|(I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1/2}\right\|_{\text{op}} \leq 1$ (with $\|\cdot\|_{\text{op}}$ the operator norm), and that

$$\begin{aligned} \left\| (I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1/2} (\lambda\mu K^{-1})^{1/2} \right\|_{\text{op}}^2 &= \lambda\mu \left\| K^{-1/2} (I + \lambda\mathcal{L} + \mu\lambda K^{-1})^{-1} K^{-1/2} \right\|_{\text{op}} \\ &= \lambda\mu \left\| (K + \lambda K^{1/2} \mathcal{L} K^{1/2} + \mu\lambda)^{-1} \right\|_{\text{op}} \leq 1. \end{aligned}$$

We continue the proof with

$$\begin{aligned} \left\| K^{-1/2} (I + \lambda\mathcal{L})^{-1} g_\rho \right\| &= \left\| \sum_{i=1}^r \frac{\lambda_i}{\lambda + \lambda_i} \langle g_\rho, e_i \rangle K^{-1/2} e_i \right\| \leq \sum_{i=1}^r \frac{\lambda_i}{\lambda + \lambda_i} \left| \langle g_\rho, e_i \rangle \right| \left\| K^{-1/2} e_i \right\| \\ &\leq \sum_{i=1}^r \left| \langle g_\rho, e_i \rangle \right| \left\| K^{-1/2} e_i \right\|_{L^2} \leq \left\|g_\rho\right\|_{L^2} \left(\sum_{i \leq r} \left\| K^{-1/2} e_i \right\|_{L^2}^2 \right)^{1/2}. \end{aligned}$$

Putting all the pieces together ends the proof. □

Vector concentration

We are left with the study of the variance $\left\|\hat{g}_p - g_{\lambda,\mu}\right\|$. To ease derivations, we denote $C = \Sigma + \lambda L$, $\hat{C} = \hat{\Sigma} + \lambda \hat{L}$, $\theta_\rho = S^\star g_\rho$, $\hat{\theta}_\rho = \widehat{S^\star g_\rho}$ and P the projection from $\mathcal{H}$ to Span $\left\{k_{X_i}\right\}_{i \leq p}$. We have, for Tikhonov regularization

$$\begin{aligned}\left\|\hat{g}_p - g_{\lambda,\mu}\right\|_{L^2} &= \left\|S\left(P(P\hat{C}P + \lambda\mu)^{-1}P\hat{\theta}_\rho - (C+\lambda\mu)^{-1}\theta_\rho\right)\right\|_{L^2} \\ &= \left\|\Sigma^{1/2}\left(P(P\hat{C}P + \lambda\mu)^{-1}P\hat{\theta}_\rho - (C+\lambda\mu)^{-1}\theta_\rho\right)\right\|_{\mathcal{H}}.\end{aligned}$$

We begin by isolating the dependency to labeled data

$$\begin{aligned}\left\|\hat{g}_p - g_{\lambda,\mu}\right\|_{L^2} \leq &\left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}} \\ &\cdots + \left\|\Sigma^{1/2}\left(P(P\hat{C}P + \lambda\mu)^{-1}\theta_\rho - (C+\lambda\mu)^{-1}\theta_\rho\right)\right\|_{\mathcal{H}}.\end{aligned} \tag{9.24}$$

We will control the first term here, and the second term in the following subsection.

Lemma 76 (Vector term)**.** *When* $\left\|(C+\lambda\mu)^{-1/2}(\hat{C} - C)(C+\lambda\mu)^{-1/2}\right\|_{\text{op}} \leq 1/2$*, we have*

$$\left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}} \leq 2\left\|(C+\lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}}. \tag{9.25}$$

Proof. We begin with the splitting

$$\begin{aligned}\left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}} \leq &\left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(C+\lambda\mu)^{1/2}\right\|_{\text{op}} \\ &\cdots \times \left\|(C+\lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}}.\end{aligned}$$

The first term will concentrate toward a matrix smaller than identity, while the second term concentrates toward zero. We can make those considerations more formal. Following basic properties with the Löwner order on operators, we have

$$\begin{aligned}& (C+\lambda\mu)^{-1/2}(C - \hat{C})(C+\lambda\mu)^{-1/2} \leq t \\ \Rightarrow\quad & \hat{C} \geq (1-t)C - t\lambda\mu \\ \Rightarrow\quad & P\hat{C}P \geq (1-t)PCP - t\lambda\mu P \geq (1-t)PCP - t\lambda\mu \\ \Rightarrow\quad & P\hat{C}P + \lambda\mu \geq (1-t)(PCP + \lambda\mu) \\ \Rightarrow\quad & (P\hat{C}P + \lambda\mu)^{-1} \leq (1-t)^{-1}(PCP + \lambda\mu)^{-1} \\ \Rightarrow\quad & (C+\lambda\mu)^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(C+\lambda\mu)^{1/2} \\ & \qquad \leq (1-t)^{-1}(C+\lambda\mu)^{1/2}P(PCP + \lambda\mu)^{-1}P(C+\lambda\mu)^{1/2} \leq (1-t)^{-1},\end{aligned}$$

where we have used the fact that the last operator is a projection. As a consequence, for any $t \in (0,1)$, we have

$$\begin{aligned}& \left\|(C+\lambda\mu)^{-1/2}(\hat{C} - C)(C+\lambda\mu)^{-1/2}\right\|_{\text{op}} \leq t \\ \Rightarrow\quad & \left\|(C+\lambda\mu)^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}(C+\lambda\mu)^{1/2}\right\|_{\text{op}} \leq (1-t)^{-1}. \\ \Rightarrow\quad & \left\|\Sigma^{1/2}P(P\hat{C}P + \lambda\mu)^{-1}P(C+\lambda\mu)^{1/2}\right\|_{\text{op}} \leq (1-t)^{-1}.\end{aligned}$$

Where the last implication follows from the fact that $C + \lambda\mu = \Sigma + \lambda L + \lambda\mu \geq \Sigma$. □

Basis concentration

We are left with the study of the basis concentration with the number of unlabeled data.

Lemma 77 (Basis term)**.** *When* $\left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1/2}\right\|_{\text{op}} \leq 1/2$*, we have*

$$\begin{aligned}&\left\|\Sigma^{1/2}(P(P\hat{C}P+\lambda\mu)^{-1}-(C+\lambda\mu)^{-1})\theta_\rho\right\|_{\mathcal{H}}\\&\qquad\leq 3\left\|C^{1/2}(I-P)\right\|_{\text{op}}\|g_\lambda\|_{\mathcal{H}}+2\left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}.\end{aligned} \tag{9.26}$$

Notice that Assumptions 17 and 18 imply $\|g_\lambda\|_{\mathcal{H}} \leq c_a \left\|g_\rho\right\|_{L^2} < +\infty$*.*

Proof. First of all, using that $A^{-1}-B^{-1} = A^{-1}(B-A)B^{-1}$, notice that

$$\begin{aligned}&\left\|\Sigma^{1/2}(P(P\hat{C}P+\lambda\mu)^{-1}-(C+\lambda\mu)^{-1})\theta_\rho\right\|_{\mathcal{H}}\\&=\left\|\Sigma^{1/2}P(P\hat{C}P+\lambda\mu)^{-1}P(C-\hat{C}P)(C+\lambda\mu)^{-1}\theta_\rho-\Sigma^{1/2}(I-P)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}\\&\leq\left\|\Sigma^{1/2}P(P\hat{C}P+\lambda\mu)^{-1}P(\hat{C}P-C)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}+\left\|\Sigma^{1/2}(I-P)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}\\&\leq\left\|\Sigma^{1/2}P(P\hat{C}P+\lambda\mu)^{-1}P(C+\lambda\mu)^{1/2}\right\|_{\text{op}}\left\|(C+\lambda\mu)^{-1/2}P(\hat{C}P-C)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}\\&\qquad\cdots+\left\|\Sigma^{1/2}(I-P)\right\|_{\text{op}}\left\|(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}.\end{aligned}$$

Because $\Sigma \preceq \Sigma+\lambda L = C$, we have $\left\|\Sigma^{1/2}(I-P)\right\|_{\text{op}} \leq \left\|C^{1/2}(I-P)\right\|_{\text{op}}$, and we also have

$$\left\|(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}} \leq \left\|C^{-1}\theta_\rho\right\|_{\mathcal{H}} = \left\|K^{-1/2}SC^{-1}\theta_\rho\right\|_{\mathcal{H}} = \left\|K^{-1/2}g_\lambda\right\|_{L^2} = \|g_\lambda\|_{\mathcal{H}}.$$

Recall, that, for any $t \in (0,1)$, we have already shown that

$$\begin{aligned}&\left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1/2}\right\|_{\text{op}} \leq t\\\Rightarrow\quad&\left\|\Sigma^{1/2}P(P\hat{C}P+\lambda\mu)^{-1}P(C+\lambda\mu)^{1/2}\right\|_{\text{op}} \leq (1-t)^{-1}.\end{aligned}$$

We are left with one last term to work on

$$\begin{aligned}\left\|(C+\lambda\mu)^{-1/2}P(\hat{C}P-C)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}} \leq&\left\|(C+\lambda\mu)^{-1/2}P(\hat{C}-C)P(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}\\&\cdots+\left\|(C+\lambda\mu)^{-1/2}C(I-P)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}.\end{aligned}$$

We control the first term with the fact for A, B, C three self-adjoint operators and x a vector we have

$$\|APBPCx\| = \|APBPCx\otimes xCPBPA\|_{\text{op}}^{1/2},$$

and that

$$\begin{aligned}&PCx\otimes xCP \preceq Cx\otimes xC\\\Rightarrow\quad&PBPCx\otimes xCPBP \preceq BPCx\otimes xCPB \preceq BCx\otimes xCB\\\Rightarrow\quad&APBPCx\otimes xCPBPA \preceq ABCx\otimes xCBA,\end{aligned}$$

so that

$$\left\|(C+\lambda\mu)^{-1/2}P(\hat{C}-C)P(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}} \leq \left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}.$$

We control the second term with

$$\begin{aligned}&\left\|(C+\lambda\mu)^{-1/2}C(I-P)(C+\lambda\mu)^{-1}\theta_\rho\right\|\\&\qquad\leq\left\|(C+\lambda\mu)^{-1/2}C^{1/2}\right\|\left\|C^{1/2}(I-P)\right\|\left\|(C+\lambda\mu)^{-1}\theta_\rho\right\|.\end{aligned}$$

Using that $(C+\lambda\mu)^{-1/2}C^{1/2} \preceq I$, we can add up everything to get the lemma.

For the part concerning $\|g_\lambda\|_{\mathcal{H}}$, notice that

$$\|g_\lambda\|_{\mathcal{H}} = \left\|K^{-1/2} g_\lambda\right\|_{L^2} = \left\|\sum_{i=1}^{r} \frac{\lambda_i}{\lambda_i + \lambda} \langle g_\rho, e_i \rangle K^{-1/2} e_i\right\|_{L^2} \leq \sum_{i=1}^{r} |g_\rho| \, e_i \left\|K^{-1/2} e_i\right\|$$
$$\leq \|g_\rho\|_{L^2} \left(\sum_{i=1}^{d} \left\|K^{-1/2} e_i\right\|_{L^2}^2\right)^{1/2} = c_a \|g_\rho\|_{L^2},$$

with c_a defined as before. □

Conclusion

Based on the last subsections, we have proved the following proposition.

Proposition 78 (Risk decomposition)**.** *When* $\left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1/2}\right\| \leq 1/2$*, Under the assumptions 17 and 18,*

$$\begin{aligned} \left\|\hat{g}_p - g_\rho\right\|_{L^2}^2 \leq 4\lambda^2 \left\|\mathcal{L} g_\rho\right\|_{L^2}^2 + 4\lambda\mu c_a^2 \|g_\rho\|_{L^2}^2 + 8 \left\|(C + \lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}}^2 \\ \cdots + 12 c_a^2 \left\|C^{1/2}(I - P)\right\|_{\text{op}}^2 \|g_\rho\|_{L^2}^2 + 8 \left\|(C + \lambda\mu)^{-1/2}(\hat{C} - C)(C + \lambda\mu)^{-1} \theta_\rho\right\|_{\mathcal{H}}^2. \end{aligned} \tag{9.27}$$

We are left with the quantification of the different convergences when the number of labeled and unlabeled data grows toward infinity. We will quantify those convergences based on concentration inequalities.

9.E.3 Probabilistic inequalities

In this subsection, we bound each term appearing in (9.27) based on concentration inequalities.

Vector concentration

The concentration of $\hat{\theta}_\rho = \widehat{S^\star g_\rho}$ toward $\theta_\rho = S^\star g_\rho$ is controlled through Bernstein inequality.

Theorem 17 (Concentration in Hilbert space (Yurinskii, 1970))**.** *Let denote by* $\mathcal{A}$ *a Hilbert space and by* (ξ_i) *a sequence of independent random vectors in* $\mathcal{A}$ *such that* $\mathbb{E}[\xi_i] = 0$*, that are bounded by a constant* M*, with finite variance* $\sigma^2 = \mathbb{E}[\sum_{i=1}^n \|\xi_i\|^2]$*. For any* $t > 0$*,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \xi_i\right\| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2 + 2tM/3}\right).$$

Proposition 79 (Vector concentration)**.** *When* $|Y|$ *is bounded by a constant* $c_{\mathcal{Y}}$*, and* $x \to k(x,x)$ *by a constant* κ^2*, we have, with* $\mathcal{D}_{n_\ell} \sim \rho^{\otimes n_\ell}$ *the dataset generating the labeled data*

$$\mathbb{P}_{\mathcal{D}_{n_\ell}}\left(\left\|(C + \lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho)\right\|_{\mathcal{H}} \geq t\right) \leq 2 \exp\left(-\frac{n_\ell t^2}{2\sigma_\ell^2 (\mu\lambda)^{-1} + 2t c_{\mathcal{Y}} (\mu\lambda)^{-1/2} \kappa / 3}\right), \tag{9.28}$$

where $\sigma_\ell^2 \leq c_{\mathcal{Y}}^2 \operatorname{Tr}(\Sigma)$ *is a variance parameter to relate with the variance of* $Y(I + \lambda\mathcal{L})^{-1}\delta_X$ *(where the randomness is inherited from* $(X,Y) \sim \rho$*).*

Proof. Recall that

$$(C + \lambda\mu)^{-1/2}(\hat{\theta}_\rho - \theta_\rho) = (\Sigma + \lambda L + \lambda\mu)^{-1/2}(n_\ell^{-1} \sum_{i=1}^{n_\ell} Y_i k_{X_i} - \mathbb{E}_\rho[Y k_X])$$

We want to apply Bernstein inequality to the vector $\xi_i = (\Sigma + \lambda L + \mu\lambda)^{-1/2} Y_i k_{X_i}$, after centering it. Let us denote by $c_{\mathcal{Y}}$ a bound on $|Y|$, $c_{\mathcal{Y}} \in \mathbb{R}$ exists since we have supposed ρ of compact support. We have

$$\begin{aligned}
\sigma^2 &= \mathbb{E}[\sum_{i=1}^{n_\ell} \|\xi_i - \mathbb{E}[\xi_i]\|^2] = n_\ell \, \mathbb{E}[\|\xi - \mathbb{E}[\xi_i]\|^2] \leq n_\ell \, \mathbb{E}[\|\xi\|^2] \\
&= n_\ell \, \mathbb{E}_{(X,Y)\sim\rho} \left[Y^2 \left\langle k_X, (\Sigma + \lambda L + \mu\lambda)^{-1} k_X \right\rangle \right] \\
&\leq n_\ell c_{\mathcal{Y}}^2 \, \mathbb{E}_{X\sim\rho_{\mathcal{X}}} \left[\left\langle k_X, (\Sigma + \lambda L + \mu\lambda)^{-1} k_X \right\rangle \right] \\
&= n_\ell c_{\mathcal{Y}}^2 \operatorname{Tr}\left((\Sigma + \lambda L + \mu\lambda)^{-1} \Sigma \right) \\
&\leq n_\ell c_{\mathcal{Y}}^2 \operatorname{Tr}(\Sigma) \left\| (\Sigma + \lambda L + \mu\lambda)^{-1} \right\|_{\mathrm{op}} \leq n_\ell c_{\mathcal{Y}}^2 \operatorname{Tr}(\Sigma) (\mu\lambda)^{-1}.
\end{aligned}$$

Note that we have proceed with a generic upper bound, but we expect this variance, which is related to the variance of $Y(I + \lambda\mathcal{L} + \lambda\mu K^{-1})^{-1}\delta_X$ to be potentially much smaller – if we remove the term in P the vector concentration is the concentration of the vector $S(S^*S + \lambda S^\star \mathcal{L} S + \lambda\mu)^{-1} Y k_X \simeq K^{1/2}(K + \lambda K^{1/2}\mathcal{L}K^{1/2} + \lambda\mu)^{-1} K^{1/2} S^{-\star} Y k_X = (I + \lambda L + \lambda\mu K^{-1})^{-1} Y S^{-\star} k_X \simeq (I + \lambda L + \lambda\mu K^{-1})^{-1} Y \delta_X$. To capture this fact, we will write $\sigma^2 \leq n_\ell \sigma_\ell^2 (\mu\lambda)^{-1}$, with $\sigma_\ell = c_{\mathcal{Y}} \operatorname{Tr}(\Sigma)^{1/2}$ in our analysis, but potentially much smaller under refined hypothesis and in practice. Similarly to the bound on the variance, we have

$$\|\xi - \mathbb{E}[\xi]\| \leq \|\xi\| = \left\| (\Sigma + \lambda L + \mu\lambda)^{-1/2} Y_i k_{X_i} \right\| \leq (\mu\lambda)^{-1/2} c_{\mathcal{Y}} \kappa,$$

with κ an upper bound on $k(x,x)^{1/2}$ for $x \in \operatorname{supp} \rho_{\mathcal{X}}$. As a consequence, applying Bernstein concentration inequality, we get, for any $t > 0$,

$$\mathbb{P}_{\mathcal{D}_{n_\ell}} \left(\left\| n_\ell^{-1} \sum_{i=1}^{n_\ell} \xi_i - E[\xi_i] \right\| \geq t \right) \leq 2 \exp\left(- \frac{n_\ell t^2}{2\sigma_\ell^2 (\mu\lambda)^{-1} + 2t c_{\mathcal{Y}} (\mu\lambda)^{-1/2} \kappa / 3} \right).$$

This ends the proof. □

Operator concentration

The convergence of $\hat{C}$ toward C is controlled with Bernstein inequality for self-adjoint operators.

Theorem 18 (Bernstein inequality for self-adjoint (Minsker, 2017))**.** *Let $\mathcal{A}$ be a separable Hilbert space, and (ξ_i) a sequence of independent random self-adjoint operators on $\mathcal{A}$. Assume that (ξ_i) are bounded by $M \in \mathbb{R}$, in the sense that almost everywhere $\|\xi\|_{\mathrm{op}} < M$, and have a finite variance $\sigma^2 = \left\| \sum_{i=1}^n \mathbb{E}[\xi_i^2] \right\|_{\mathrm{op}}$. For any $t > 0$,*

$$\mathbb{P}\left(\left\| \sum_{i=1}^n (\xi_i - \mathbb{E}[\xi_i]) \right\|_{\mathrm{op}} > t \right) \leq 2 \left(1 + 6 \, \frac{\sigma^2 + Mt/3}{t^2} \right) \frac{\operatorname{Tr}\left(\sum_{i=1}^n \mathbb{E}[\xi_i^2] \right)}{\left\| \sum_{i=1}^n \mathbb{E}[\xi_i^2] \right\|_{\mathrm{op}}} \exp\left(- \frac{t^2}{2\sigma^2 + 2tM/3} \right).$$

Proposition 80 (Operator concentration)**.** *When $x \to k(x,x)$ is bounded by κ^2, and $x \to \partial_{1,j}\partial_{2,j} k(x,x)$ is bounded by κ_j^2, we have*

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_n} \left(\left\| (C + \mu\lambda)^{-\frac{1}{2}} (C - \hat{C}) (C + \mu\lambda)^{-\frac{1}{2}} \right\|_{\mathrm{op}} > 1/2 \right) &\leq \left(2 + 56 \, \frac{\kappa^2 + \lambda \sum_{i=1}^d \kappa_j^2}{\lambda\mu n} \right) \\
\cdots \times (1 + \lambda\mu \, \|C\|_{\mathrm{op}}^{-1}) \frac{\kappa^2 + \lambda \sum_{j=1}^d \kappa_j^2}{\lambda\mu} &\exp\left(- \frac{\lambda\mu n}{10 \left(\kappa^2 + \lambda \sum_{j=1}^d \kappa_j^2 \right)} \right).
\end{aligned} \tag{9.29}$$

Proof. We want to apply the precedent concentration inequality to

$$\xi_i = (\Sigma + \lambda L + \lambda\mu)^{-1/2} (k_{X_i} \otimes k_{X_i} + \lambda \sum_{j=1}^d \partial_j k_{X_i} \otimes \partial_j k_{X_i}) (\Sigma + \lambda L + \lambda\mu)^{-1/2},$$

since we have, based on the fact that $C = \Sigma + \lambda L$ and that $\Sigma = \mathbb{E}[k_X \otimes k_X]$ and $L = \mathbb{E}[\sum_{j=1}^n \partial_j k_X \otimes \partial_j k_X]$,

$$\left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1/2}\right\|_{\text{op}} = n^{-1}\left\|\sum_{i=1}^n \xi_i - \mathbb{E}[\xi_i]\right\|_{\text{op}}.$$

We bound ξ with

$$\begin{aligned}
\|\xi\|_{\text{op}} &= \left\|(C+\mu\lambda)^{-\frac{1}{2}}\left(k_X\otimes k_X + \lambda\sum_{j=1}^d \partial_j k_X \otimes \partial_j k_X\right)(C+\mu\lambda)^{-\frac{1}{2}}\right\|_{\text{op}} \\
&\leq \text{Tr}\left((C+\mu\lambda)^{-\frac{1}{2}}\left(k_X\otimes k_X + \lambda\sum_{j=1}^d \partial_j k_X \otimes \partial_j k_X\right)(C+\mu\lambda)^{-\frac{1}{2}}\right) \\
&= \text{Tr}\left((C+\mu\lambda)^{-\frac{1}{2}} k_X \otimes k_X (C+\mu\lambda)^{-\frac{1}{2}}\right) \\
&\qquad \cdots + \lambda\sum_{j=1}^d \text{Tr}\left((C+\mu\lambda)^{-\frac{1}{2}}\partial_j k_X \otimes \partial_j k_X (C+\mu\lambda)^{-\frac{1}{2}}\right) \\
&= \left\|(C+\mu\lambda)^{-\frac{1}{2}}k_X\right\|_{\mathcal{H}}^2 + \lambda\sum_{j=1}^d \left\|(C+\mu\lambda)^{-\frac{1}{2}}\partial_j k_X\right\|_{\mathcal{H}}^2 \leq (\lambda\mu)^{-1}\left(\kappa^2 + \lambda\sum_{j=1}^d \kappa_j^2\right).
\end{aligned}$$

With κ^2 an upper bound on the kernel k and κ_j^2 an upper bound on $\partial_{1,j}\partial_{2,j}k$. For the variance we have, using Löwner order,

$$\begin{aligned}
\mathbb{E}[\xi^2] &\leq \sup_{X\in\mathcal{X}} \|\xi(X)\|_{\text{op}}\,\mathbb{E}[\xi] \leq (\lambda\mu)^{-1}\left(\kappa^2 + \lambda\sum_{j=1}^d \kappa_j^2\right)\mathbb{E}[\xi] \\
&= (\lambda\mu)^{-1}\left(\kappa^2 + \lambda\sum_{j=1}^d \kappa_j^2\right)(C+\lambda)^{-1}C \leq (\lambda\mu)^{-1}\left(\kappa^2 + \lambda\sum_{j=1}^d \kappa_j^2\right).
\end{aligned}$$

Therefore, we get for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_n}\left(\left\|(C+\mu\lambda)^{-\frac{1}{2}}(C-\hat{C})(C+\mu\lambda)^{-\frac{1}{2}}\right\|_{\text{op}} > t\right) \\
&\quad \leq 2\left(1 + 6\,\frac{(\kappa^2 + \lambda\sum_{i=1}^d \kappa_j^2)(1+t/3)}{\lambda\mu n t^2}\right)\frac{\|C\|_{\text{op}} + \lambda\mu}{\|C\|_{\text{op}}}\,\text{Tr}\left((C+\lambda)^{-1}C\right) \\
&\quad \cdots \times \exp\left(-\frac{nt^2}{2(\lambda\mu)^{-1}\left(\kappa^2 + \lambda\sum_{j=1}^d \kappa_j^2\right)(1+t/3)}\right).
\end{aligned}$$

Remark that

$$\text{Tr}\left((C+\mu\lambda)^{-1}C\right) \leq \left\|(C+\mu\lambda)^{-1}\right\|_{\text{op}}\text{Tr}(C) \leq (\lambda\mu)^{-1}(\kappa^2 + \lambda\sum_{j=1}^d \kappa_j^2).$$

Taking $t = 1/2$ ends the lemma. □

Basis concentration

Similarly, we could control $\left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}$ by using concentration of self-adjoint, yet this will lead to laxer bounds, than using concentration on vectors.

Proposition 81 (Basis concentration)**.** *When $x \to k(x,x)$ is bounded by κ^2, $x \to \partial_{1,j}\partial_{2,j}k(x,x)$ is bounded by κ_j^2, with Assumptions 17 and 18, we have*

$$\mathbb{P}_{\mathcal{D}_n}\left(\left\|(C+\mu\lambda)^{-\frac{1}{2}}(C-\hat{C})(C+\mu\lambda)^{-1}\theta_\rho\right\|_{\mathcal{H}} > t\right) \leq 2\exp\left(-\frac{\mu\lambda n t^2}{2c_1(c_1 + \lambda^{1/2}\mu^{1/2}t/3)}\right), \tag{9.30}$$

with $c_1 = (\kappa^2 + \lambda \sum_{i=1}^d \kappa_j^2) c_a \left\|g_\rho\right\|_{L^2}$.

Proof. We want to apply Bernstein concentration inequality to the vectors

$$\xi_i = (C + \mu\lambda)^{-1/2} \left(k_{X_i} \otimes k_{X_i} + \lambda \sum_{j=1}^d \partial_j k_{X_i} \otimes \partial_j k_{X_i} \right) (C + \lambda\mu)^{-1} \theta_\rho,$$

since

$$\left\| (C + \mu\lambda)^{-\frac{1}{2}} (C - \hat{C}) (C + \mu\lambda)^{-1} \theta_\rho \right\|_{\mathcal{H}} = n^{-1} \left\| \sum_{i=1}^n \xi_i - \mathbb{E}[\xi_i] \right\|_{\mathcal{H}}.$$

We bound ξ, reusing prior derivations, with

$$\begin{aligned}
\|\xi_i\|_{\mathcal{H}} &= \left\| (C + \mu\lambda)^{-1/2} \left(k_{X_i} \otimes k_{X_i} + \lambda \sum_{j=1}^d \partial_j k_{X_i} \otimes \partial_j k_{X_i} \right) (C + \lambda\mu)^{-1} \theta_\rho \right\|_{\mathcal{H}} \\
&\leq \left\| (C + \mu\lambda)^{-1/2} \right\|_{\mathrm{op}} \left\| \left(k_{X_i} \otimes k_{X_i} + \lambda \sum_{j=1}^d \partial_j k_{X_i} \otimes \partial_j k_{X_i} \right) \right\|_{\mathrm{op}} \left\| (C + \mu\lambda)^{-1} \theta_\rho \right\|_{\mathcal{H}}. \\
&\leq (\mu\lambda)^{-1/2} (\kappa^2 + \lambda \sum_{i=1}^d \kappa_j^2) c_a \left\| g_\rho \right\|_{L^2}.
\end{aligned}$$

For the variance, we have, similarly to prior derivations,

$$\begin{aligned}
\mathbb{E}[\|\xi\|^2] &\leq \sup_{X \in \mathcal{X}} \left\| k_X \otimes k_X + \lambda \sum_{j=1}^d \partial_j k_X \otimes \partial_j k_X \right\|_{\mathrm{op}} \left\| (C + \lambda\mu)^{-1} \theta_\rho \right\|^2 \\
&\qquad \cdots \times \mathbb{E} \left[\left\| (C + \mu\lambda)^{-1} \left(k_X \otimes k_X + \lambda \sum_{j=1}^d \partial_j k_{X_i} \otimes \partial_j k_X \right)_{\mathrm{op}} \right\| \right] \\
&\leq \left(\kappa^2 + \lambda \sum_{i=1}^d \kappa_i^2 \right) c_a^2 \left\| g_\rho \right\|_{L^2}^2 \\
&\qquad \cdots \mathbb{E} \left[\left\| (C + \mu\lambda)^{-1} k_X \otimes k_X \right\|_{\mathrm{op}} + \lambda \sum_{j=1}^d \left\| (C + \mu\lambda)^{-1} \partial_j k_{X_i} \otimes \partial_j k_X \right\|_{\mathrm{op}} \right] \\
&= \left(\kappa^2 + \lambda \sum_{i=1}^d \kappa_i^2 \right) c_a^2 \left\| g_\rho \right\|_{L^2}^2 \mathrm{Tr} \left((C + \mu\lambda)^{-1} C \right) \\
&\leq (\lambda\mu)^{-1} \left(\kappa^2 + \lambda \sum_{i=1}^d \kappa_i^2 \right)^2 c_a^2 \left\| g_\rho \right\|_{L^2}^2.
\end{aligned}$$

As a consequence, using Bernstein inequality,

$$\mathbb{P} \left(n^{-1} \left\| \sum_{i=1}^n \xi_i - \mathbb{E}[\xi_i] \right\| > t \right) \leq 2 \exp \left(- \frac{\mu\lambda n t^2}{2 c_1 (c_1 + \lambda^{1/2} \mu^{1/2} t / 3)}, \right)$$

with $c_1 = (\kappa^2 + \lambda \sum_{i=1}^d \kappa_j^2) c_a \left\| g_\rho \right\|_{L^2}$. Note that we have bounded naively the variable ξ and its variance, but we have shown how appears $\sup_{X \in \mathcal{X}} \left\| (C + \lambda\mu)^{-1} k_X \right\| + \lambda \sum_{i=1}^d \left\| (C + \lambda\mu)^{-1} \partial_i k_X \right\|$ and $\mathrm{Tr}((C + \lambda\mu)^{-1} C)$, which under interpolation and capacity assumptions could be controlled in a better fashion. □

Low-rank approximation

We now switch to Nyström approximation.

Proposition 82 (Low-rank approximation)**.** *When $x \to k(x,x)$ is bounded by κ^2, for any $p \in \mathbb{N}$ and $t > 0$, we have*

$$\mathbb{P}_{\mathcal{D}_p}\left(\left\|(I-P)\Sigma^{1/2}\right\|^2 > t\right) \leq \left(2 + \frac{116\kappa^2}{tp}\right)(2 + t\left\|\Sigma\right\|_{\text{op}}^{-1})\frac{\kappa^2}{t}\exp\left(-\frac{pt}{10\kappa^2}\right),$$

Proof. Reusing Proposition 3 of Rudi et al. (2015), for any $\gamma > 0$, we have, with P the projection on $\text{Span}\left\{k_{X_i}\right\}_{i\leq p}$ and $\hat{\Sigma} = p^{-1}\sum_{i=1}^{p} k_{X_i} \otimes k_{X_i}$,

$$\left\|(I-P)\Sigma^{1/2}\right\|^2 \leq \gamma\left\|(\hat{\Sigma}+\gamma)^{-1/2}\Sigma^{1/2}\right\|_{\text{op}}^2 \leq \gamma\left\|\Sigma^{1/2}(\hat{\Sigma}+\gamma)^{-1}\Sigma^{1/2}\right\|_{\text{op}}.$$

As a consequence, skipping derivations that can be retaken from our precedent proofs,

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_p}\left(\left\|(I-P)\Sigma^{1/2}\right\|^2 > t\right) &\leq \inf_{\gamma>0}\mathbb{P}_{\mathcal{D}_p}\left(\gamma\left\|\Sigma^{1/2}(\hat{\Sigma}+\gamma)^{-1}\Sigma^{1/2}\right\|_{\text{op}} > t\right) \\
&\leq \inf_{\gamma>0}\mathbb{P}_{\mathcal{D}_p}\left(\left\|(\Sigma+\gamma)^{-1/2}(\hat{\Sigma}-\Sigma)(\Sigma+\gamma)^{-1/2}\right\|_{\text{op}} > (1-\gamma t^{-1})\right) \\
&\leq \inf_{\gamma>0}\left(2 + 56\,\frac{\kappa^2}{\gamma p}\right)(1+\gamma\left\|\Sigma\right\|_{\text{op}}^{-1})frac\kappa^2\gamma\exp\left(-\frac{p\gamma u^2}{2\kappa^2(1+u/3)}\right),
\end{aligned}$$

with $u = (1-\gamma t^{-1})$. Taking $\gamma = t/2$, this term is simplified as

$$\mathbb{P}_{\mathcal{D}_p}\left(\left\|(I-P)\Sigma^{1/2}\right\|^2 > t\right) \leq \left(2 + 116\,\frac{\kappa^2}{tp}\right)(2 + t\left\|\Sigma\right\|_{\text{op}}^{-1})\frac{\kappa^2}{t}\exp\left(-\frac{pt}{10\kappa^2}\right),$$

which is the object of this proposition. □

Lemma 83. *When $L \leq c_d\Sigma^a$, we have*

$$\left\|(I-P)C^{1/2}\right\|_{\text{op}}^2 \leq \left\|(I-P)\Sigma^{1/2}\right\|_{\text{op}}^2 + c_d\lambda\left\|(I-P)\Sigma^{1/2}\right\|_{\text{op}}^{2a}. \tag{9.31}$$

Proof. This follows from the fact that

$$\begin{aligned}
\left\|C^{1/2}(I-P)\right\|_{\text{op}}^2 &= \left\|(I-P)C(I-P)\right\|_{\text{op}} = \left\|(I-P)(\Sigma+\lambda L)(I-P)\right\|_{\text{op}} \\
&\leq \left\|(I-P)\Sigma(I-P)\right\|_{\text{op}} + \lambda\left\|(I-P)L(I-P)\right\|_{\text{op}} \\
&\leq \left\|(I-P)\Sigma(I-P)\right\|_{\text{op}} + \lambda c_d\left\|(I-P)\Sigma^a(I-P)\right\|_{\text{op}} \\
&= \left\|(I-P)\Sigma^{1/2}\right\|_{\text{op}}^2 + \lambda c_d\left\|(I-P)\Sigma^{a/2}\right\|_{\text{op}}^2 \\
&= \left\|(I-P)\Sigma^{1/2}\right\|_{\text{op}}^2 + \lambda c_d\left\|(I-P)^a\Sigma^{a/2}\right\|_{\text{op}}^2 \\
&\leq \left\|(I-P)\Sigma^{1/2}\right\|_{\text{op}}^2 + \lambda c_d\left\|(I-P)\Sigma^{1/2}\right\|_{\text{op}}^{2a},
\end{aligned}$$

where we used the fact that $(I-P)^a = (I-P)$ and that $\|A^sB^s\| \leq \|AB\|^s$ for $s \in [0,1]$ and A, B positive self-adjoint. □

9.E.4 Averaged excess of risk - ending the proof

Based on the precedent excess of risk decomposition, and precedent concentration inequalities, we have all the elements to derive convergence rates of our algorithm. We will enunciate this convergence in terms of the averaged excess of risk of $\mathbb{E}_{\mathcal{D}_n}\left[\left\|\hat{g}_p - g_\rho\right\|_{L^2}^2\right]$.

Lemma 84. *Under Assumptions 17 and 18,*

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}\left[\left\|\hat{g}_p - g_\rho\right\|_{L^2}^2\right] &\leq 4c_{\mathcal{Y}}^2\,\mathbb{P}\left(\left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1/2}\right\| \leq 1/2\right)\\ &\cdots + 4\lambda^2\left\|\mathcal{L}g_\rho\right\|_{L^2}^2 + 4\lambda\mu c_a^2\left\|g_\rho\right\|_{L^2}^2\\ &\cdots + 8\,\mathbb{E}_{\mathcal{D}_n}\left[\left\|(C+\lambda\mu)^{-1/2}(\hat{\theta}_\rho-\theta_\rho)\right\|_{\mathcal{H}}^2\right] + 12c_a^2\left\|g_\rho\right\|_{L^2}^2\,\mathbb{E}_{\mathcal{D}_n}\left[\left\|C^{1/2}(I-P)\right\|_{\mathrm{op}}^2\right]\\ &\cdots + 8\,\mathbb{E}_{\mathcal{D}_n}\left[\left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}^2\right].\end{aligned} \tag{9.32}$$

Proof. We proceed using the fact that $\mathbb{E}[X] = \mathbb{E}[X\,|\,{}^cA]\,\mathbb{P}({}^cA) + \mathbb{E}[X\,|\,A]\,\mathbb{P}(A) \leq \sup X\,\mathbb{P}({}^cA) + \mathbb{E}[X\,|\,A]\,\mathbb{P}(A)$, with $A = \left\{\mathcal{D}_n \,\middle|\, \left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1/2}\right\| \leq 1/2\right\}$,

$$\mathbb{E}_{\mathcal{D}_n}\left[\left\|\hat{g}_p - g_\rho\right\|_{L^2}^2\right] \leq \sup_{\mathcal{D}_n}\left\|\hat{g}_p - g_\rho\right\|^2\mathbb{P}\left({}^cA\right) + \mathbb{E}_{\mathcal{D}_n}\left[\left\|\hat{g}_p - g_\rho\right\|^2\,\middle|\,A\right]\mathbb{P}(A).$$

When Y is bounded by $c_{\mathcal{Y}}$, because g_ρ is a convex combination of Y, we know that $\left\|g_\rho\right\|_{L^2} \leq c_{\mathcal{Y}}$, as a consequence, we can clip $\hat{g}_p$ to $[-c_{\mathcal{Y}}, c_{\mathcal{Y}}]$, which will only improve the estimation of g_ρ, as a consequence, we can consider the clipping estimate for which we have $\sup_{\mathcal{D}_n}\left\|\hat{g}_p - g_\rho\right\|_{L^2}^2 \leq 4c_{\mathcal{Y}}^2$. Regarding the second part, we have already decomposed the risk under the event $A = \left\{\mathcal{D}_n \,\middle|\, \left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1/2}\right\| \leq 1/2\right\}$. As a consequence, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}\left[\left\|\hat{g}_p - g_\rho\right\|_{L^2}^2\right] &\leq 4c_{\mathcal{Y}}^2\,\mathbb{P}({}^cA) + 4\lambda^2\left\|\mathcal{L}g_\rho\right\|_{L^2}^2\mathbb{P}(A) + 4\lambda\mu c_a^2\left\|g_\rho\right\|_{L^2}^2\mathbb{P}(A)\\ &\cdots + 8\,\mathbb{E}_{\mathcal{D}_n}\left[\left\|(C+\lambda\mu)^{-1/2}(\hat{\theta}_\rho-\theta_\rho)\right\|_{\mathcal{H}}^2\,\middle|\,A\right]\mathbb{P}(A)\\ &\cdots + 12c_a^2\left\|g_\rho\right\|_{L^2}^2\,\mathbb{E}_{\mathcal{D}_n}\left[\left\|C^{1/2}(I-P)\right\|_{\mathrm{op}}^2\,\middle|\,A\right]\mathbb{P}(A)\\ &\cdots + 8\,\mathbb{E}_{\mathcal{D}_n}\left[\left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}^2\,\middle|\,A\right]\mathbb{P}(A).\end{aligned}$$

To control the conditional expectation, we use that, when X is positive

$$\mathbb{E}\left[X\,|\,A\right]P(A) = \mathbb{E}[X] - \mathbb{E}\left[X\,|\,{}^cA\right]\mathbb{P}({}^cA) \leq \mathbb{E}[X].$$

This ends the proof. □

Based on deviation inequalities, we can control expectations based on the equality, for X positive, $\mathbb{E}[X] = \int_0^{+\infty}\mathbb{P}(X > t)\,\mathrm{d}t$.

Lemma 85. *In the setting of the paper,*

$$\mathbb{E}_{\mathcal{D}_n}\left[\left\|(C+\lambda\mu)^{-1/2}(\hat{\theta}_\rho-\theta_\rho)\right\|_{\mathcal{H}}^2\right] \leq 8\sigma_\ell^2(n_\ell\mu\lambda)^{-1} + 8c_{\mathcal{Y}}^2\kappa^2(n_\ell^2\mu\lambda)^{-1}. \tag{9.33}$$

Proof. First, recall that

$$\begin{aligned}\mathbb{P}\left(\left\|(C+\lambda\mu)^{-1/2}(\hat{\theta}_\rho-\theta_\rho)\right\|_{\mathcal{H}} > t\right) &\leq 2\exp\left(-\frac{n_\ell t^2}{2\sigma_\ell^2(\mu\lambda)^{-1} + 2tc_{\mathcal{Y}}(\lambda\mu)^{-1/2}\kappa/3}\right)\\ &\leq 2\exp\left(-\frac{n_\ell t^2}{2\max\left(2\sigma_\ell^2(\mu\lambda)^{-1}, 2tc_{\mathcal{Y}}(\lambda\mu)^{-1/2}\kappa/3\right)}\right)\\ &\leq 2\exp\left(-\frac{n_\ell\mu\lambda t^2}{4\sigma_\ell^2}\right) + 2\exp\left(-\frac{3n_\ell\mu^{1/2}\lambda^{1/2}t}{4c_{\mathcal{Y}}\kappa}\right).\end{aligned}$$

As a consequence

$$\begin{aligned}\mathbb{E}\left[\left\|(C+\lambda\mu)^{-1/2}(\hat{\theta}_\rho-\theta_\rho)\right\|_{\mathcal{H}}^2\right] &= \int_0^{+\infty}\mathbb{P}\left(\left\|(C+\lambda\mu)^{-1/2}(\hat{\theta}_\rho-\theta_\rho)\right\|_{\mathcal{H}}^2 > t\right)\mathrm{d}t\\ &\leq 2\int\exp\left(-\frac{n_\ell\mu\lambda t}{4\sigma_\ell^2}\right)\mathrm{d}t + 2\int\exp\left(-\frac{3n_\ell\mu^{1/2}\lambda^{1/2}t^{1/2}}{4c_{\mathcal{Y}}\kappa}\right)\mathrm{d}t.\\ &= 8\sigma_\ell^2(n_\ell\mu\lambda)^{-1} + \frac{64c_{\mathcal{Y}}^2\kappa^2}{9}(n_\ell^2\mu\lambda)^{-1}.\end{aligned}$$

This is the result stated in the lemma. □

Lemma 86. *In the setting of the paper,*

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}\left[\left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}^2\right] &\leq 8(\kappa^2+\lambda\partial\kappa^2)^2c_a^2\left\|g_\rho\right\|_{L^2}^2\\ &\quad\cdots\times\left((\mu\lambda n)^{-1}+(\mu\lambda n^2)^{-1}\right),\end{aligned} \tag{9.34}$$

with $\partial\kappa^2 = \sum_{i=1}^d \kappa_i^2$.

Proof. Let us denote by A the quantity $\left\|(C+\lambda\mu)^{-1/2}(\hat{C}-C)(C+\lambda\mu)^{-1}\theta_\rho\right\|_{\mathcal{H}}$, and $\partial\kappa^2 = \sum_{i=1}^d \kappa_j^2$. Recall that

$$\begin{aligned}\mathbb{P}(A>t) &\leq 2\exp\left(-\frac{\mu\lambda nt^2}{2c_1(c_1+\lambda^{1/2}\mu^{1/2}t/3)}\right)\\ &\leq 2\exp\left(-\frac{\mu\lambda nt^2}{4c_1^2}\right)+2\exp\left(-\frac{3(\mu\lambda)^{1/2}nt}{4c_1}\right).\end{aligned}$$

We conclude the proof similarly to the precedent lemma. □

Lemma 87. *Under Assumption 19,*

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}\left[\left\|C^{1/2}(I-P)\right\|_{\mathrm{op}}^2\right] &\leq \left(\frac{10\kappa^2\log(p)}{p}+\frac{10^a\kappa^{2a}c_d\lambda\log(p)^a}{p^a}\right)\\ &\quad\cdots\times\left(1+\frac{2\kappa^2}{\|\Sigma\|_{\mathrm{op}}\log(p)}\left(1+\frac{6}{\log(p)}\right)\left(\frac{1}{p}+\frac{1}{5\log(p)}\right)\right).\end{aligned} \tag{9.35}$$

Proof. Once again, this result comes from integration of the tail bound obtained on $\left\|C^{1/2}(I-P)\right\|_{\mathrm{op}}^2$ through the one we have on $\left\|\Sigma^{1/2}(I-P)\right\|_{\mathrm{op}}^2$ and the fact that $\left\|C^{1/2}(I-P)\right\|_{\mathrm{op}}^2 \leq \left\|\Sigma^{1/2}(I-P)\right\|_{\mathrm{op}}^2 + c_d\lambda\left\|\Sigma^{1/2}(I-P)\right\|_{\mathrm{op}}^{2a}$. For any $a,b>0$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}\left[\left\|\Sigma^{1/2}(I-P)\right\|_{\mathrm{op}}^2\right] &= \int_0^\infty\mathbb{P}_{\mathcal{D}_n}\left(\left\|\Sigma^{1/2}(I-P)\right\|_{\mathrm{op}}^2 > t\right)\mathrm{d}t\\ &\leq \int_0^\infty\min\left\{1,2\kappa^2\|\Sigma\|_{\mathrm{op}}^{-1}\left(1+\frac{58\kappa^2}{tp}\right)\left(1+\frac{2\kappa^2}{t}\right)\exp\left(-\frac{pt}{10\kappa^2}\right)\right\}\mathrm{d}t\\ &= \frac{10\kappa^2a}{p}\int_0^\infty\min\left\{1,2\kappa^2\|\Sigma\|_{\mathrm{op}}^{-1}\left(1+\frac{58}{10au}\right)\left(1+\frac{p}{5au}\right)\exp(-au)\right\}\mathrm{d}u\\ &\leq \frac{10\kappa^2a}{p}\left(b+\int_b^\infty 2\kappa^2\|\Sigma\|_{\mathrm{op}}^{-1}\left(1+\frac{6}{au}\right)\left(1+\frac{p}{5au}\right)\exp(-au)\,\mathrm{d}u\right)\\ &\leq \frac{10\kappa^2}{p}\left(ab+2\kappa^2\|\Sigma\|_{\mathrm{op}}^{-1}\left(1+\frac{6}{ab}\right)\left(1+\frac{p}{5ab}\right)\exp(-ab)\right).\end{aligned}$$

This last quantity is optimized for $ab = \log(p)$, which leads to the first part of the lemma. Similarly,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}\left[\left\|\Sigma^{1/2}(I-P)\right\|_{\text{op}}^{2a}\right] &= \int_0^\infty \mathbb{P}_{\mathcal{D}_n}\left(\left\|\Sigma^{1/2}(I-P)\right\|_{\text{op}}^{2a} > t\right)\mathrm{d}t \\
&= \int_0^\infty \mathbb{P}_{\mathcal{D}_n}\left(\left\|\Sigma^{1/2}(I-P)\right\|_{\text{op}}^{2} > t^{1/a}\right)\mathrm{d}t \\
&\leq \int_0^\infty \min\left\{1, 2\kappa^2\left\|\Sigma\right\|_{\text{op}}^{-1}\left(1+\frac{58\kappa^2}{t^{1/a}p}\right)\left(1+\frac{2\kappa^2}{t^{1/a}}\right)\exp\left(-\frac{pt^{1/a}}{10\kappa^2}\right)\right\}\mathrm{d}t \\
&= \frac{10^a\kappa^{2a}ac^a}{p^a}\int_0^\infty \min\left\{u^{a-1}, 2\kappa^2\left\|\Sigma\right\|_{\text{op}}^{-1}\left(1+\frac{58}{10cu}\right)\left(1+\frac{p}{5cu}\right)\frac{1}{u^{1-a}}\exp\left(-cu\right)\right\}\mathrm{d}u \\
&\leq \frac{10^a\kappa^{2a}ac^a}{p^a}\left(\frac{b^a}{a}+\int_b^\infty 2\kappa^2\left\|\Sigma\right\|_{\text{op}}^{-1}\left(1+\frac{6}{cu}\right)\left(1+\frac{p}{5cu}\right)\frac{1}{u^{1-a}}\exp\left(-cu\right)\mathrm{d}u\right) \\
&\leq \frac{10^a\kappa^{2a}}{p^a}\left((cb)^a + 2\kappa^2\left\|\Sigma\right\|_{\text{op}}^{-1}\left(1+\frac{6}{cb}\right)\left(1+\frac{p}{5cb}\right)\frac{1}{(cb)^{1-a}}\exp\left(-cb\right)\right).
\end{aligned}$$

Once again this is optimized for $cb = \log(p)$. □

Remark 88 (Leverage scores)**.** *Out of simplicity, we only present a low rank approximation with random subsampling. Yet, we can improve the result by considering subsampling based on leverage scores. If we consider the Gaussian kernel, $Sk_x \in L^2$ can be thought of as a function that is a little bump around $x \in \mathcal{X}$. In essence, subsampling based on leverage scores, consists in representing the solution on a subsampled sequence $(k_{X_i})_{i\in I}$ where the X_i are far from one another so that the bump functions (Sk_{X_i}) can approximate a maximum of functions. (Rudi et al., 2015) shows that with leverage scores, we can take $p = (\mu\lambda)^\gamma \log(n)$, with γ linked with the capacity of the RKHS linked with the kernel k.*

If we add all derivations, we have derived the following theorem.

Theorem 19. *Under Assumptions 17, 18 and 19,*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\left[\left\|\hat{g}-g_\rho\right\|_{L^2}^2\right] \\
&\quad\leq 8c_{\mathcal{Y}}^2\left(1+28\,\frac{\kappa^2+\lambda\partial\kappa^2}{\lambda\mu n}\right)\left(1+\lambda\mu\left\|C\right\|_{\text{op}}^{-1}\right)\frac{\kappa^2+\lambda\partial\kappa^2}{\lambda\mu}\exp\left(-\frac{\lambda\mu n}{10\left(\kappa^2+\lambda\partial\kappa^2\right)}\right) \\
&\quad\cdots + 4\lambda^2\left\|\mathcal{L}g_\rho\right\|_{L^2}^2 + 4\lambda\mu c_a^2\left\|g_\rho\right\|_{L^2}^2 + 64\sigma_\ell^2(n_\ell\mu\lambda)^{-1} + 57c_{\mathcal{Y}}^2\kappa^2(n_\ell^2\mu\lambda)^{-1} \\
&\quad\cdots + 64(\kappa^2+\lambda\partial\kappa^2)^2c_a^2\left\|g_\rho\right\|_{L^2}^2(\mu\lambda n)^{-1} + 57(\kappa^2+\lambda\partial\kappa^2)^2c_a^2\left\|g_\rho\right\|_{L^2}^2(\mu\lambda n^2)^{-1} \\
&\quad\cdots + 12c_a^2\left\|g_\rho\right\|_{L^2}^2\left(\frac{10\kappa^2\log(p)}{p}+\frac{10^a\kappa^{2a}c_d\lambda\log(p)^a}{p^a}\right) \\
&\qquad\cdots\times\left(1+\frac{2\kappa^2}{\left\|\Sigma\right\|_{\text{op}}\log(p)}\left(1+\frac{6}{\log(p)}\right)\left(\frac{1}{p}+\frac{1}{5\log(p)}\right)\right).
\end{aligned} \tag{9.36}$$

where $c_{\mathcal{Y}}$ is an upper bound on Y, κ^2 is an upper bound on $x \to k(x,x)$, $\partial\kappa^2 = \sum_{i=1}^d \kappa_i^2$ with κ_i^2 a bound on $x \to \partial_{1i}\partial_{2i}\partial k_{x_i}$, c_d and a the constants appearing in Assumption 19, c_a a constant such that $\|g\|_{\mathcal{H}} \leq c_a\|g\|_{L^2}$ and $\sigma_\ell^2 \leq c_{\mathcal{Y}}^2\kappa^2$ a variance parameter linked with the variance of $Y(I+\lambda\mathcal{L})^{-1}\delta_X$.

Theorem 16 is a corollary of this theorem.

Part IV

Active Labeling

Chapter 10

Streaming Stochastic Gradients

The following is a reproduction of Cabannes et al. (2022).

The workhorse of machine learning is stochastic gradient descent. To access stochastic gradients, it is common to consider iteratively input/output pairs of a training dataset. Interestingly, it appears that one does not need full supervision to access stochastic gradients, which is the main motivation of this paper. After formalizing the "active labeling" problem, which generalizes active learning based on partial supervision, we provide a streaming technique that provably minimizes the ratio of generalization error over the number of samples. We illustrate our technique in depth for robust regression.

10.1 Introduction

A large amount of the current hype around artificial intelligence was fueled by the recent successes of supervised learning. Supervised learning consists in designing an algorithm that maps inputs to outputs by learning from a set of input/output examples. When accessing many samples, and given enough computation power, this framework is able to tackle complex tasks. Interestingly, many of the difficulties arising in practice do not emerge from choosing the right statistical model to solve the supervised learning problem, but from the problem of collecting and cleaning enough data (see Chapters 1 and 2 of Géron, 2017, for example). Those difficulties are not disjoint from the current trends toward data privacy regulations (Council of European Union, 2016). This fact motivates this work, where we focus on how to efficiently collect information to carry out the learning process.

In this paper, we formalize the "active labeling" problem for weak supervision, where the goal is to learn a target function by acquiring the most informative dataset given a restricted budget for annotation. We focus explicitly on weak supervision that comes as a set of label candidates for each input, aiming to partially supervise input data in the most efficient way to guide a learning algorithm. We also restrict our study to the streaming variant where, for each input, only a single partial information can be collected about its corresponding output. The crux of this work is to leverage the fact that full supervision is not needed to acquire unbiased stochastic gradients, and perform stochastic gradient descent.

The following summarizes our contributions.

1. First, we introduce the "active labeling" problem, which is a relevant theoretical framework that encompasses many useful problems encountered by practitioners trying to annotate their data in the most efficient fashion, as well as its streaming variation, in order to deal with privacy preserving issues. This is the focus of Section 10.2.
2. Then, in Section 10.3, we give a high-level framework to access unbiased stochastic gradients with weak information only. This provides a simple solution to the streaming "active labeling" problem.
3. Finally, we detail this framework for a robust regression task in Section 10.4, and provide an algorithm whose optimality is proved in Section 10.5.

As a proof of concept, we provide numerical simulations in Section 10.6. We conclude with a high-level discussion around our methods in Section 10.7.

Related work. Active query of information is relevant to many settings. The most straightforward applications are searching games, such as Bar Kokhba or twenty questions (Walsorth, 1882). We refer to Pelc

(2002) for an in-depth survey of such games, especially when liars introduce uncertainty, and their relations with coding on noisy channels. But applications are much more diverse, *e.g.* for numerical simulation (Chevalier et al., 2014), database search (Qarabaqi and Riedewald, 2014), or shape recognition (Geman and Jedynak, 1993), to name a few.

In terms of motivations, many streams of research can be related to this problem, such as experimental design (Chernoff, 1959), statistical queries (Kearns, 1998; Fotakis et al., 2021), crowdsourcing (Doan et al., 2011), or aggregation methods in weak supervision (Ratner et al., 2020). More precisely, "active labeling"[1] consists in having several inputs and querying partial information on the labels. It is close to active learning (Settles, 2010; Dasgupta, 2011; Hanneke, 2014), where there are several inputs, but exact outputs are queried; and to active ranking (Valiant, 1975; Ailon, 2011; Braverman et al., 2019), where partial information is queried, but there is only one input. The streaming variant introduces privacy preserving constraints, a problem that is usually tackled through the notion of differential privacy (Dwork et al., 2006).

In terms of formalization, we build on the partial supervision formalization of Cabannes et al. (2020b), which casts weak supervision as sets of label candidates and generalizes semi-supervised learning (Chapelle et al., 2006). Finally, our sequential setting with a unique final reward is similar to combinatorial bandits in a pure-exploration setting (Garivier and Kaufmann, 2016; Fiez et al., 2019).

10.2 The "active labeling" problem

Supervised learning is traditionally modeled in the following manner. Consider $\mathcal{X}$ an input space, $\mathcal{Y}$ an output space, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a loss function, and $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ a joint probability distribution. The goal is to recover the function

$$f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f) := \mathbb{E}_{(X,Y)\sim\rho}[\ell(f(X), Y)], \tag{10.1}$$

yet, without accessing ρ, but a dataset of independent samples distributed according to ρ, $\mathcal{D}_n = (X_i, Y_i)_{i\leq n} \sim \rho^{\otimes n}$. In practice, accessing data comes at a cost, and it is valuable to understand the cheapest way to collect a dataset allowing to discriminate f^*.

We shall suppose that the input data $(X_i)_{i\leq n}$ are easy to collect, yet that labeling those inputs to get outputs $(Y_i)_{i\leq n}$ demands a high amount of work. For example, it is relatively easy to scrap the web or medical databases to access radiography images, but labeling them by asking radiologists to recognize tumors on zillions of radiographs will be both time-consuming and expensive. As a consequence, *we assume the $(X_i)_{i\leq n}$ given but the $(Y_i)_{i\leq n}$ unknown*. As getting information on the labels comes at a cost (*e.g.*, paying a pool of label workers, or spending your own time), given a budget constraint, what information should we query on the labels?

To quantify this problem, we will assume that *we can sequentially and adaptively query T information of the type $\mathbf{1}_{Y_{i_t}\in S_t}$, for any index $i_t \in \{1, \cdots, n\}$ and any set of labels $S_t \subset \mathcal{Y}$ (belonging to a specified set of subsets of $\mathcal{Y}$).* Here, $t \in \{1, \cdots, T\}$ indexes the query sequence, and $T \in \mathbb{N}$ is a fixed budget. The goal is to optimize the design of the sequence (i_t, S_t) in order to get the best estimate of f^* in terms of risk minimization (10.1). In the following, we give some examples to make this setting more concrete.

Example 18 (Classification with attributes)**.** *Suppose that a labeler is asked to provide fine-grained classes on images (Krause et al., 2016; Zheng et al., 2019), such as the label "caracal" in Figure 10.1. This would be difficult for many people. Yet, it is relatively easy to recognize that the image depicts a "feline" with "tufted-ears" and "sandy color". As such, a labeler can give the weak information that Y belongs to the set "feline", $S_1 = \{$"cat", "lion", "tiger", $\ldots\}$, and the set "tufted ears", $S_2 = \{$"Great horned owl", "Aruacana chicken", $\ldots\}$. This is enough to recognize that $Y \in S_1 \cap S_2 = \{$"caracal"$\}$. The question $\mathbf{1}_{Y\in S_1}$, corresponds to asking if the image depicts a feline. Literature on hierarchical classification and autonomic taxonomy construction provides interesting ideas for this problem (*e.g.*, Cesa-Bianchi et al., 2006; Gangaputra and Geman, 2006).*

Example 19 (Ranking with partial ordering)**.** *Consider a problem where for a given input x, characterizing a user, we are asked to deduce their preferences over m items. Collecting such a label requires knowing the exact ordering of the m items induced by a user. This might be hard to ask for. Instead, one can easily ask*

[1] Note that the wording "active labeling" has been more or less used as synonymous of "active learning" (*e.g.*, Wang and Shang, 2014). In contrast, we use "active labeling" to design "active weakly supervised learning".

Figure 10.1: Recognizing fine-grained classes is difficult, but recognizing attributes is easy.

the user which items they prefer in a collection of a few items. The user's answer will give weak information about the labels, which can be modeled as knowing $\mathbf{1}_{Y_i \in S} = 1$*, for* S *the set of total orderings that satisfy this partial ordering. We refer the curious reader to active ranking and dueling bandits for additional contents (Jamieson and Nowak, 2011; Bengs et al., 2021).*

Example 20 (Pricing a product)**.** *Suppose that we want to sell a product to a consumer characterized by some features* x*, this consumer is ready to pay a price* $y \in \mathbb{R}$ *for this product. We price it* $f(x) \in \mathbb{R}$*, and we observe* $\mathbf{1}_{f(x)<y}$*, that is if the consumer is willing to buy this product at this price tag or not (Cesa-Bianchi et al., 2019; Liu et al., 2021). Although, in this setting, the goal is often to minimize the regret, which contrasts with our pure exploration setting.*

As a counter-example, our assumptions are not set to deal with missing data, *i.e.* if some coordinates of some input feature vectors X_i are missing (Rubin, 1976). Typically, this happens when input data comes from different sources (*e.g.*, when trying to predict economic growth from country information that is self-reported).

Streaming variation. The special case of the active labeling problem we shall consider consists in its variant without resampling. This corresponds to the online setting where one can only ask one question by sample, formally $i_t = t$. This setting is particularly appealing for privacy concerns, in settings where the labels (Y_i) contain sensitive information that should not be revealed totally. For example, some people might be more comfortable giving a range over a salary rather than the exact value; or in the context of polling, one might not call back a previous respondent characterized by some features X_i to ask them again about their preferences captured by Y_i. Similarly, the streaming setting is relevant for web marketing, where inputs model new users visiting a website, queries model sets of advertisements chosen by an advertising company, and one observes potential clicks.

10.3 Weak information as stochastic gradients

In this section, we discuss how stochastic gradients can be accessed through weak information.

Suppose that we model $f = f_\theta$ for some Hilbert space $\Theta \ni \theta$. With some abuse of notations, let us denote $\ell(x, y, \theta) := \ell(f_\theta(x), y)$. We aim to minimize $\mathcal{R}(\theta) = \mathbb{E}_{(X,Y)}\left[\ell(X, Y, \theta)\right]$. Assume that $\mathcal{R}$ is differentiable (or sub-differentiable) and denote its gradients by $\nabla_\theta \mathcal{R}$.

Definition 89 (Stochastic gradient)**.** *A stochastic gradient of* $\mathcal{R}$ *is any random function* $G : \Theta \to \Theta$ *such that* $\mathbb{E}[G(\theta)] = \nabla_\theta \mathcal{R}(\theta)$*. Given some step size function* $\gamma : \mathbb{N} \to \mathbb{R}^*$*, a stochastic gradient descent (SGD) is a procedure,* $(\theta_t) \in \Theta^{\mathbb{N}}$*, initialized with some* θ_0 *and updated as* $\theta_{t+1} = \theta_t - \gamma(t)G(\theta_t)$*, where the realization of* $G(\theta_t)$ *given* θ_t *is independent of the previous realizations of* $G(\theta_s)$ *given* θ_s*.*

In supervised learning, SGD is usually performed with the stochastic gradients $\nabla_\theta \ell(X, Y, \theta)$. More generally, stochastic gradients are given by

$$G(\theta) = \mathbf{1}_{\nabla_\theta \ell(X,Y,\theta) \in T} \cdot \tau(T), \tag{10.2}$$

for $\tau : \mathcal{T} \to \Theta$ with $\mathcal{T} \subset 2^\Theta$ a set of subsets of Θ, and T a random variable on $\mathcal{T}$, such that

$$\forall\, \theta \in \Theta, \quad \mathbb{E}_T[\mathbf{1}_{\theta \in T} \cdot \tau(T)] = \theta. \tag{10.3}$$

Stated otherwise, if you have a way to image a vector θ from partial measurements $\mathbf{1}_{\theta \in T}$ such that you can reconstruct this vector in a linear fashion (10.3), then it provides you a generic strategy to get an unbiased stochastic estimate of this vector from a partial measurement (10.2).

For $\psi : \mathcal{Y} \to \Theta$ a function from $\mathcal{Y}$ to Θ (*e.g.*, $\psi = \nabla_\theta(X, \cdot, \theta)$), a question $\mathbf{1}_{\psi(Y) \in T}$ translates into a question $\mathbf{1}_{Y \in S}$ for some set $S = \psi^{-1}(T) \subset \mathcal{Y}$, meaning that the stochastic gradient (10.2) can be evaluated from a single query. As a proof of concept, we derive a generic implementation for T and τ in Appendix 10.B. This provides a generic SGD scheme to learn functions from weak queries when there are no constraints on the sets to query.

Remark 90 (Cutting plane methods)**.** *While we provide here a descent method, one could also develop cutting-plane/ellipsoid methods to localize θ^* according to weak information, which corresponds to the techniques developed for pricing by Cohen et al. (2020) and related literature.*

10.4 Median regression

In this section, we focus on efficiently acquiring weak information providing stochastic gradients for regression problems. In particular, we motivate and detail our methods for the absolute deviation loss.

Motivated by seminal works on censored data (Tobin, 1958), we shall suppose that *we query half-spaces*. For an output $y \in \mathcal{Y} = \mathbb{R}^m$, and any hyper-plane $z + u^\perp \subset \mathbb{R}^m$ for $z \in \mathbb{R}^m$, $u \in \mathbb{S}^{m-1}$, we can ask a labeler to tell us which half-space y belongs to. Formally, *we access the quantity* $\operatorname{sign}(\langle y - z, u\rangle)$ *for a given unit cost*.

Least-squares. For regression problems, it is common to look at the mean square loss

$$\ell(X, Y, \theta) = \|f_\theta(X) - Y\|^2, \qquad \nabla_\theta \ell(X, Y, \theta) = 2(f_\theta(X) - Y)^\top D f_\theta(X),$$

where $D f_\theta(x) \in \mathcal{Y} \otimes \Theta$ denotes the Jacobian of $\theta \to f_\theta(x)$. In rich parametric models, it is preferable to ask questions on $Y \in \mathcal{Y}$ rather than on gradients in Θ which is a potentially much bigger space. If we assume that Y and $f_\theta(X)$ are bounded in ℓ^2-norm by $M \in \mathbb{R}_+$, we can adapt (10.2) and (10.3) through the fact that for any $z \in \mathcal{Y}$, such that $\|z\| \leq 2M$, as proven in Appendix 10.B,

$$\mathbb{E}_{U,V}\left[\mathbf{1}_{\langle z, U\rangle \geq V} \cdot U\right] = c_1 \cdot z, \quad \text{where} \quad c_1 = \mathbb{E}_{U,V}\left[\mathbf{1}_{\langle e_1, U\rangle \geq V} \cdot \langle e_1, U\rangle\right] = \frac{\pi^{3/2}}{2M(m^2 + 4m + 3)},$$

for U uniform on the sphere $\mathbb{S}^{m-1}$ and V uniform on $[0, 2M]$. Applied to $z = f_\theta(X) - Y$, it designs an SGD procedure by querying information of the type $\mathbf{1}_{\langle Y, U\rangle < \langle f_\theta(X), U\rangle - V}$.

A case for median regression. Motivated by robustness purposes, we will rather expand on median regression. In general, we would like to learn a function that, given an input, replicates the output of I/O samples generated by the joint probability ρ. In many instances, X does not characterize all the sources of variations of Y, *i.e.* input features are not rich enough to characterize a unique output, leading to randomness in the conditional distributions $(Y|X)$. When many targets can be linked to a vector $x \in \mathcal{X}$, how to define a consensual $f(x)$? For analytical reasons, statisticians tend to use the least-squares error which corresponds to asking for $f(x)$ to be the mean of the distribution $(Y|X = x)$. Yet, means are known to be too sensitive to rare but large outputs (see *e.g.*, Huber, 1981), and cannot be defined as good and robust consensus in a world of heavy-tailed distributions. This contrasts with the median, which, as a consequence, is often much more valuable to summarize a range of values. For instance, median income is preferred over mean income as a population indicator (see *e.g.*, US Census Bureau, 2021).

Median regression. The geometric median is variationally defined through the absolute deviation loss, leading to

$$\ell(X, Y, \theta) = \|f_\theta(X) - Y\|, \qquad \nabla_\theta \ell(X, Y, \theta) = \left(\frac{f_\theta(X) - Y}{\|f_\theta(X) - Y\|}\right)^\top D f_\theta(X). \tag{10.4}$$

Similarly to the least-squares case, we can access weakly supervised stochastic gradients through the fact that for $z \in \mathbb{S}^{m-1}$, as shown in Appendix 10.B,

$$\mathbb{E}_U \left[\operatorname{sign} \left(\langle z, U \rangle \right) \cdot U \right] = c_2 \cdot z, \quad \text{where} \quad c_2 = \mathbb{E}_U \left[\operatorname{sign} \left(\langle e_1, U \rangle \right) \cdot \langle e_1, U \rangle \right] = \frac{\sqrt{\pi} \Gamma(\frac{m-1}{2})}{m \Gamma(\frac{m}{2})}, \tag{10.5}$$

where U is uniformly drawn on the sphere $\mathbb{S}^{m-1}$, and Γ is the gamma function. This suggests Algorithm 2.

Algorithm 2: Median regression with SGD.

Data: A model f_θ for $\theta \in \Theta$, some data $(X_i)_{i \leq n}$, a labeling budget T, a step size rule $\gamma : \mathbb{N} \to \mathbb{R}_+$
Result: A learned parameter $\hat{\theta}$ and the predictive function $\hat{f} = f_{\hat{\theta}}$.
Initialize θ_0.
for $t \leftarrow 1$ **to** T **do**
  Sample U_t uniformly on $\mathbb{S}^{m-1}$.
  Query $\varepsilon = \operatorname{sign}(\langle Y_t - z, U_t \rangle)$ for $z = f_{\theta_{t-1}}(X_t)$.
  Update the parameter $\theta_t = \theta_{t-1} + \gamma(t) \varepsilon \cdot U_t^\top (D f_{\theta_{t-1}}(X_t))$.
Output $\hat{\theta} = \theta_T$, or some average, *e.g.*, $\hat{\theta} = T^{-1} \sum_{t=1}^{T} \theta_t$.

10.5 Statistical analysis

In this section, we quantify the performance of Algorithm 2 by proving optimal rates of convergence when the median regression problem is approached with (reproducing) kernels. For simplicity, we will assume that f^* can be parametrized by a linear model (potentially of infinite dimension).

Assumption 20. *Assume that the solution $f^* : \mathcal{X} \to \mathbb{R}^m$ of the median regression problem* (10.1) *and* (10.4) *can be parametrized by some separable Hilbert space $\mathcal{H}$, and a bounded feature map $\varphi : \mathcal{X} \to \mathcal{H}$, such that, for any $i \in [m]$, there exists some $\theta_i^* \in \mathcal{H}$ such that $\langle f^*(\cdot), e_i \rangle_{\mathcal{Y}} = \left\langle \theta_i^*, \varphi(\cdot) \right\rangle_{\mathcal{H}}$, where (e_i) is the canonical basis of $\mathbb{R}^m$. Written into matrix form, there exists $\theta^* \in \mathcal{Y} \otimes \mathcal{H}$, such that $f^*(\cdot) = \theta^* \varphi(\cdot)$.*

The curious reader can easily relax this assumption in the realm of reproducing kernel Hilbert spaces following the work of Pillaud-Vivien et al. (2018a). Under the linear model of Assumption 20, Algorithm 2 is specified with $u^\top D f_\theta(x) = u \otimes \varphi(x)$. Note that rather than working with $\Theta = \mathcal{Y} \otimes \mathcal{H}$ which is potentially infinite-dimensional, empirical estimates can be represented in the finite-dimensional space $\mathcal{Y} \otimes \operatorname{Span} \{\varphi(X_i)\}_{i \leq n}$, and well approximated by small-dimensional spaces to ensure efficient computations (Williams and Seeger, 2000; Meanti et al., 2020).

One of the key points of SGD is that gradient descent is so gradual that one can use noisy or stochastic gradients without loosing statistical guarantees while speeding up computations. This is especially true when minimizing convex functions that are nor strongly-convex, *i.e.*, bounded below by a quadratic, nor smooth, *i.e.*, with Lipschitz-continuous gradient (see, *e.g.*, Bubeck, 2015). In particular, the following theorem, proven in Appendix 10.A.1, states that Algorithm 2 minimizes the population risk at a speed at least proportional to $O(T^{-1/2})$.

Theorem 20 (Convergence rates)**.** *Under Assumption 20, and under the knowledge of κ and M two real values such that $\mathbb{E}[\|\varphi(X)\|^2] \leq \kappa^2$ and $\|\theta^*\| \leq M$, with a budget $T \in \mathbb{N}$, a constant step size $\gamma = \frac{M}{\kappa \sqrt{T}}$ and the average estimate $\hat{\theta} = \frac{1}{T} \sum_{t=0}^{T-1} \theta_t$, Algorithm 2 leads to an estimate f that suffers from an excess of risk*

$$\mathbb{E} \left[\mathcal{R} \left(f_{\hat{\theta}} \right) \right] - \mathcal{R}(f^*) \leq \frac{2 \kappa M}{c_2 \sqrt{T}} \leq \kappa M m^{3/2} T^{-1/2}, \tag{10.6}$$

where the expectation is taken with respect to the randomness of $\hat{\theta}$ that depends on the dataset (X_i, Y_i) as well as the questions $(i_t, S_t)_{t \leq T}$.

While we give here a result for a fixed step size, one could retake the extensive literature on SGD to prove similar results for decaying step sizes that do not require to know the labeling budget in advance (*e.g.* setting $\gamma(t) \propto t^{-1/2}$ at the expense of an extra term in $\log(T)$ in front of the rates), as well as different averaging strategies (see *e.g.*, Bach, 2023). In practice, one might not know *a priori* the parameter M but could nonetheless find the right scaling for γ based on cross-validation.

The rate in $O(T^{-1/2})$ applies more broadly to all the strategies described in Section 10.3 as long as the loss ℓ and the parametric model f_θ ensure that $\mathcal{R}(\theta)$ is convex and Lipschitz-continuous. Although the constants appearing in front of rates depend on the complexity to reconstruct the full gradient $\nabla_\theta \ell(f_\theta(X_i, Y_i))$ from the reconstruction scheme (10.3). Those constants correspond to the second moment of the stochastic gradient. For example, for the least-squares technique described earlier one would have to replace c_2 by c_1 in (10.6).

Theorem 21, proven in Appendix 10.A.3, states that any algorithm that accesses a fully supervised learning dataset of size T cannot beat the rates in $O(T^{-1/2})$, hence any algorithm that collects weaker information on $(Y_i)_{i\leq T}$ cannot display better rates than the ones verified by Algorithm 2. This proves minimax optimality of our algorithm up to constants.

Theorem 21 (Minimax optimality)**.** *Under Assumption 20 and the knowledge of an upper bound on $\|\theta^*\| \leq M$, assuming that φ is bounded by κ, there exists a universal constant c_3 such that for any algorithm $\mathcal{A}$ that takes as input $\mathcal{D}_T = (X_i, Y_i)_{i\leq T} \sim \rho^{\otimes T}$ for any $T \in \mathbb{N}$ and output a parameter θ,*

$$\sup_{\rho\in\mathcal{M}_M} \mathbb{E}_{\mathcal{D}_T\sim\rho^{\otimes T}}\left[\mathcal{R}(f_{\mathcal{A}(\mathcal{D}_T;\rho)})\right] - \mathcal{R}(f_\rho;\rho) \geq c_3 M\kappa T^{-1/2}. \tag{10.7}$$

The supremum over $\rho \in \mathcal{M}_M$ has to be understood as the supremum over all distributions $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ such that the problem defined through the risk $\mathcal{R}(f;\rho) := \mathbb{E}_\rho[\ell(f(X), Y)]$ is minimized for f_ρ that verifies Assumption 20 with $\|\theta^*\|$ bounded by a constant M.

The same theorem applies for least-squares with a different universal constant. It should be noted that minimax lower bounds are in essence quantifying worst cases of a given class of problems. In particular, to prove Theorem 21, we consider distributions that lead to hard problems; more specifically, we assumed the variance of the conditional distribution $(Y \mid X)$ to be high. The practitioner should keep in mind that it is possible to add additional structure on the solution, leverage active learning or semi-supervised strategy such as uncertainty sampling (Nguyen et al., 2021), or Laplacian regularization (Zhu et al., 2003; Cabannes et al., 2021a), and reduce the optimal rates of convergence.

To conclude this section, let us remark that most of our derivations could easily be refined for practitioners facing a slightly different cost model for annotation. In particular, they might prefer to perform batches of annotations before updating θ rather than modifying the question strategy after each input annotation. This would be similar to mini-batching in gradient descent. Indeed, the dependency of our result on the annotation cost model and on Assumption 20 should not be seen as a limitation but rather as a proof of concept.

10.6 Numerical analysis

In this section, we illustrate the differences between our active method versus a classical passive method, for regression and classification problems. Extensive details are provided in Appendix 10.E. Our code is available online at `anonymized-url`.

Let us begin with the regression problem that consists in estimating the function f^* that maps $x \in [0, 1]$ to $\sin(2\pi x) \in \mathbb{R}$. Such a regular function, which belongs to any Hölder or Sobolev classes of functions, can be estimated with the Gaussian kernel, which would ensure Assumption 20, and that corresponds to a feature map φ such that $k(x, x') := \langle\varphi(x), \varphi(x')\rangle = \exp(-|x - x'|/(2\sigma^2))$ for any bandwidth parameter $\sigma > 0$.[2] On Figure 10.2, we focus on estimating f^* given data $(X_i)_{i\in[T]}$ that are uniform on $[0, 1]$ in the noiseless setting where $Y_i = f^*(X_i)$, based on the minimization of the absolute deviation loss. The passive baseline consists in randomly choosing a threshold $U_i \sim \mathcal{N}(0, 1)$ and acquiring the observations $(\mathbf{1}_{Y_i>U_i})_{i\in[T]}$ that can be cast as the observation of the half-space $S_i = \left\{y \in \mathcal{Y} \,\middle|\, \mathbf{1}_{y>U_i} = \mathbf{1}_{Y_i>U_i}\right\} =: s(Y_i, U_i)$. In this noiseless setting, a good baseline to learn f^* from the data (X_i, S_i) is provided by the infimum loss characterization (see Cabannes et al., 2020b)

$$f^* = \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,S)}[\inf_{y\in S} \ell(f(X), y)],$$

where the distribution over X corresponds to the marginal of ρ over $\mathcal{X}$, and the distribution over $(S \mid X = x)$ is the pushforward of $U \sim \mathcal{N}(0, 1)$ under $s(f^*(x), \cdot)$. The left plot on Figure 10.2 corresponds to an instance

[2] A noteworthy computational aspect of linear models, often refer as the "kernel trick", is that the features map φ does not need to be explicit, the knowledge of $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ being sufficient to compute all quantities of interest (Scholkopf and Smola, 2001). This "trick" can be applied to our algorithms.

of SGD on such an objective based on the data (X_i, S_i), while the right plot corresponds to Algorithm 2. We take the same hyperparameters for both plots, a bandwidth $\sigma = 0.2$ and an SGD step size $\gamma = 0.3$. We refer the curious reader to Figure 10.7 in Appendix 10.E for plots illustrating the streaming history, and to Figure 10.10 for "real-world" experiments.

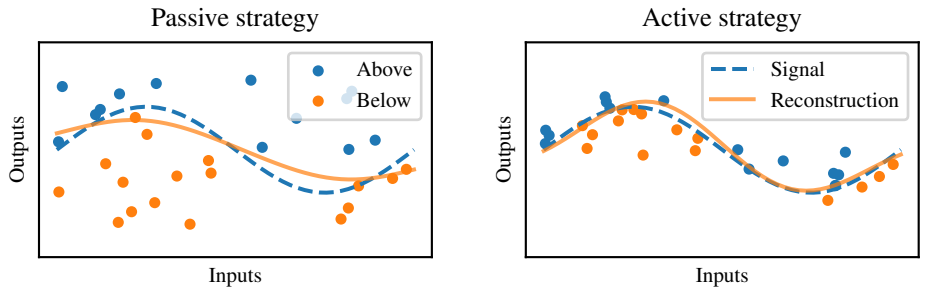

Figure 10.2: *Visual comparison of active and passive strategies.* Estimation in orange of the original signal f^* in dashed blue based on median regression in a noiseless setting. Any orange point $(x, u) \in \mathbb{R}^2$ corresponds to an observation made that u is below $f^*(x)$, while a blue point corresponds to u above $f^*(x)$. The passive strategy corresponds to acquiring information based on $(U \,|\, x)$ following a normal distribution, while the active strategy corresponds to $(u \,|\, x) = f_\theta(x)$. The active strategy reconstructs the signal much better given the budget of $T = 30$ observations.

To illustrate the versatility of our method, we approach a classification problem through the median surrogate technique presented in Proposition 91. To do so, we consider the classification problem with $m \in \mathbb{N}$ classes, $\mathcal{X} = [0, 1]$ and the conditional distribution $(Y \,|\, X)$ linearly interpolating between Dirac in y_1, y_2 and y_3 respectively for $x = 0$, $x = 1/2$ and $x = 1$ and the uniform distribution for $x = 1/4$ and $x = 3/4$; and X uniform on $\mathcal{X} \setminus ([1/4 - \varepsilon, 1/4 + \varepsilon] \cup [3/4 - \varepsilon, 3/4 + \varepsilon])$.

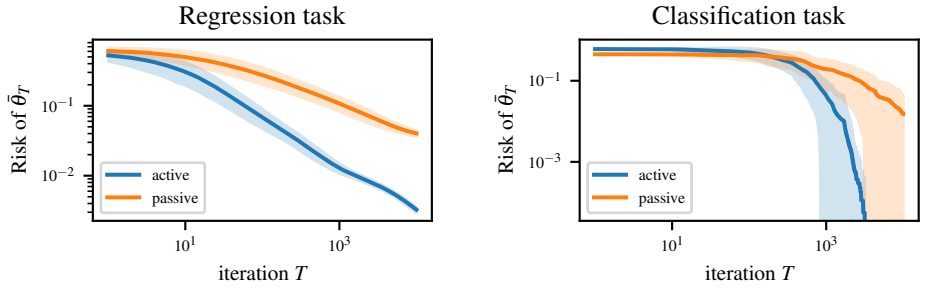

Figure 10.3: *Comparison of generalization errors of passive and active strategies* as a function of the annotation budget T. This error is computed by averaging over 100 trials. In solid is represented the average error, while the height of the dark area represents one standard deviation on each side. In order to consider the streaming setting where T is not known in advance, we consider the decreasing step size $\gamma(t) = \gamma_0/\sqrt{t}$; and to smooth out the stochasticity due to random gradients, we consider the average estimate $\bar{\theta}_t = (\theta_1 + \cdots + \theta_t)/t$. The left figure corresponds to the noiseless regression setting of Figure 10.2, with $\gamma_0 = 1$. We observe the convergence behavior in $O(T^{-1/2})$ of our active strategy. The right setting corresponds to the classification problem setting described in the main text with $m = 100$, $\varepsilon = 1/20$, and approached with the median surrogate. We observe the exponential convergence phenomenon described by Cabannes et al. (2021c); its kicks in earlier for the active strategy. The two plots are displayed with logarithmic scales on both axes.

10.7 Discussion

10.7.1 Discrete output problems

Learning problems with discrete output spaces are not as well understood as regression problems. This is a consequence of the complexity of dealing with combinatorial structures in contrast with continuous metric spaces. In particular, gradients are not defined for discrete output models. The current state-of-the-art framework to deal with discrete output problems is to introduce a continuous surrogate problem whose solution can be decoded as a solution on the original problem (Bartlett et al., 2006). For example, one could solve a classification task with a median regression surrogate problem, which is the object of the next proposition, proven in Appendix 10.C.

Proposition 91 (Consistency of median surrogate)**.** *The classification setting where $\mathcal{Y}$ is a finite space, and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the zero-one loss $\ell(y, z) = \mathbf{1}_{y \neq z}$ can be solved as a regression task through the simplex embedding of $\mathcal{Y}$ in $\mathbb{R}^{\mathcal{Y}}$ with the orthonormal basis $(e_y)_{y \in \mathcal{Y}}$. More precisely, if $g^* : \mathcal{X} \to \mathbb{R}^{\mathcal{Y}}$ is the minimizer of the median surrogate risk $\mathcal{R}_S(g) = \mathbb{E}\left[\|g(X) - e_Y\|\right]$, then $f^* : \mathcal{X} \to \mathcal{Y}$ defined as $f^*(x) = \arg\max_{y \in \mathcal{Y}} g^*_y(x)$ minimizes the original risk $\mathcal{R}(f) = \mathbb{E}\left[\ell(f(X), Y)\right]$.*

More generally, any discrete output problem can be solved by reusing the consistent least-squares surrogate of Ciliberto et al. (2020). Algorithm 2 can be adapted to the least-squares problem based on specifications at the beginning of Section 10.4. This allows using our method in an off-the-shelve fashion for all discrete output problems. In this setting, Theorem 20 can be refined under margin conditions where our approach would exhibit exponential convergence rates as illustrated on Figure 10.3. As a side note, while we are not aware of any generic theory encompassing the absolute-deviation surrogate of Proposition 91, we showcase its superiority over least-squares on at least two types of problems on Figures 10.4 and 10.5 in Appendix 10.C.

10.7.2 Supervised learning baseline with resampling

When resampling is allowed a simple baseline for the active labeling problem is provided by supervised learning. In regression problems with the query of any half-space, a method that consists in annotating each $(Y_i)_{i \leq n(T,\varepsilon)}$ up to precision ε, before using any supervised learning method to learn f from $(X_i, Y_i)_{i \leq n(T,\varepsilon)}$ could acquire $n(T, \varepsilon) \simeq T/m \log_2(\varepsilon^{-1})$ data points with a dichotomic search along all directions, assuming Y_i bounded or sub-Gaussian. In terms of minimax rates, such a procedure cannot perform better than in $n(T, \varepsilon)^{-1/2} + \varepsilon$, the first term being due to the statistical limit in Theorem 21, the second due to the incertitude ε on each Y_i that transfers to the same level of incertitude on f. Optimizing with respect to ε yields a bound in $O(T^{-1/2} \log(T)^{1/2})$. Therefore, this not-so-naive baseline is only suboptimal by a factor $\log(T)^{1/2}$. In the meanwhile, Algorithm 2 can be rewritten with resampling, as well as Theorem 20, which we prove in Appendix 10.A.2. Hence, our technique will still achieve minimax optimality for the problem "with resampling". In other terms, by deciding to acquire more imprecise information, our algorithm reduces annotation cost for a given level of generalization error (or equivalently reduces generalization error for a given annotation budget) by a factor $\log(T)^{1/2}$ when compared to this baseline.

The picture is slightly different for discrete-output problems. If one can ask any question $s \in 2^{\mathcal{Y}}$ then with a dichotomic search, one can retrieve any label with $\log_2(m)$ questions. Hence, to theoretically beat the fully supervised baseline with the SGD method described in Section 10.3, one would have to derive a gradient strategy (10.2) with a small enough second moment (*e.g.*, for convex losses that are non-smooth nor strongly convex, the increase in the second moment compared to the usual stochastic gradients should be no greater than $\log_2(m)^{1/2}$). How to best refine our technique to better take into account the discrete structure of the output space is an open question. Introducing bias that does not modify convergence properties while reducing variance eventually thanks to importance sampling is a potential way to approach this problem. A simpler idea would be to remember information of the type $Y_i \in s$ to restrict the questions asked in order to locate $f_{\theta_t}(X_i) - Y_i$ when performing stochastic gradient descent with resampling. Combinatorial bandits might also provide helpful insights on the matter. Ultimately, we would like to build an understanding of the whole distribution $(Y \mid X)$ and not only of $f^*(X)$ as we explore labels in order to refine this exploration.

10.7.3 Min-max approaches

Min-max approaches have been popularized for searching games and active learning, where one searches for the question that minimizes the size of the space where a potential guess could lie under the worst possible answer to that question. A particularly well illustrative example is the solution of the Mastermind game proposed by Knuth (1977). While our work leverages plain SGD, one could build on the vector field point-of-view of gradient descent (see, *e.g.*, Bubeck, 2015) to tackle min-max convex concave problems with similar guarantees. In particular, we could design weakly supervised losses $L(f(x), s; \mathbf{1}_{y \in s})$ and min-max games where a prediction player aims at minimizing such a loss with respect to the prediction f, while the query player aims at maximizing it with respect to the question s, that is querying information that best elicit mistakes made by the prediction player. For example, the dual norm characterization of the norm leads to the following min-max approach to the median regression

$$\operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f) = \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \max_{U \in (\mathbb{S}^{m-1})^{\mathcal{X}\times\mathcal{Y}}} \mathbb{E}_{(X,Y)\sim\rho} \left[\langle U(x,y), f(x) - y \rangle \right].$$

Such min-max formulations would be of interest if they lead to improvement of computational and statistical efficiencies, similarly to the work of Babichev et al. (2019). For classification problems, the following proposition introduces such a game and suggests its suitability. Its proof can be found in Appendix 10.D.

Proposition 92. *Consider the classification problem of learning $f^* : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{Y}$ is of finite cardinality, with the 0-1 loss $\ell(z, y) = \mathbf{1}_{z \neq y}$, minimizing the risk* (10.1) *under a distribution ρ on $\mathcal{X} \times \mathcal{Y}$. Introduce the surrogate score functions $g : \mathcal{X} \to \Delta_{\mathcal{Y}}; x \to v$ where $v = (v_y)_{y\in\mathcal{Y}}$ is a family of non-negative weights that sum to one, as well as the surrogate loss function $L : \Delta_{\mathcal{Y}} \times \mathcal{S} \times \{-1, 1\} \to \mathbb{R}; (v, S, \varepsilon) = \varepsilon(1 - 2\sum_{y\in S} v_y)$, and the min-max game*

$$\min_{g:\mathcal{X}\to\Delta_{\mathcal{Y}}} \max_{\mu:\mathcal{X}\to\Delta_{\mathcal{S}}} \mathbb{E}_{(X,Y)\sim\rho} \mathbb{E}_{S\sim\mu(x)} \left[L(g(x), S; \mathbf{1}_{Y\in S} - \mathbf{1}_{Y\notin S}) \right]. \tag{10.8}$$

When $\mathcal{S}$ contains the singletons and with the low-noise condition that $\mathbb{P}(Y \neq f^*(x) \mid X = x) < 1/2$ almost everywhere, then f^* can be learned through the relation $f^*(x) = \arg\min_{y\in\mathcal{Y}} g^*(x)_y$ for the unique minimizer g^* of (10.8)*. Moreover, the minimization of the empirical version of this objective with the stochastic gradient updates for saddle point problems provides a natural "active labeling" scheme to find this g^*.*

10.8 Conclusion

We have introduced the "active labeling" problem, which corresponds to "active partially supervised learning". We provided a solution to this problem based on stochastic gradient descent. Although our method can be used for any discrete output problem, we detailed how it works for median regression, where we show that it optimizes the generalization error for a given annotation budget. In a near future, we would like to focus on better exploiting the discrete structure of classification problems, eventually with resampling strategies.

Understanding more precisely the key issues in applications concerned with privacy, and studying how weak gradients might provide a good trade-off between learning efficiently and revealing too much information also provide interesting follow-ups. Finally, regarding dataset annotation, exploring different paradigms of weakly supervised learning would lead to different active weakly supervised learning frameworks. While this work is based on partial labeling, similar formalization could be made based on other weak supervision models, such as aggregation (*e.g.*, Ratner et al., 2020), or group statistics (Dietterich et al., 1997). In particular, annotating a huge dataset is often done by bagging inputs according to predicted labels and correcting errors that can be spotted on those bags of inputs (Deng et al., 2009). We left for future work the study of variants of the "active labeling" problem that model those settings.

Appendix

10.A Proofs of the statistical analysis

In the following proofs, we assume $\mathcal{X}$ to be Polish and $\mathcal{Y} = \mathbb{R}^m$, so to define the joint probability $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$. Moreover, we assume that $\mathbb{E}[\|Y\|] < +\infty$ in order to define the risk of median regression. We consider $\mathcal{H}$ to be a Hilbert space that is separable (*i.e.* only the origin is in all the neighborhood of the origin), and φ to be a measurable mapping from $\mathcal{X}$ to $\mathcal{H}$.

In terms of notations, we denote $\{1, 2, \cdots, n\}$ by $[n]$ for any $n \in \mathbb{N}^*$, and by $(x_i)_{i\leq n}$ the family $(x_1, \cdots, x_n)$ for any sequence (x_i). The unit sphere in $\mathbb{R}^m$ is denoted by $\mathbb{S}^{m-1}$. The symbol $\otimes$ denotes tensors, and is extended to product measures in the notation $\rho^{\otimes n} = \rho \times \rho \times \cdots \times \rho$. We have used the isometry between trace-class linear mappings from $\mathcal{H}$ to $\mathcal{Y}$ and the tensor space $\mathcal{Y} \otimes \mathcal{H}$, which generalizes the matrix representation of linear map between two finite-dimensional vector spaces. This space inherits from the Hilbertian structure of $\mathcal{H}$ and $\mathcal{Y}$ and we denote by $\|\cdot\|$ the Hilbertian norm that generalizes the Frobenius norm on linear maps between Euclidean spaces.

10.A.1 Upper bound for stochastic gradient descent

This subsection is devoted to the proof of Theorem 20. For simplicity, we will work with the rescaled step size $\gamma_t := c_2\gamma(t)$ rather than the step size described in the main text $\gamma(t)$.

Convergence of stochastic gradient descent for non-smooth problems is a known result. For completeness, we reproduce and adapt a usual proof to our setting. For $t \in \mathbb{N}$, let us introduce the random functions

$$\mathcal{R}_t(\theta) = c_2^{-1} \left|\langle \theta\varphi(X_t) - Y_t, U_t\rangle\right|, \qquad \text{where} \qquad c_2 = \mathbb{E}_U[|\langle e_1, U\rangle|] = \mathbb{E}_U[\operatorname{sign}(\langle e_1, U\rangle)\, \langle e_1, U\rangle]$$

for $(X_t, Y_t) \sim \rho$, U_t uniform on the sphere $\mathbb{S}^{m-1} \subset \mathcal{Y}$. Those random functions all average to $\mathcal{R}(\theta) = \mathbb{E}_\rho\, \mathbb{E}_U[c_2^{-1} \left|\langle \theta\varphi(X) - Y, U\rangle\right|] = \mathbb{E}_\rho[\|\theta\varphi(X) - Y\|]$. After a random initialization $\theta_0 \in \Theta$, the stochastic gradient update rule can be written for any $t \in \mathbb{N}$ as

$$\theta_{t+1} = \theta_t - \gamma_t \nabla\mathcal{R}_t(\theta_t),$$

where $\nabla\mathcal{R}_t$ denotes any sub-gradients of $\mathcal{R}_t$. We can compute

$$\nabla\mathcal{R}_t(\theta_t) = c_2^{-1}\nabla \left|\langle \theta\varphi(X_t) - Y_t, U_t\rangle\right| = c_2^{-1} \operatorname{sign}\left(\langle \theta\varphi(X_t) - Y_t, U_t\rangle\right) U_t \otimes \varphi(X_t).$$

This corresponds to the gradient written in Algorithm 2.

Let us now express the recurrence relation on $\|\theta_{t+1} - \theta^*\|$. We have

$$\begin{aligned}\|\theta_{t+1} - \theta^*\|^2 &= \|\theta_t - \gamma_t\nabla\mathcal{R}_t(\theta_t) - \theta^*\|^2 \\ &= \|\theta_t - \theta^*\|^2 + \gamma_t^2 \|\nabla\mathcal{R}_t(\theta_t)\|^2 - 2\gamma_t \left\langle \nabla\mathcal{R}_t(\theta_t), \theta_t - \theta^*\right\rangle.\end{aligned}$$

Because $\mathcal{R}_t$ is convex, it is above its tangents

$$\mathcal{R}_t(\theta^*) \geq \mathcal{R}_t(\theta_t) + \left\langle \nabla\mathcal{R}_t(\theta_t), \theta^* - \theta_t\right\rangle.$$

Hence,

$$\|\theta_{t+1} - \theta^*\|^2 \leq \|\theta_t - \theta^*\|^2 + \gamma_t^2 \|\nabla\mathcal{R}_t(\theta_t)\|^2 + 2\gamma_t(\mathcal{R}_t(\theta^*) - \mathcal{R}_t(\theta_t)).$$

This allows bounding the excess of risk as

$$2(\mathcal{R}_t(\theta_t) - \mathcal{R}_t(\theta^*)) \leq \frac{1}{\gamma_t}(\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2) + \gamma_t c_2^{-2} \|\varphi(X_t)\|^2 .$$

where we used the fact that $\|\nabla \mathcal{R}_t\| = c_2^{-1} \|\varphi(X_t)\|$. Let us multiply this inequality by $\eta_t > 0$ and sum from $t = 0$ to $t = T - 1$, we get

$$\begin{aligned} 2(\sum_{t=0}^{T-1} \eta_t \mathcal{R}_t(\theta_t) - \sum_{t=0}^{T-1} \eta_t \mathcal{R}_t(\theta^*)) &\leq \sum_{t=0}^{T-1} \frac{\eta_t}{\gamma_t}(\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2) + \sum_{t=0}^{T-1} \eta_t \gamma_t c_2^{-2} \|\varphi(X_t)\|^2 \\ &= \frac{\eta_0}{\gamma_0} \|\theta_0 - \theta^*\|^2 - \frac{\eta_{T-1}}{\gamma_{T-1}} \|\theta_T - \theta^*\|^2 + \sum_{t=1}^{T-1} \left(\frac{\eta_t}{\gamma_t} - \frac{\eta_{t-1}}{\gamma_{t-1}} \right) \|\theta_t - \theta^*\|^2 + \sum_{t=0}^{T-1} \eta_t \gamma_t c_2^{-2} \|\varphi(X_t)\|^2 . \end{aligned}$$

From here, there is several options to obtain a convergence result, either one assume $\|\theta_t - \theta^*\|$ bounded and take $\eta_t \gamma_{t-1} \geq \eta_{t-1} \gamma_t$; or one take $\eta_t = \gamma_t$ but at the price of paying an extra $\log(T)$ factor in the bound; or one take γ_t and η_t independent of t. Since we suppose the annotation budget given, we will choose γ_t and η_t independent of t, only depending on T.

$$2(\sum_{t=0}^{T-1} \eta \mathcal{R}_t(\theta_t) - \sum_{t=0}^{T-1} \eta \mathcal{R}_t(\theta^*)) \leq \frac{\eta}{\gamma} \|\theta_0 - \theta^*\|^2 + \sum_{t=0}^{T-1} \eta \gamma c_2^{-2} \|\varphi(X_t)\|^2 .$$

Let now take the expectation with respect to all the random variables, for the risk

$$\begin{aligned} \mathbb{E}_{(X_s, Y_s, U_s)_{s \leq t}} [\mathcal{R}_t(\theta_t)] &= \mathbb{E}_{(X_s, Y_s, U_s)_{s \leq t}} \left[\mathbb{E}_{(X_t, Y_t)} \left[\mathbb{E}_{U_t} \left[\mathcal{R}_t(\theta_t) \mid \theta_t \right] \middle| \theta_t \right] \right] \\ &= \mathbb{E}_{(X_s, Y_s, U_s)_{s \leq t}} \left[\mathcal{R}(\theta_t) \right] = \mathbb{E}[\mathcal{R}(\theta_t)]. \end{aligned}$$

For the variance, $\mathbb{E}[\|\varphi(X_s)\|^2] = \mathbb{E}[\|\varphi(X)\|^2] = \kappa^2$.

Let us fix T and consider $\eta_t = 1/T$, by Jensen we can bound the following averaging

$$\begin{aligned} 2\left(\mathcal{R}\left(\sum_{t=0}^{T-1} \eta_t \theta_t \right) - \mathcal{R}(\theta^*) \right) \leq 2\left(\sum_{t=0}^{T-1} \eta_t \mathcal{R}\left(\theta_t\right) - \mathcal{R}(\theta^*) \right) &= 2\,\mathbb{E}\left[\sum_{t=0}^{T-1} \eta_t (\mathcal{R}_t\left(\theta_t\right) - \mathcal{R}_t(\theta^*)) \right] \\ &\leq \frac{1}{T\gamma} \|\theta_0 - \theta^*\|^2 + \gamma c_2^{-2} \kappa^2. \end{aligned}$$

Initializing θ_0 to zero, we can optimize the resulting quantity to get the desired result.

10.A.2 Upper bound for resampling strategy

For resampling strategies, the proof is built on classical statistical learning theory considerations. Let us decompose the risk between estimation and optimization errors. Recall the expression of the risk $\mathcal{R}$, the function taking as inputs measurable functions from $\mathcal{X}$ to $\mathcal{Y}$ and outputting a real number

$$\mathcal{R}(f) = \mathbb{E}_\rho[\|f(X) - Y\|].$$

Let us denote by $\mathcal{F}$ the class of functions from $\mathcal{X}$ to $\mathcal{Y}$ we are going to work with. Let f_n be our estimate of f^* which maps almost every $x \in \mathcal{X}$ to the geometric median of $(Y \mid X)$. Denote by $\mathcal{R}^*_{\mathcal{D}_n}$ the best value that can be achieved by our class of functions to minimize the empirical average absolute deviation

$$\mathcal{R}^*_{\mathcal{D}_n} = \inf_{f \in \mathcal{F}} \mathcal{R}_{\mathcal{D}_n}(f).$$

Assumption 20 states that we have a well-specified model $\mathcal{F}$ to estimate the median, *i.e.* $f^* \in \mathcal{F}$. Hence, the excess of risk can be decomposed as an estimation and an optimization error, without approximation error (it is not difficult to add an approximation error, but it will make the derivations longer and the convergence rates harder to parse for the reader). Using the fact that $\mathcal{R}_{\mathcal{D}_n}(f^*) \geq \mathcal{R}^*_{\mathcal{D}_n}$ by definition of the infimum, we have

$$\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \underbrace{\mathcal{R}(f_n) - \mathcal{R}_{\mathcal{D}_n}(f_n) + \mathcal{R}_{\mathcal{D}_n}(f^*) - \mathcal{R}(f^*)}_{\text{estimation error}} + \underbrace{\mathcal{R}_{\mathcal{D}_n}(f_n) - \mathcal{R}^*_{\mathcal{D}_n}}_{\text{optimization error}} . \tag{10.9}$$

Estimation error. Let us begin by controlling the estimation error. We have two terms in it. $\mathcal{R}_{\mathcal{D}_n}(f^*) - \mathcal{R}(f^*)$ can be controlled with a concentration inequality on the empirical average of $\|f^*(X) - Y\|$ around its population mean. Assuming sub-Gaussian moments of Y, it can be done with Bernstein inequality.

$\mathcal{R}_{\mathcal{D}_n}(f_n) - \mathcal{R}(f_n)$ is harder to control as f_n depends on $\mathcal{D}_n$, so we can not use the same technique. The classical technique consists in going for the brutal uniform majoration,

$$\mathcal{R}(f_n) - \mathcal{R}_{\mathcal{D}_n}(f_n) \leq \sup_{f \in \mathcal{F}} \left(\mathcal{R}(f) - \mathcal{R}_{\mathcal{D}_n}(f) \right), \tag{10.10}$$

where $\mathcal{F}$ denotes the set of functions that f_n could be in concordance with our algorithm. While this bound could seem highly suboptimal, when the class of functions is well-behaved, we can indeed control the deviation $\mathcal{R}(f) - \mathcal{R}_{\mathcal{D}_n}(f)$ uniformly over this class without losing much (indeed for any class of functions, it is possible to build some really adversarial distribution ρ so that this supremum behaves similarly to the concentration we are looking for (Vapnik, 1995; Anthony and Bartlett, 1999)). This is particularly the case for our model linked with Assumption 20. Expectations of supremum processes have been extensively studied, allowing to get satisfying upper bounds (note that when the $\|f(X) - Y\|$ is bounded, deviation of the quantity of interest around its expectation can be controlled through McDiarmid inequality). In the statistical learning literature, it is usual to proceed with Rademacher complexity.

Lemma 93 (Uniform control of functions deviation with Rademacher complexity)**.** *The expectation of the excess of risk can be bounded as*

$$\frac{1}{2}\,\mathbb{E}_{\mathcal{D}_n}\left[\sup_{f \in \mathcal{F}} \left(\mathcal{R}(f) - \mathcal{R}_{\mathcal{D}_n}(f) \right)\right] \leq \mathfrak{R}_n(\mathcal{F}, \ell, \rho) := \frac{1}{n}\,\mathbb{E}_{\mathcal{D}_n,(\sigma_i)}\left[\sup_{f \in \mathcal{F}} \sigma_i \ell(f(X_i), Y_i)\right], \tag{10.11}$$

where $(\sigma_i)_{i \leq n}$ is defined as a family of Bernoulli independent variables taking value one or minus one with equal probability, and $\mathfrak{R}_n(\mathcal{F}, \ell, \rho)$ is called Rademacher complexity.

Proof. This results from the reduction to larger supremum and a symmetrization trick,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}\left[\sup_{f \in \mathcal{F}} \left(\mathcal{R}(f) - \mathcal{R}_{\mathcal{D}_n}(f) \right)\right] &= \mathbb{E}_{\mathcal{D}_n}\left[\sup_{f \in \mathcal{F}} \left(\mathbb{E}_{\mathcal{D}'_n}\, \mathcal{R}_{\mathcal{D}'_n}(f) - \mathcal{R}_{\mathcal{D}_n}(f) \right)\right] \\
&\leq \mathbb{E}_{\mathcal{D}_n}\, \mathbb{E}_{\mathcal{D}'_n}\left[\sup_{f \in \mathcal{F}} \left(\mathcal{R}_{\mathcal{D}'_n}(f) - \mathcal{R}_{\mathcal{D}_n}(f) \right)\right] \\
&= \mathbb{E}_{(X_i,Y_i),(X'_i,Y'_i)}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} \ell(f(X'_i), Y'_i) - \ell(f(X_i), Y_i) \right)\right] \\
&= \mathbb{E}_{(X_i,Y_i),(X'_i,Y'_i),(\sigma_i)}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} \sigma_i \left(\ell(f(X'_i), Y'_i) - \ell(f(X_i), Y_i) \right) \right)\right] \\
&\leq 2\,\mathbb{E}_{(X_i,Y_i),(\sigma_i)}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} \sigma_i \left(\ell(f(X_i), Y_i) \right) \right)\right],
\end{aligned}$$

which ends the proof. □

In our case, we want to compute the Rademacher complexity for ℓ given by the norm of $\mathcal{Y}$, and $\mathcal{F} = \{x \to \theta\varphi(x) \,|\, \theta \in \mathcal{Y} \otimes \mathcal{H}, \|\theta\| < M\}$, for $M > 0$ a parameter to specify in order to make sure that $\|\theta^*\| < M$, where the norm has to be understood as the ℓ^2-product norm on $\mathcal{Y} \otimes \mathcal{H} \simeq \mathcal{H}^m$. Working with linear models and Lipschitz losses is a well-known setting, allowing to derive directly the following bound.

Lemma 94 (Rademacher complexity of linear models with Lipschitz losses)**.** *The complexity of the linear class of vector-valued function $\mathcal{F} = \{x \to \theta\varphi(x) \,|\, \theta \in \mathcal{Y} \otimes \mathcal{H}, \|\theta\| < M\}$ is bounded as*

$$\mathbb{E}_{(\sigma_i)}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} \sigma_i \|f(x_i) - y_i\| \right)\right] \leq M\kappa n^{-1/2}. \tag{10.12}$$

Proof. This proposition is usually split in two. First using the fact that the composition of a space of functions with a Lipschitz function does not increase the entropy of the subsequent space (Vitushkin, 1954). Then bounding the Rademacher complexity of linear models. We refer to Maurer (2016) for a self-contained proof of this result (stated in its Section 4.3). □

Adding all the pieces together we have proven the following proposition, using the fact that the previous bound also applies to $\sup_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{D}_n}(f) - \mathcal{R}(f)$ by symmetry, hence it can be used for the deviation of $\mathcal{R}_{\mathcal{D}_n}(f^*) - \mathcal{R}(f^*)$.

Proposition 95 (Control of the estimation error)**.** *Under Assumption 20, with the model of computation $\mathcal{F} = \{x \in \mathcal{X} \to \theta\varphi(x) \in \mathcal{Y} \mid \|\theta\| \leq M\}$, the generalization error of f_n is controlled by a term in $n^{-1/2}$ plus an optimization error on the empirical risk minimization*

$$\mathbb{E}_{\mathcal{D}_n}\left[\mathcal{R}(f_n) - \mathcal{R}(f^*)\right] \leq \frac{4M\kappa}{n^{1/2}} + \mathbb{E}_{\mathcal{D}_n}\left[\mathcal{R}_{\mathcal{D}_n}(f_n) - \mathcal{R}^*_{\mathcal{D}_n}\right], \tag{10.13}$$

as long as $f^* \in \mathcal{F}$.

Note that this result can be refined using regularized risk (Sridharan et al., 2008), which would be useful under richer (stronger or weaker) source assumptions (*e.g.*, Caponnetto and De Vito, 2006). Such a refinement would allow switching from a constraint $\|\theta\| < M$ to define $\mathcal{F}$ to a regularization parameter $\lambda\|\theta\|^2$ added in the risk without restrictions on $\|\theta\|$, which would be better aligned with the current practice of machine learning. Under Assumption 20, this will not fundamentally change the result. The estimation error can be controlled with the derivation in Appendix 10.A.1, where stochastic gradients correspond to random sampling of a coefficient $i_t \leq n$ plus the choice of a random U_t. For the option without resampling, there exists an acceleration scheme specific to different losses in order to benefit from the strong convexity (*e.g.*, Bach and Moulines, 2013).

10.A.3 Lower bound

In this section, we prove Theorem 21. Let us consider any algorithm $\mathcal{A} : \cup_{n\in\mathbb{N}}(\mathcal{X}\times\mathcal{Y})^n \to \Theta$ that matches a dataset $\mathcal{D}_n$ to an estimate $\theta_{\mathcal{D}_n} \in \Theta$. Let us consider jointly a distribution ρ and a parameter θ such that Assumption 20 holds, that is $f_\rho := \arg\min_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_\rho[\ell(f(X), Y)] = f_\theta$. We are interested in characterizing for each algorithm the worst excess of risk it can achieve with respect to an adversarial distribution. The best worst performance that can be achieved by algorithms matching datasets to parameter can be written as

$$\mathcal{E} = \inf_{\mathcal{A}} \sup_{\theta\in\Theta, \rho\in\Delta_{\mathcal{X}\times\mathcal{Y}}; f_\rho = f_\theta} \mathbb{E}_{\mathcal{D}_n\sim\rho^{\otimes n}}\left[\mathbb{E}_{(X,Y)\sim\rho}\left[\ell(f_{\mathcal{A}(\mathcal{D}_n)}(X), Y) - \ell(f_\theta(X), Y)\right]\right]. \tag{10.14}$$

This provides a lower bound to upper bounds such as (10.6) that can be derived for any algorithm. There are many ways to get lower bounds on this quantity. Ultimately, we want to quantify the best certainty one can have on an estimate θ based on some observations $(X_i, Y_i)_{i\leq n}$. In particular, the algorithms $\mathcal{A}$ can be seen as rules to discriminate a model θ from observations $\mathcal{D}_n$ made under ρ_θ, and where the error is measured through the excess of risk $\mathcal{R}(f_{\hat\theta}, \rho_\theta) - \mathcal{R}(f_\theta; \rho_\theta)$ where $\mathcal{R}(f;\rho) = \mathbb{E}_\rho[\ell(f(X), Y)]$ and ρ_θ is a distribution parametrized by θ such that $f_\theta = f_\rho$.

Let us first characterize the measure of error. Surprisingly, when in presence of Gaussian noise or uniform noise, the excess of risk behaves like a quadratic metric between parameters.

Lemma 96 (Quadratic behavior of the median regression excess of risk with Gaussian noise)**.** *Consider the random variable $Y \sim \mathcal{N}(\mu, \sigma^2 I_m)$, denote by $\hat\mu$ an estimate of μ, the excess of risk can be developed as*

$$\mathbb{E}_{\mathcal{N}(\mu,\sigma^2 I_m)}[\|\hat\mu - Y\| - \|\mu - Y\|] = \frac{c_4\|\hat\mu - \mu\|^2}{\sigma} + o\left(\frac{\|\hat\mu - \mu\|^3}{\sigma^2}\right), \tag{10.15}$$

where $c_4 = \Gamma(\frac{m+1}{2})/(2\sqrt{2}\Gamma(\frac{m+2}{2})) \geq (m+2)^{-1/2}/2$.

Proof. With this specific noise model, one can do the following derivations.

$$\mathbb{E}_{\mathcal{N}(\mu,\sigma^2 I_m)}[\|\hat\mu - Y\|] = \mathbb{E}_{\mathcal{N}(0,I_m)}[\|\hat\mu - \mu - \sigma Y\|] = \sigma\,\mathbb{E}_{\mathcal{N}(0,I_m)}\left[\left\|\frac{\hat\mu - \mu}{\sigma} - Y\right\|\right].$$

We recognize the mean of a non-central χ-distribution of parameter $k = m$ and $\lambda = \left\|\frac{\hat{\mu}-\mu}{\sigma}\right\|$. It can be expressed through the generalized Laguerre functions, which allows us to get the following Taylor expansion

$$\begin{aligned}\mathbb{E}_{\mathcal{N}(\mu,\sigma^2 I_m)}[\|\hat{\mu}-Y\|] &= \frac{\sqrt{\pi}\sigma}{\sqrt{2}} L_{\frac{1}{2}}^{(\frac{m-2}{2})}\left(-\frac{\|\hat{\mu}-\mu\|^2}{2\sigma^2}\right) \\ &= \frac{\sqrt{\pi}\sigma}{\sqrt{2}}\left(L_{\frac{1}{2}}^{(\frac{m-2}{2})}(0) + \frac{\|\hat{\mu}-\mu\|^2}{2\sigma^2} L_{-\frac{1}{2}}^{(\frac{m}{2})}(0)\right) + o\left(\frac{\|\hat{\mu}-\mu\|^3}{\sigma^2}\right).\end{aligned}$$

Hence, the following expression of the excess of risk,

$$\begin{aligned}\mathbb{E}_{\mathcal{N}(\mu,\sigma^2 I_m)}[\|\hat{\mu}-Y\| - \|\mu-Y\|] &= \frac{\sqrt{\pi}\,\|\hat{\mu}-\mu\|^2}{2\sqrt{2}\sigma} L_{-\frac{1}{2}}^{(\frac{m}{2})}(0) + o\left(\frac{\|\hat{\mu}-\mu\|^3}{\sigma^2}\right) \\ &= \frac{\Gamma(\frac{m+1}{2})\,\|\hat{\mu}-\mu\|^2}{2\sqrt{2}\Gamma(\frac{m+2}{2})\sigma} + o\left(\frac{\|\hat{\mu}-\mu\|^3}{\sigma^2}\right).\end{aligned}$$

Note that in dimension one, the calculation can be done explicitly by computing integrals with the error function.

$$\begin{aligned}\mathbb{E}_{\mathcal{N}(\mu,\sigma^2)}[\|\hat{\mu}-Y\|] &= \sigma\,\mathbb{E}_{\mathcal{N}(0,1)}\left[Y - \frac{\hat{\mu}-\mu}{\sigma} + 2\mathbf{1}_{Y<\frac{\hat{\mu}-\mu}{\sigma}}\left(\frac{\hat{\mu}-\mu}{\sigma} - Y\right)\right] \\ &= \mu - \hat{\mu} + 2(\hat{\mu}-\mu)\,\mathbb{E}_{\mathcal{N}(0,1)}\left[\mathbf{1}_{Y<\frac{\hat{\mu}-\mu}{\sigma}}\right] - 2\sigma\,\mathbb{E}_{\mathcal{N}(0,1)}\left[Y\mathbf{1}_{Y<\frac{\hat{\mu}-\mu}{\sigma}}\right] \\ &= \mu - \hat{\mu} + 2(\hat{\mu}-\mu)\left(\frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{\hat{\mu}-\mu}{\sqrt{2}\sigma}\right)\right) - \frac{\sqrt{2}\sigma}{\sqrt{\pi}}\int_{-\infty}^{\frac{\hat{\mu}-\mu}{\sigma}} y e^{-\frac{y^2}{2}}\,\mathrm{d}y \\ &= (\hat{\mu}-\mu)\operatorname{erf}\left(\frac{\hat{\mu}-\mu}{\sqrt{2}\sigma}\right) - \frac{\sqrt{2}\sigma}{\sqrt{\pi}} e^{-\frac{(\hat{\mu}-\mu)^2}{2\sigma^2}},\end{aligned}$$

where we used the error function, which is the symmetric function defined for $x \in \mathbb{R}_+$ as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,\mathrm{d}t = \frac{2}{\sqrt{2\pi}}\int_0^{\sqrt{2}x} e^{-\frac{u^2}{2}}\,\mathrm{d}u = 2\,\mathbb{E}_{\mathcal{N}(0,1)}[\mathbf{1}_{0\le Y\le\sqrt{2}x}].$$

Developing those two functions in the Taylor series leads to the same quadratic behavior. □

Let us now add a context variable.

Lemma 97 (Reduction to least-squares)**.** *For $\mathcal{Y} = \mathbb{R}^m$, there exists a $\sigma_m > 0$, such that if φ is bounded by κ, and f^* belongs to the class of functions $\mathcal{F} = \{x \to \theta\varphi(x) \mid \theta \in \mathcal{Y}\otimes\mathcal{H}, \|\theta\| \le M\}$, and the conditional distribution are distributed as $(Y \mid X) \sim \mathcal{N}(f^*(x), \sigma^2 I_m)$, with $\sigma > 2M\kappa\sigma_m$,*

$$\forall\, f \in \mathcal{F}, \qquad \mathcal{R}(f) - \mathcal{R}(f^*) \ge \frac{c_4\,\|f - f^*\|^2_{L^2(\rho_{\mathcal{X}})}}{2\sigma}. \tag{10.16}$$

Proof. According to the precedent lemma, there exists σ_m such that $\|\hat{\mu}-\mu\|\,\sigma^{-1} \le \sigma_m^{-1}$ leads to[3]

$$\mathbb{E}_{\mathcal{N}(\mu,\sigma^2 I_m)}[\|\hat{\mu}-Y\| - \|\mu-Y\|] \ge \frac{c_4\,\|\hat{\mu}-\mu\|^2}{2\sigma}.$$

Let f and $f^* \in \mathcal{F}$ be parametrized by θ and θ^*. For a given x, setting $\hat{\mu} = f_\theta(x) = \theta\varphi(x)$ and $\mu = f_{\theta^*}(x)$, we get that, using the operator norm,

$$\|\hat{\mu}-\mu\| = \|(\theta-\theta^*)\varphi(x)\| \le \|\theta-\theta^*\|_{\mathrm{op}}\,\|\varphi(x)\| \le \|\theta-\theta^*\|\,\|\varphi(x)\| \le 2M\kappa.$$

Hence, as soon as $2M\kappa \le \sigma\sigma_m^{-1}$, we have that for almost all $x \in \mathcal{X}$

$$\mathbb{E}_Y\left[\|f(X)-Y\| - \|f^*(X)-Y\| \mid X = x\right] \ge \frac{c_4\,\|f(X)-f^*(X)\|^2}{2\sigma}.$$

The result follows from integration over $\mathcal{X}$. □

[3]This best value for σ_m can be derived by studying the Laguerre function, which we will not do in this paper.

We now have a characterization of the excess of risk that will allow us to reuse lower bounds for least-squares regression. We will follow the exposition of Bach (2023) that we reproduce and comment here for completeness. It is based on the generalized Fano's method (Ibragimov and Khas'minskii, 1977; Birgé, 1983).

Learnability over a class of functions depends on the size of this class of functions. For least-squares regression with a Hilbert class of functions, the right notion of size is given by the Kolmogorov entropy. Let us call ε-packing of $\mathcal{F}$ with a metric d any family $(f_i)_{i\leq N} \in \mathcal{F}^N$ such that $d(f_i, f_j) > \varepsilon$. The logarithm of the maximum cardinality of an ε-packing defines the ε-capacity of the class of functions $\mathcal{F}$. We refer the interested reader to Theorem 6 in Kolmogorov and Tikhomirov (1959) to make a link between the notions of capacity and entropy of a space. To be perfectly rigorous, the least-squares error in not a norm on the space of L^2 functions, but we will call it a *quasi-distance* as it verifies symmetry, positive definiteness and the inequality $d(x, y) \leq K(d(x, z) + d(z, y))$ for $K \geq 1$. Let us define an ε-packing with respect to a quasi-distance similarly as before.

The ε-capacity of a space $\mathcal{F}$ gives a lower bound on the number of information to transmit in order to recover a function in $\mathcal{F}$ up to precision ε. We will leverage this fact in order to show our lower bound. Let us first reduce the problem to a statistical test.

Lemma 98 (Reduction to statistical testing)**.** *Let us consider a class of functions $\mathcal{F}$ and an ε-packing $(f_i)_{i\leq N}$ of $\mathcal{F}$ with respect to a quasi-distance $d(\cdot,\cdot)$ verifying the triangular inequality up to a multiplicative factor K. Then the minimax optimality of an algorithm $\mathcal{A}$ that takes as input the dataset $\mathcal{D}_n = (X_i, Y_i)_{i\leq n}$ and output a function in $\mathcal{F}$ can be related to the minimax optimality of an algorithm $\mathcal{C}$ that takes an input the dataset $\mathcal{D}_n$ and output an index $j \in [m]$ through*

$$\inf_{\mathcal{A}} \sup_{\rho} \mathbb{E}_{\mathcal{D}_n\sim\rho^{\otimes n}} \left[d\left(f_{\mathcal{A}(\mathcal{D}_n)}, f_\rho\right)\right] \geq \frac{\varepsilon}{2K} \inf_{\mathcal{C}} \sup_{i\in[N]} \mathbb{P}_{\mathcal{D}_n\sim(\rho_i)^{\otimes n}} \left(\mathcal{C}(\mathcal{D}_n) \neq i\right), \tag{10.17}$$

where the supremum over ρ has to be understood as taken over all measures whose marginals can be written $\mathcal{N}(f^*(x), \sigma)$ for σ bigger than a threshold σ_m and $f^* \in \mathcal{F}$, and the supremum over ρ_i taken over the same type of measures with $f^* \in (f_i)_{i\leq N}$.

Proof. Consider an algorithm $\mathcal{A}$ that takes as input a dataset $\mathcal{D}_n = (X_j, Y_j)_{j\leq n}$ and output a function $f \in \mathcal{F}$. We would like to see $\mathcal{A}$ as deriving from a classification rule and relate the classification and regression errors. The natural classification rule associated with the algorithm $\mathcal{A}$ can be defined through π the projection from $\mathcal{F}$ to $[N]$ that minimizes $d\left(f, f_{\pi(f)}\right)$. The classification error and regression error made by $\pi \circ \mathcal{A}$ can be related thanks to the ε-packing property. For any index $j \in [N]$

$$d\left(f_{\pi\circ\mathcal{A}(\mathcal{D}_n)}, f_j\right) \geq \varepsilon \mathbf{1}_{\pi\circ\mathcal{A}(\mathcal{D}_n)\neq j}.$$

The error made by $f_{\mathcal{A}}(\mathcal{D}_n)$ relates to the one made by $f_{\pi\circ\mathcal{A}(\mathcal{D}_n)}$ thanks to the modified triangular inequality, using the definition of the projection

$$d\left(f_{\pi\circ\mathcal{A}(\mathcal{D}_n)}, f_j\right) \leq K\left(d\left(f_{\pi\circ\mathcal{A}(\mathcal{D}_n)}, f_{\mathcal{A}(\mathcal{D}_n)}\right) + d\left(f_{\mathcal{A}(\mathcal{D}_n)}, f_j\right)\right) \leq 2Kd\left(f_{\mathcal{A}(\mathcal{D}_n)}, f_j\right).$$

Finally,

$$d\left(f_{\mathcal{A}(\mathcal{D}_n)}, f_j\right) \geq \frac{\varepsilon}{2K} \mathbf{1}_{\pi\circ\mathcal{A}(\mathcal{D}_n)\neq j}.$$

Assuming that the data were generated by a ρ_i and taking the expectation, the supremum over ρ_i and the infimum over $\mathcal{A}$ leads to

$$\inf_{\mathcal{A}} \sup_{\rho_i} \mathbb{E}_{\mathcal{D}_n\sim\rho_i^{\otimes n}} \left[d\left(f_{\mathcal{A}(\mathcal{D}_n)}, f_i\right)\right] \geq \frac{\varepsilon}{2K} \inf_{\mathcal{C}=\pi\circ\mathcal{A}} \sup_{(\rho_i)} \mathbb{P}_{\mathcal{D}_n\sim\rho_i^{\otimes n}} \left(\mathcal{C}(\mathcal{D}_n) \neq i\right).$$

Because $\pi \circ \mathcal{A}$ are part of classification rules (indeed it parametrizes all the classification rules, simply consider $\mathcal{A}$ that matches a dataset to one of the functions $(f_i)_{i\leq N}$), and because the distributions ρ_i are part of the distributions ρ defined in the lemma, this last equation implies the stated result. □

One of the harshest inequalities in the last proof is due to the usage of the ε-packing condition without considering error made by $d\left(f_{\pi\circ\mathcal{A}(\mathcal{D}_n)}, f_j\right)$ that might be much worse than ε. We will later add a condition on the ε-packings to ensure that the (f_i) are not too far from each other. This will not be a major problem when considering small balls in big dimension spaces.

Results from statistical testing

In this section, we expand on lower bounds for statistical testing. We refer the curious reader to Cover and Thomas (1991). We begin by relaxing the supremum by an average

$$\inf_{\mathcal{C}} \sup_{i\in[N]} \mathbb{P}_{\mathcal{D}_n\sim(\rho_i)^{\otimes n}}\left(\mathcal{C}(\mathcal{D}_n)\neq i\right) = \inf_{\mathcal{C}} \sup_{p\in\Delta_N} \sum_{i=1}^{N} p_i\, \mathbb{P}_{\mathcal{D}_n\sim(\rho_i)^{\otimes n}}\left(\mathcal{C}(\mathcal{D}_n)\neq i\right) \tag{10.18}$$

$$\geq \inf_{\mathcal{C}} \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_{\mathcal{D}_n\sim(\rho_i)^{\otimes n}}\left(\mathcal{C}(\mathcal{D}_n)\neq i\right). \tag{10.19}$$

The last quantity can be seen as the best measure of error that can be achieved by a decoder $\mathcal{C}$ of a signal $i\in[N]$ based on noisy observations $\mathcal{D}_n$ of the signal. A lower bound on such a similar quantity is the object of Fano's inequality (Fano, 1968).

Lemma 99 (Fano's inequality)**.** *Let (X,Y) be a couple of random variables in $\mathcal{X}\times\mathcal{Y}$ with $\mathcal{X}$, $\mathcal{Y}$ finite, and $\hat{X}:\mathcal{Y}\to\mathcal{X}$ be a classification rule. Then, the error $e=e(X,Y)=\mathbf{1}_{X\neq\hat{X}(Y)}$ verifies*

$$H\left(X\,|\,Y\right)\leq H(e)+\mathbb{P}(e)\log(|\mathcal{X}|-1)\leq\log(2)+\mathbb{P}(e)\log(|\mathcal{X}|).$$

Where for $(X,Y)\in\Delta_{\mathcal{X}\times\mathcal{Y}}$, $H(X)$ and $H\left(X\,|\,Y\right)$ denotes the entropy and conditional entropy, defined as, with the convention $0\log 0=0$,

$$H(X)=-\sum_{x\in\mathcal{X}}\mathbb{P}(X=x)\log(\mathbb{P}(X=x)),$$
$$H\left(X\,|\,Y\right)=-\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\mathbb{P}(X=x,Y=y)\log(\mathbb{P}\left(X=x\,|\,Y=y\right)).$$

Proof. This lemma is actually the result of two properties. The first part of the proof is due to some manipulation of the entropy, consisting in showing that

$$H\left(X\,\middle|\,\hat{X}(Y)\right)\leq H(e)+\mathbb{P}(e)\log(|\mathcal{X}|-1). \tag{10.20}$$

Let us first recall the following additive property of entropy

$$\begin{aligned}
H\left(X,Y\,|\,Z\right) &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}}\mathbb{P}(X=x,Y=y,Z=z)\log(\mathbb{P}\left(X=x,Y=y\,|\,Z=z\right))\\
&= -\sum_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}}\mathbb{P}(X=x,Y=y,Z=z)\log(\mathbb{P}\left(Y=y\,|\,X=x,Z=z\right))\\
&\qquad -\sum_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}}\mathbb{P}(X=x,Y=y,Z=z)\log(\mathbb{P}\left(X=x\,|\,Z=z\right))\\
&= H\left(Y\,|\,X,Z\right)+H\left(X\,|\,Z\right).
\end{aligned}$$

Using this chain rule, we get

$$\begin{aligned}
H\left(e,X\,\middle|\,\hat{X}\right) &= H\left(e\,\middle|\,X,\hat{X}\right)+H\left(X\,\middle|\,\hat{X}\right)\\
&= H\left(X\,\middle|\,e,\hat{X}\right)+H\left(e\,\middle|\,\hat{X}\right)
\end{aligned}$$

Because e is a function of $\hat{X}$ and X one can check that $H\left(e\,\middle|\,X,\hat{X}\right)=0$,

$$\begin{aligned}
H\left(e\,\middle|\,X,\hat{X}\right) &= -\sum_{e,X,\hat{X}}\mathbb{P}(X,\hat{X})\,\mathbb{P}\left(e\,\middle|\,X,\hat{X}\right)\log(\mathbb{P}\left(e\,\middle|\,X,\hat{X}\right))\\
&= -\sum_{e,X,\hat{X}}\mathbb{P}(X,\hat{X})\mathbf{1}_{e=\mathbf{1}_{X\neq\hat{X}}}\log(\mathbf{1}_{e=\mathbf{1}_{X\neq\hat{X}}}) = -\sum_{e,X,\hat{X}}\mathbb{P}(X,\hat{X})\cdot 0=0.
\end{aligned}$$

Using Jensen inequality for the logarithm, we get

$$\begin{aligned}
H\left(X \,\middle|\, e, \hat{X}\right) &= -\sum_{X,e,\hat{X}} \mathbb{P}(X,e,\hat{X}) \log(\mathbb{P}\left(X \,\middle|\, e, \hat{X}\right)) \\
&= -\sum_{x,x'} \mathbb{P}(X=x, e=0, \hat{X}=x') \log(\mathbb{P}\left(X=x \,\middle|\, e=0, \hat{X}=x'\right)) \\
&\qquad\qquad - \mathbb{P}(X=x, e=1, \hat{X}=x') \log(\mathbb{P}\left(X=x \,\middle|\, e=1, \hat{X}=x'\right)) \\
&= -\sum_{x,x'} \mathbb{P}\left(X=x, \hat{X}=x'\right) \mathbf{1}_{x=x'} \log(\mathbf{1}_{x=x'}) \\
&\qquad\qquad - \mathbb{P}(e=1)\mathbf{1}_{x\neq x'}\, \mathbb{P}(X=x, \hat{X}=x') \log(\mathbb{P}\left(X=x \,\middle|\, \hat{X}=x'\right)) \\
&= \mathbb{P}(e=1) \sum_{x'} \mathbb{P}(\hat{X}=x') \sum_{x\neq x'} \mathbb{P}\left(X=x \,\middle|\, \hat{X}=x'\right) \log\left(\frac{1}{\mathbb{P}\left(X=x \,\middle|\, \hat{X}=x'\right)}\right) \\
&\leq \mathbb{P}(e=1) \sum_{x'} \mathbb{P}(\hat{X}=x') \log\left(\sum_{x\neq x'} \mathbb{P}\left(X=x \,\middle|\, \hat{X}=x'\right) \frac{1}{\mathbb{P}\left(X=x \,\middle|\, \hat{X}=x'\right)}\right) \\
&= \mathbb{P}(e=1) \log(|\mathcal{X}|-1).
\end{aligned}$$

Using that conditioning reduces the entropy, which follows again from Jensen inequality,

$$\begin{aligned}
H(X) - H\left(X \,|\, Y\right) &= \sum_{x,y} \mathbb{P}(X=x, Y=y) \log\left(\frac{\mathbb{P}\left(X=x \,|\, Y=y\right)}{\mathbb{P}\left(X=x\right)}\right) \\
&= -\sum_{x,y} \mathbb{P}(X=x, Y=y) \log\left(\frac{\mathbb{P}\left(X=x\right)\mathbb{P}\left(Y=y\right)}{\mathbb{P}\left(X=x, Y=y\right)}\right) \\
&\geq -\log\left(\sum_{x,y} \mathbb{P}(X=x, Y=y) \frac{\mathbb{P}\left(X=x\right)\mathbb{P}\left(Y=y\right)}{\mathbb{P}\left(X=x, Y=y\right)}\right) = 0,
\end{aligned}$$

we get

$$H\left(e \,\middle|\, \hat{X}\right) \leq H(e) \leq \log(2).$$

Hence, we have proven that

$$H\left(X \,\middle|\, \hat{X}\right) \leq \mathbb{P}(e=1)\log(|\mathcal{X}|-1) + H(e).$$

The rest of the proof follows from the so-called data processing inequality, that is

$$H\left(X \,\middle|\, \hat{X}(Y)\right) \geq H\left(X \,|\, Y\right). \tag{10.21}$$

We will not derive it here, since it will not be used in the following. □

In our case, a slight modification of the proof of Fano's inequality leads to the following Proposition.

Lemma 100 (Generalized Fano's method)**.** *For any family of distributions $(\rho_i)_{i\leq N}$ on $\mathcal{X}\times\mathcal{Y}$ with $N\in\mathbb{N}^*$, any classification rule $\mathcal{C}:\mathcal{D}_n\to[N]$ cannot beat the following average lower bound*

$$\inf_{\mathcal{C}} \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_{\mathcal{D}_n\sim\rho_i^{\otimes n}}(\mathcal{C}(\mathcal{D}_n)\neq i)\log(N-1) \geq \log(N) - \log(2) - \frac{n}{N^2}\sum_{i,j\in[N]} K\left(\rho_i \,||\, \rho_j\right), \tag{10.22}$$

where $K\left(p \,||\, q\right)$ is the Kullback-Leibler divergence defined for any measure p absolutely continuous with respect to a measure q as

$$K\left(p \,||\, q\right) = \mathbb{E}_{X\sim q}\left[-\log\left(\frac{\mathrm{d}p(X)}{\mathrm{d}q(X)}\right)\right].$$

Proof. Let us consider the joint variable (X,Y) where X is a uniform variable on $[N]$ and $(Y \,|\, X)$ is distributed according to $\rho_X^{\otimes n}$. For any classification rule $\hat{X}:\mathcal{D}_n\to[N]$, using (10.20) we get

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_{\mathcal{D}_n\sim\rho_i^{\otimes n}}\left(\hat{X}(\mathcal{D}_n)\neq i\right) = \mathbb{P}(\hat{X}\neq X)\log(N-1) \geq H\left(X \,\middle|\, \hat{X}\right) - \log(2).$$

We should work on $H\left(X \,\middle|\, \hat{X} \,\middle|\, X\right)$ with similar ideas to the data processing inequality. First of all, using the chain rule for entropy

$$H\left(X \,\middle|\, \hat{X}\right) = H(X, \hat{X}) - H(\hat{X}) = H(X) + (H(X, \hat{X}) - H(X) - H(\hat{X})) = \log(N) - I(X, \hat{X}),$$

where I is the mutual information defined as, for X and Z discrete

$$\begin{aligned} I(X, Z) &= H(X) + H(Z) - H(X, Z) = \sum_{x,z} \mathbb{P}\left(X = x, Z = z\right) \log\left(\frac{\mathbb{P}(X = x, Z = z)}{\mathbb{P}(X = x)\,\mathbb{P}(Z = z)}\right) \\ &= \sum_{x} \mathbb{P}(X = x) \sum_{z} \mathbb{P}\left(Z = z \,|\, X = x\right) \log\left(\frac{\mathbb{P}\left(Z = z \,|\, X = x\right))}{\mathbb{P}(Z = z)}\right). \end{aligned}$$

Similarly, one can define the mutual information for continuous variables. In particular, we are interested in the case where X is discrete and Y is continuous, denote by $\mu_{\mathcal{Y}}$ the marginal of (X, Y) over Y and by $\mu|_x$ the conditional $(Y \,|\, X = x)$.

$$I(X, Y) = \sum_{x} \mathbb{P}(X = x) \int_{y} \mu|_x(\mathrm{d}y) \log\left(\frac{\mu|_x(\mathrm{d}y)}{\mu(\mathrm{d}y)}\right).$$

Let us show the following version of the data processing inequality

$$I(X, \hat{X}(Y)) \leq I(X, Y). \tag{10.23}$$

To do so, we will use the conditional independence of X and $\hat{X}$ given Y, which leads to

$$\begin{aligned} \mathbb{P}\left(X = x \,\middle|\, \hat{X} = x'\right) &= \int \mathbb{P}\left(X = x \,|\, Y = \mathrm{d}y\right) \mathbb{P}\left(Y = \mathrm{d}y \,\middle|\, \hat{X} = z\right) \\ &= \int \frac{\mathbb{P}(X = x)\mu|_x(\mathrm{d}y)}{\mu(\mathrm{d}y)} \mathbb{P}\left(Y = \mathrm{d}y \,\middle|\, \hat{X} = z\right). \end{aligned}$$

Hence, using Jensen inequality,

$$\begin{aligned} I(X, \hat{X}) &= H(X) - H\left(X \,\middle|\, \hat{X}\right) \\ &= H(X) + \sum_{z} \mathbb{P}(\hat{X} = z) \sum_{x} \mathbb{P}(X = x) \log(\mathbb{P}\left(X = x \,\middle|\, \hat{X} = z\right)) \\ &= H(X) + \sum_{z} \mathbb{P}(\hat{X} = z) \sum_{x} \mathbb{P}(X = x) \log\left(\int \frac{\mathbb{P}(X = x)\mu|_x(\mathrm{d}y)}{\mu(\mathrm{d}y)} \mathbb{P}\left(Y = \mathrm{d}y \,\middle|\, \hat{X} = z\right)\right) \\ &\leq H(X) + \sum_{z} \mathbb{P}(\hat{X} = z) \sum_{x} \mathbb{P}(X = x) \int \mathbb{P}\left(Y = \mathrm{d}y \,\middle|\, \hat{X} = z\right) \log\left(\frac{\mathbb{P}(X = x)\mu|_x(\mathrm{d}y)}{\mu(\mathrm{d}y)}\right) \\ &= H(X) + \sum_{x} \mathbb{P}(X = x) \int \mu(\mathrm{d}y) \log\left(\frac{\mathbb{P}(X = x)\mu|_x(\mathrm{d}y)}{\mu(\mathrm{d}y)}\right) \\ &= \sum_{x} \mathbb{P}(X = x) \left(\int \mu(\mathrm{d}y) \log\left(\frac{\mathbb{P}(X = x)\mu|_x(\mathrm{d}y)}{\mu(\mathrm{d}y)}\right) - \log(P(X = x)\right) \\ &= \sum_{x} \mathbb{P}(X = x) \int \mu(\mathrm{d}y) \log\left(\frac{\mu|_x(\mathrm{d}y)}{\mu(\mathrm{d}y)}\right) \\ &= I(X, Y). \end{aligned}$$

We continue by computing the value of $I(X, Y)$, by definition and using Jensen inequality, we get

$$\begin{aligned} I(X, Y) &= \frac{1}{N} \sum_{i \in [N]} \int_{\mathcal{D}_n \sim \rho_i^{\otimes n}} \rho_i^{\otimes n}(\mathrm{d}\mathcal{D}_n) \log\left(\frac{\rho_i^{\otimes n}(\mathrm{d}\mathcal{D}_n)}{\frac{1}{N}\sum_{j \in [N]} \rho_j^{\otimes n}(\mathrm{d}\mathcal{D}_n)}\right) \\ &\leq \frac{1}{N} \sum_{i \in [N]} \int_{\mathcal{D}_n \sim \rho_i^{\otimes n}} \rho_i^{\otimes n}(\mathrm{d}\mathcal{D}_n) \frac{1}{N} \sum_{j \in [N]} \log\left(\frac{\rho_i^{\otimes n}(\mathrm{d}\mathcal{D}_n)}{\rho_j^{\otimes n}(\mathrm{d}\mathcal{D}_n)}\right) = \frac{1}{N^2} \sum_{i,j \in [N]} K\left(\rho_i^{\otimes n} \,\middle|\, \rho_j^{\otimes n}\right). \end{aligned}$$

We conclude from the fact that for p and q two distributions on a space $\mathcal{Z}$, we have

$$\begin{aligned} K\left(p^{\otimes n} \,||\, q^{\otimes n}\right) &= \int_{\mathcal{Z}^n} -\log\left(\frac{\mathrm{d}p^{\otimes n}(z_1,\cdots,z_n)}{\mathrm{d}q^{\otimes n}(z_1,\cdots,z_n)}\right) q^{\otimes n}(\mathrm{d}z_1,\cdots,\mathrm{d}z_n) \\ &= \int_{\mathcal{Z}^n} -\log\left(\frac{\prod_{i\leq n}\mathrm{d}p(z_i)}{\prod_{i\leq n}\mathrm{d}q(z_i)}\right) q^{\otimes n}(\mathrm{d}z_1,\cdots,\mathrm{d}z_n) \\ &= \sum_{i\leq n}\int_{\mathcal{Z}^n} -\log\left(\frac{\mathrm{d}p(z_i)}{\mathrm{d}q(z_i)}\right) q^{\otimes n}(\mathrm{d}z_1,\cdots,\mathrm{d}z_n) \\ &= \sum_{i\leq n}\int_{\mathcal{Z}} -\log\left(\frac{\mathrm{d}p(z_i)}{\mathrm{d}q(z_i)}\right) q(\mathrm{d}z_i) = nK\left(p \,||\, q\right). \end{aligned}$$

This explains the result. □

Let us assemble all the results proven thus far. In order to reduce our excess risk to a quadratic metric, we have assumed that the conditional distribution $\rho_i|_x$ to be Gaussian noise. In order to integrate this constraint into the precedent derivations, we leverage the following lemma.

Lemma 101 (Kullback-Leibler divergence with Gaussian noise)**.** *If ρ_i and ρ_j are two different distributions on $\mathcal{X}\times\mathcal{Y}$ such that there marginal over $\mathcal{X}$ are equal and the conditional distributions $(Y \mid X = x)$ are respectively equal to $\mathcal{N}(f_i(x), \sigma I_m)$ and $\mathcal{N}(f_j(x), \sigma I_m)$, then*

$$K\left(\rho_i \,||\, \rho_j\right) = \frac{1}{2\sigma^2}\left\|f_i - f_j\right\|^2_{L^2(\rho_{\mathcal{X}})}.$$

Proof. We proceed with

$$\begin{aligned} K\left(\rho_i \,||\, \rho_j\right) &= \int_{\mathcal{X}} \mathbb{E}_{Y\sim\mathcal{N}(f_j(x),\sigma I_m)}\left[\frac{\|Y - f_i(x)\|^2 - \left\|Y - f_j(x)\right\|^2}{2\sigma^2}\right]\rho_j(\mathrm{d}x) \\ &= \int_{\mathcal{X}} \mathbb{E}_{Y\sim\mathcal{N}\left(\frac{f_j(x)-f_i(x)}{\sqrt{2}\sigma}, I_m\right)}\left[\|Y\|^2\right] - \mathbb{E}_{Y\sim\mathcal{N}(0,I_m)}\left[\|Y\|^2\right]\rho_j(\mathrm{d}x) \\ &= \int_{\mathcal{X}}\left(m + \frac{\left\|f_j(x) - f_i(x)\right\|^2}{2\sigma^2} - m\right)\rho_j(\mathrm{d}x) = \frac{\left\|f_j - f_i\right\|^2_{L^2(\rho_{\mathcal{X}})}}{2\sigma^2}, \end{aligned}$$

where we have used the fact that the mean of a non-central χ-square variable of parameter (m,μ^2) is $m+\mu^2$. One could also develop the first two squared norms and use the fact that for any vector $u\in\mathbb{R}^m$, $\mathbb{E}[\langle Y - f_i(x), u\rangle] = 0$ to get the result. □

Combining the different results leads to the following proposition.

Lemma 102. *Under Assumption 20 with $\mathcal{F} = \{x\in\mathcal{X}\to\theta\varphi(x)\in\mathcal{Y} \mid \|\theta\|\leq M\}$ and φ bounded by κ, for any family $(f_i)_{i\leq N_\varepsilon}\in\mathcal{F}^N$ and any $\sigma > 2M\kappa\sigma_m$*

$$\begin{aligned} &\inf_{\mathcal{A}}\sup_{\rho}\mathbb{E}_{\mathcal{D}_n\sim\rho^{\otimes n}}\left[\mathcal{R}(f_{\mathcal{A}(\mathcal{D}_n)};\rho)\right] - \mathcal{R}^*(\rho) \\ &\qquad\geq \frac{\min_{i,j\in[N]}\left\|f_i - f_j\right\|^2_{L^2(\rho_{\mathcal{X}})}}{16(m+2)^{1/2}\sigma}\left(1 - \frac{\log(2)}{\log(N)} - \frac{n\max_{i,j\in[N]}\left\|f_i - f_j\right\|^2_{L^2(\rho_{\mathcal{X}})}}{2\sigma^2\log(N)}\right), \end{aligned}$$

for any algorithm $\mathcal{A}$ that maps a dataset $\mathcal{D}_n\in(\mathcal{X}\times\mathcal{Y})^n$ to a parameter $\theta\in\Theta$.

Covering number for linear model

We are left with finding a good packing of the space induced by Assumption 20. To do so, we shall recall some property of reproducing kernel methods.

Lemma 103 (Linear models are ellipsoids)**.** *For $\mathcal{H}$ a separable Hilbert space and $\varphi : \mathcal{X} \to \mathcal{H}$ bounded, the class of functions $\mathcal{F} = \{x \in \mathcal{X} \to \theta\varphi(x) \in \mathcal{Y} \,|\, \|\theta\| \leq M\}$ can be characterized by*

$$\mathcal{F} = \left\{ f : \mathcal{X} \to \mathcal{Y} \,\middle|\, \left\| K^{-1/2} f \right\|_{L^2(\rho_{\mathcal{X}})} \leq M \right\}, \tag{10.24}$$

where $\rho_{\mathcal{X}}$ is any distribution on $\mathcal{X}$ and K is the operator on $L^2(\rho_{\mathcal{X}})$ that map f to

$$Kf(x') = \int_{x \in \mathcal{X}} \langle \varphi(x), \varphi(x') \rangle \, f(x) \rho_{\mathcal{X}}(\mathrm{d}x),$$

whose image is assumed to be dense in L^2.

Proof. This follows for isometry between elements in $\mathcal{H}$ and elements in L^2. More precisely, let us define

$$\begin{array}{rccc} S: & \mathcal{Y} \otimes \mathcal{H} & \to & L^2(\mathcal{X}, \mathcal{Y}, \rho_{\mathcal{X}}) \\ & \theta & \to & x \to \theta\varphi(x). \end{array}$$

The adjoint of S is characterized by

$$\begin{array}{rccc} S^*: & L^2(\mathcal{X}, \mathcal{Y}, \rho_{\mathcal{X}}) & \to & \mathcal{Y} \otimes \mathcal{H} \\ & f & \to & \mathbb{E}[f(x) \otimes \varphi(X)], \end{array}$$

which follows from the fact that for $\theta \in \mathcal{Y} \otimes \mathcal{H}$, $f \in L^2$ we have

$$\begin{aligned} \langle \theta, S^* f \rangle_{\mathcal{Y} \otimes \mathcal{H}} = \langle S\theta, f \rangle_{L^2} &= \sum_{i=1}^{m} \int_{\mathcal{X}} f_i(x) \, \langle \theta_i, \varphi(x) \rangle_{\mathcal{H}} \, \rho_{\mathcal{X}}(\mathrm{d}x) \\ &= \sum_{i=1}^{m} \langle \theta_i, \mathbb{E}[f_i(X)\varphi(X)] \rangle_{\mathcal{H}} = \langle \theta, \mathbb{E}[f(X) \otimes \varphi(X)] \rangle_{\mathcal{Y} \otimes \mathcal{H}}. \end{aligned}$$

When SS^* is compact and dense in L^2, we have

$$\|\theta\|_{\mathcal{Y} \otimes \mathcal{H}} = \left\| (SS^*)^{-1/2} S\theta \right\|_{L^2(\rho_{\mathcal{X}})}.$$

The compactness allows considering spectral decomposition hence fractional powers. We continue by observing that $SS^* = K$, which follows from

$$(SS^* f)(x') = (S\,\mathbb{E}[f(X) \otimes \varphi(X)])(x') = \mathbb{E}[f(X) \otimes \varphi(X)]\varphi(x') = \mathbb{E}[\langle \varphi(X), \varphi(x') \rangle \, f(X)].$$

The compactness of K derives from the fact that

$$\|Kf(x')\|^2 = \|\mathbb{E}[\langle \varphi(X), \varphi(x') \rangle \, f(X)]\|^2 \leq \mathbb{E}[\|\langle \varphi(X), \varphi(x') \rangle \, f(X)\|^2] \leq \kappa^2 \, \|f\|_{L^2}^2.$$

Hence, $\|K\|_{\mathrm{op}} \leq \kappa^2$. Indeed, it is not hard to prove that the trace of K is bounded by $m\kappa^2$, hence K is not only compact but trace-class. □

It should be noted that the condition on K being dense in $L^2(\rho_{\mathcal{X}})$ is not restrictive, as indeed all the problem is only seen through the lens of φ and $\rho_{\mathcal{X}}$: one can replace $\mathcal{X}$ by $\operatorname{supp} \rho_{\mathcal{X}}$ and $L^2(\rho_{\mathcal{X}})$ by the closure of the range of K in $L^2(\rho_{\mathcal{X}})$ without modifying nor the analysis, nor the original problem.

We should study packing in the ellipsoid $\mathcal{F} = \left\{ f \in L^2 \,\middle|\, \|K^{-1/2} f\|_{L^2(\rho_{\mathcal{X}})} \leq M \right\}$. It is useful to split the ellipsoid between a projection on a finite dimensional space that is isomorphic to the Euclidean space $\mathbb{R}^k$ and on a residual space R where the energies $(\|f|_R\|_{L^2(\rho_{\mathcal{X}})})_{f \in \mathcal{F}}^2$ are uniformly small. We begin with the following packing lemma, sometimes referred to as Gilbert-Varshamov bound (Gilbert, 1952; Varshamov, 1957) which corresponds to a more generic result in coding theory.

Lemma 104 (ℓ_2^2-packing of the hypercube)**.** *For any $k \in \mathbb{N}^*$, there exists a $k/4$-packing of the hypercube $\{0,1\}^k$, with respect to Hamming distance, of cardinality $N = \exp(k/8)$.*

Proof. Let us consider $\varepsilon > 0$, and a maximal ε-packing $(x_i)_{i\leq N}$ of the hypercube with respect to the distance $d(x,y) = \sum_{i\in[k]} \mathbf{1}_{x_i\neq y_i} = \|x-y\|_1 = \|x-y\|_2^2$. By maximality, we have $\{0,1\}^k \subset \cup_{i\in[N]} B_d(x_i,\varepsilon)$, hence

$$2^k \leq N \left|\left\{x \in \{0,1\}^k \,\middle|\, \|x\|_1 \leq \varepsilon\right\}\right|.$$

This inequality can be rewritten with Z a binomial variable of parameter $(k, 1/2)$ as $1 \leq N\,\mathbb{P}(Z \leq \varepsilon)$. Using Hoeffding inequality (Hoeffding, 1963), when $\varepsilon = k/4$ we get

$$N^{-1} \leq \mathbb{P}(Z \leq k/4) = \mathbb{P}(Z - \mathbb{E}[Z] \leq k/4) \leq \exp\left(-\frac{2k^2}{4^2 k}\right) = \exp\left(-k/8\right).$$

This is the desired result. □

Lemma 105 (Packing of infinite-dimensional ellipsoids)**.** *Let $\mathcal{F}$ be the function in $L^2(\rho_{\mathcal{X}})$ such that $\left\|K^{-1/2} f\right\|_{L^2(\rho_{\mathcal{X}})} \leq M$ for K a compact operator and M any positive number. For any $k \in \mathbb{N}^*$, it is possible to find a family of $N \geq \exp(k/8)$ elements $(f_i)_{i\in[N]}$ in $\mathcal{F}$ such that for any $i \neq j$,*

$$\frac{kM^2}{\sum_{i\leq k} \lambda_i^{-1}} \leq \left\|f_i - f_j\right\|_{L^2(\rho_{\mathcal{X}})}^2 \leq \frac{4kM^2}{\sum_{i\leq k} \lambda_i^{-1}}, \tag{10.25}$$

where $(\lambda_i)_{i\in\mathbb{N}}$ are the ordered (with repetition) eigenvalues of K.

Proof. Let us denote by $(\lambda_i)_{i\in\mathbb{N}}$ the eigenvalues of K and $(u_i)_{i\in\mathbb{N}}$ in L^2 the associated eigenvectors. Consider $(a_s)_{s\in[N]}$ a k-packing of the hypercube $\{-1,1\}^k$ for $N \geq \exp(k/8)$ with respect to the ℓ_2^2 quasi-distance and define for any $a \in \{a_s\}$

$$f_a = \frac{M}{c} \sum_{s=1}^{k} a_i u_i,$$

with $c^2 = \sum_{i=1}^k \lambda_i^{-1}$. We verify that

$$\left\|K^{-1/2} f_a\right\|_{L^2}^2 = \frac{M^2}{c^2} \sum_{i=1}^{k} \lambda_i^{-1} = M^2.$$

$$\|f_a - f_b\|_{L^2}^2 = \frac{M^2}{c^2} \sum_{i=1}^{k} |a_i - b_i|^2 = \frac{M^2}{c^2} \|a_i - b_i\|_2^2 \in \frac{M^2}{c^2} \cdot [k, 4k].$$

This is the object of the lemma. □

So far, we have proven the following lower bound.

Lemma 106. *Under Assumption 20 with $\mathcal{F} = \{x \in \mathcal{X} \to \theta\varphi(x) \in \mathcal{Y} \mid \|\theta\| \leq M\}$ and φ bounded by κ, for any family $(f_i)_{i\leq N_\varepsilon} \in \mathcal{F}^N$ and any $\sigma > 2M\kappa\sigma_m$ and $km > 10$,*

$$\inf_{\mathcal{A}} \sup_{\rho} \mathbb{E}_{\mathcal{D}_n\sim\rho^{\otimes n}}[\mathcal{R}(f_{\mathcal{A}(\mathcal{D}_n)};\rho)] - \mathcal{R}^*(\rho) \geq \frac{1}{128} \min\left\{\frac{M^2}{\sigma m^{1/2} \sum_{i\leq k} (k\lambda_i)^{-1}}, \frac{\sigma k m^{1/2}}{32n}\right\},$$

for any algorithm $\mathcal{A}$ that maps a dataset $\mathcal{D}_n \in (\mathcal{X}\times\mathcal{Y})^n$ to a parameter $\theta \in \Theta$, and where (λ_i) are the ordered eigenvalue of the operator K on $L^2(\mathcal{X}, \mathbb{R}, \rho_{\mathcal{X}})$ that maps any function f to the function Kf defines for $x' \in \mathcal{X}$ as

$$(Kf)(x') = \int_{x\in\mathcal{X}} \langle\varphi(x), \varphi(x')\rangle\, f(x)\rho_{\mathcal{X}}(\mathrm{d}x).$$

In particular, when $\lambda_i = \kappa^2 i^{-a}/\zeta(\alpha)$, where ζ denotes the Riemann zeta function, we get the following bounds. If we optimize with respect to σ, there exists $n_\alpha \in \mathbb{N}$ such that for any $n > n_\alpha$.

$$\inf_{\mathcal{A}} \sup_{\rho} \mathbb{E}_{\mathcal{D}_n\sim\rho^{\otimes n}}[\mathcal{R}(f_{\mathcal{A}(\mathcal{D}_n)};\rho)] - \mathcal{R}^*(\rho) \geq \frac{M\kappa}{725\zeta(\alpha)^{1/2} n^{1/2}}. \tag{10.26}$$

If we fix $\sigma = \beta M\kappa$ with $\beta \geq 2$, and we optimize with respect to k, there exists a constant c_β and an integer n_0 such that for $n > n_0$ we have

$$\inf_{\mathcal{A}} \sup_{\rho} \mathbb{E}_{\mathcal{D}_n\sim\rho^{\otimes n}}[\mathcal{R}(f_{\mathcal{A}(\mathcal{D}_n)};\rho)] - \mathcal{R}^*(\rho) \geq \frac{M\kappa c_\beta}{\zeta(\alpha)^{\frac{1}{1+\alpha}} n^{\frac{\alpha}{\alpha+1}}}. \tag{10.27}$$

Proof. Reusing Lemma 102, with the same notations, we have the lower bound in

$$\frac{\min \left\|f_i - f_j\right\|^2}{16\sigma(m+2)^{1/2}} \left(1 - \frac{\log(2)}{\log(N)} - \frac{n \max \left\|f_i - f_j\right\|^2}{2\sigma^2 \log(N)}\right).$$

Let K and $K_{\mathcal{Y}}$ be the self-adjoint operators on $L^2(\mathcal{X}, \mathbb{R}, \rho_{\mathcal{X}})$ and $L^2(\mathcal{X}, \mathcal{Y}, \rho_{\mathcal{X}})$ respectively, both defined through the formula

$$(Kf)(x') = \int_{x \in \mathcal{X}} \langle \varphi(x), \varphi(x') \rangle \, f(x) \rho_{\mathcal{X}}(\mathrm{d}x).$$

When K is compact, it admits an eigenvalue decomposition $K = \sum_{i \in \mathbb{N}} \lambda_i u_i \otimes u_i$ where the equality as to be understood as the convergence of operator with respect to the operator norm based on the L^2-topology. It follows from the product structure of $L^2(\mathcal{X}, \mathcal{Y}, \rho_{\mathcal{X}}) \simeq L^2(\mathcal{X}, \mathbb{R}, \rho_{\mathcal{X}})^m$ that $K_{\mathcal{Y}} = \sum_{i \in \mathbb{R}, j \in [m]} \sum_{i \in \mathbb{N}, j \in [m]} \lambda_i (e_i \otimes y_j) \otimes (e_i \otimes u_j)$ with (e_j) the canonical basis of $\mathcal{Y} = \mathbb{R}^m$. As a consequence, if $(\lambda_i)_{i \in \mathbb{N}}$ are the ordered eigenvalues of K then $(\lambda_{\lfloor i/m \rfloor})$ are the ordered eigenvalues of $K_{\mathcal{Y}}$. Hence, with Lemmas 103 and 105, it is possible to find $N = \exp(km/8)$ functions in $\mathcal{F}$ such that

$$\frac{kmM^2}{m \sum_{i \leq k} \lambda_i^{-1}} \leq \left\|f_i - f_j\right\|^2_{L^2(\rho_{\mathcal{X}})} \leq \frac{4kmM^2}{m \sum_{i \leq k} \lambda_i^{-1}}.$$

If we multiply those functions by $\eta \in [0, 1]$ we get a lower bound in

$$\frac{\eta^2 M^2}{16\sigma(m+2)^{1/2} \sum_{i \leq k} (k\lambda_i)^{-1}} \left(1 - \frac{8\log(2)}{km} - \frac{16M^2 n\eta^2}{\sigma^2 km \sum_{i \leq k} (k\lambda_i)^{-1}}\right).$$

Making sure that the last two terms are smaller than one fourth and one half respectively we get the following conditions on k and η, with $\Lambda_k = \sum_{i \leq k} (k\lambda_i)^{-1}$,

$$km \geq 32\log(2), \qquad 32M^2 n\eta^2 \leq \sigma^2 km\Lambda_k.$$

Using the fact that $\eta < 1$, the lower bound becomes

$$\frac{M^2}{128\sigma m^{1/2}\Lambda_k} \min\left\{1, \frac{\sigma^2 km\Lambda_k}{32M^2 n}\right\} = \frac{1}{128} \min\left\{\frac{M^2}{\sigma m^{1/2}\Lambda_k}, \frac{\sigma km^{1/2}}{32n}\right\},$$

as long as $km > 10$. When $\lambda_i^{-1} = i^\alpha \zeta(\alpha)/\kappa^2$, since $\Lambda_k \leq \lambda_k^{-1}$, we simplify the last expression as

$$\frac{1}{128} \min\left\{\frac{M^2\kappa^2}{\sigma m^{1/2} k^\alpha \zeta(\alpha)}, \frac{\sigma km^{1/2}}{32n}\right\}.$$

Optimizing with respect to σ leads to

$$\sigma^2 = \frac{32nM^2\kappa^2}{mk^{1+\alpha}\zeta(\alpha)} \geq 4M^2\kappa^2\sigma_m.$$

This gives

$$n_{\alpha,m} = m\zeta(\alpha)\sigma_m^2/8.$$

The dependency of n_α to m can be removed since any problem with $\mathcal{Y} = \mathbb{R}^m$ can be cast as a problem in $\mathbb{R}^{m+1}$ by adding a spurious coordinate. Taking $k = 1$ and $m = 10$ leads to the result stated in the lemma. When $n < n_\alpha$, one can artificially multiply the bound by $n_\alpha^{1/2}$, since an optimal algorithm can not do better with fewer data. After checking that one can take $\sigma_1 \geq 1$, this leads to a bound in

$$\frac{M\kappa}{2048n^{1/2}}.$$

Optimizing with respect to k leads to $k^{\alpha+1} = 32M^2\kappa^2 n/(\sigma^2 m\zeta(\alpha))$ and a bound in

$$\frac{(\sigma m^{1/2})^{\frac{\alpha-1}{\alpha+1}} (M\kappa)^{\frac{2}{\alpha+1}}}{128(32n)^{\frac{\alpha}{\alpha+1}} \zeta(\alpha)^{\frac{1}{\alpha+1}}}.$$

The condition $k > \min\left\{10m^{-1}, 1\right\}$ and $\sigma \geq 2M\kappa\sigma_m$ translates into the condition

$$4M^2\kappa^2\sigma_m^2 \leq \sigma^2 \leq \frac{32M^2\kappa^2 n}{m\zeta(\alpha)} \min\left\{1, \frac{m^{1+\alpha}}{10^{1+\alpha}}\right\}.$$

We deduce that $\sigma_m = O(m^{-1/2})$, otherwise we would not respect the upper bound derived with Rademacher complexity (or have made a mistake somewhere). Once again we can remove the dependency to m. Considering $\sigma = \beta M\kappa$ leads to the result stated in the lemma. □

Controlling eigenvalues decay

Based on Lemma 106, in order to prove Theorem 21, we only need to show that there exists a mapping φ, an input space $\mathcal{X}$ and a distribution $\rho_{\mathcal{X}}$ such that the integral operator K introduced in the lemma verifies the assumption on its eigenvalues. Notice that we show in the proof of Lemma 106 that the universal constant c_3 can be taken as $c_3 = 2^{-11}$.

To proceed, let us consider any infinite dimensional Hilbert space $\mathcal{H}$ with a basis $(e_i)_{i\in\mathbb{N}}$, $\mathcal{X} = \mathbb{N}$ and $\varphi : \mathbb{N} \to \mathcal{H}; i \to \kappa e_i$. For $a : \mathbb{N} \to \mathbb{R}$ we have

$$(Ka)(i) = \sum_{j\in\mathbb{N}} \langle\varphi(i), \varphi(j)\rangle\, a(j)\rho(j) = \kappa^2 a(i)\rho(i).$$

Hence, the eigenvalues of K are $(\kappa^2\rho_{\mathcal{X}}(i))_{i\leq n}$. It suffices to consider $\rho_{\mathcal{X}}(i) = i^{-\alpha}/\zeta(\alpha)$ to conclude.

The eigenvalue decay in $O(i^{-\alpha})$ can also be witnessed in many regression problems. One way to build those cases is to turn a sequence of non-negative real values into a one-periodic function h from $\mathbb{R}^d$ to $\mathbb{R}$ thanks to the inverse Fourier transform. Using Bochner (1933), one can construct a map φ such that the convolution operator linked with h corresponds to the operator K. When ρ is uniform on $[0,1]^d$, diagonalizing this convolution operator with the Fourier functions and using the property in Lemma 103 shows that the class of functions $\mathcal{F}$ are akin to Sobolev spaces. Similar behavior can be proven when $\mathcal{X} = \mathbb{R}^d$ and $\rho_{\mathcal{X}}$ is absolutely continuous with respect to the Lebesgue measure and has bounded density (Widom, 1963). We refer the curious reader to Scholkopf and Smola (2001) or Bach (2023) for details.

10.B Unbiased weakly supervised stochastic gradients

In this section, we provide a generic scheme to acquire unbiased weakly supervised stochastic gradients, as well as specifications of the formula given in the main text for least-squares and median regression.

10.B.1 Generic implementation

Suppose that Θ is finite dimensional, or that it can be approximated by a finite dimensional space without too much approximation error. For example, in the realm of scalar-valued kernel methods, it is usual to consider either the random finite dimensional space $\mathrm{Span}\,\{\varphi(x_i)\}_{i\leq n}$ for (x_i) the data points, or the finite dimension space linked to the first eigenspaces of the operator $\mathbb{E}[\varphi(X)\otimes\varphi(X)]$. In the context of neural networks, the parameter space is always finite-dimensional.

Suppose also that, given θ, we know an upper bound M_θ on the amplitude of $\nabla_\theta\ell(f_\theta(x), y)$, or that we know how to handle clipped gradients at amplitude M_θ for SGD. Then, similarly to the least-squares method proposed in the main text, we can access weakly supervised gradient through the formula

$$\nabla_\theta\ell(f_\theta(x), y) = \frac{2M_\theta(|\Theta|^2 + 4\,|\Theta| + 3)}{\pi^{3/2}}\, \mathbb{E}_{U\sim\mathcal{U}(B_\Theta), V\sim\mathcal{U}([0,M_\theta])}[\mathbf{1}_{y\in(z\to\langle U, \nabla_\theta\ell(f_\theta(x),z)\rangle)^{-1}([V,\infty))}U],$$

where B_Θ is the unit ball of Θ.

This scheme is really generic, and we do not advocate for it in practice as one may hope to leverage specific structure of the loss function and the parametric model in a more efficient way. This formula is rather a proof of concept to illustrate that our technique can be applied generically, and is not specific to least-squares or median regression.

10.B.2 Specific implementations

Let us prove the two formulas to get stochastic gradients for both least-squares and median regression. We begin with median regression. Consider $z \in \mathbb{S}^{m-1}$, and let us denote

$$x = \mathbb{E}_U[\mathrm{sign}(\langle z, U\rangle)U].$$

The direction $x/\|x\| \in \mathbb{S}^{m-1}$ is characterized by the argmax over the sphere of the linear form

$$y \to \langle \mathbb{E}_U[\mathrm{sign}(\langle z, U\rangle)U], y\rangle = \mathbb{E}_U[\mathrm{sign}(\langle z, U\rangle)\,\langle U, y\rangle].$$

This linear form has a unique maximizer on $\mathbb{S}^{m-1}$ and by invariance by symmetry over the axis z, this maximizer is aligned with z, hence $x = c_x \cdot z$. We compute the amplitude with the formula, because z is a unit vector

$$c_x = \langle x, z\rangle = \mathbb{E}_U[\mathrm{sign}(\langle z, U\rangle)\,\langle U, z\rangle].$$

By invariance by rotation of both the uniform distribution and the scalar product, c_x is actually a constant, it is equal to its value $c_2 = c_{e_1}$.

The same type of reasoning applies for the least-squares case. Consider $z \in \mathbb{R}^m$, and denote

$$x = \mathbb{E}_{U,V}\left[\mathbf{1}_{\langle z,U\rangle \geq V} \cdot U\right].$$

For the same reasons as before $x = c_x \cdot u$ for $u = z/\|z\|$, and c_x verifies

$$\begin{aligned} c_x = \langle x, u\rangle &= \mathbb{E}_{U,V}[\mathbf{1}_{\langle z,U\rangle \geq V}\,\langle U, u\rangle] = \mathbb{E}_U[\mathbb{E}_V[\mathbf{1}_{\langle z,U\rangle \geq V}]\,\langle U, u\rangle] \\ &= \mathbb{E}_U[\mathbf{1}_{\langle z,U\rangle > 0}\frac{\langle z, U\rangle}{M}\,\langle U, u\rangle] = \frac{\|z\|}{M}\,\mathbb{E}_U[\mathbf{1}_{\langle u,U\rangle > 0}\,\langle U, u\rangle^2]. \end{aligned}$$

Hence,

$$x = \frac{1}{M}\,\mathbb{E}_U[\mathbf{1}_{\langle u,U\rangle > 0}\,\langle U, u\rangle^2] \cdot z = c_1 \cdot z.$$

This explains the formula for least-squares.

Lemma 107 (Constant for the uniform strategy)**.** *Under the uniform distribution on the sphere*

$$c_2 = \mathbb{E}_{u\sim\mathbb{S}^{m-1}}\left[|\langle u, e_1\rangle|\right] = \frac{\sqrt{\pi}\,\Gamma(\frac{m-1}{2})}{m\,\Gamma(\frac{m}{2})} \geq \frac{\sqrt{2\pi}}{m^{3/2}}. \tag{10.28}$$

Proof. Let us compute $c_2 = \mathbb{E}_{u\sim\mathbb{S}^{m-1}}\left[|\langle u, e_1\rangle|\right]$. This constant can be written explicitly as

$$c_2 = \frac{\int_{x\in\mathbb{S}^{m-1}} |x_1|\,\mathrm{d}x}{\int_{x\in\mathbb{S}^{m-1}} \mathrm{d}x}.$$

Remark that for any function $f : \mathbb{R} \to \mathbb{R}$, we have

$$\int_{\mathbb{S}^{m-1}} f(x_1)\,\mathrm{d}x = \int_{x_1\in[-1,1]} f(x_1)\,\mathrm{d}x_1 \int_{\tilde{x}\in\sqrt{1-x_1^2}\cdot\mathbb{S}^{m-2}} \mathrm{d}\tilde{x} = \int_{x_1\in[-1,1]} f(x_1)(1-x_1^2)^{\frac{m-2}{2}}\,\mathrm{d}x_1 \int_{\tilde{x}\in\mathbb{S}^{m-2}} \mathrm{d}\tilde{x}.$$

By denoting S_m the surface of the m-sphere, the last integral is nothing but S_{m-2}. By setting $f(x) = 1$, we can retrieve by recurrence the expression of S_m. In our case, $f(x) = |x|$, so we compute, with $u = 1 - x^2$

$$\int_{x_1\in[-1,1]} |x_1|\,(1-x_1^2)^{\frac{m-2}{2}}\,\mathrm{d}x_1 = 2\int_{x_1\in[0,1]} x_1(1-x_1^2)^{\frac{m-2}{2}}\,\mathrm{d}x_1 = \int_{u=0}^{1} u^{\frac{m-2}{2}}\,\mathrm{d}u = \frac{1}{m}.$$

This leads to

$$c_2 = \frac{S_{m-2}}{mS_{m-1}} = \frac{\sqrt{\pi}\,\Gamma(\frac{m-1}{2})}{m\,\Gamma(\frac{m}{2})}.$$

The ratio S_{m-2}/S_{m-1} can be expressed with the integral corresponding to $f = 1$, but it is common knowledge that $S_{m-1} = {2\pi^{m/2}}/{\Gamma(m/2)}$. □

Lemma 108 (Constant for least-squares)**.** *Under the uniform distributions on* $[0, M]$ *and the sphere*

$$c_1 = \mathbb{E}_{y\sim[0,M]}\,\mathbb{E}_{u\sim\mathbb{S}^{m-1}}\left[\mathbf{1}_{\langle u,e_1\rangle>v}\,\langle u, e_1\rangle\right] = \frac{\pi^{3/2}}{M(m^2+4m+3)}. \tag{10.29}$$

Proof. Similarly to the previous case, this constant can be written explicitly as

$$c_1 = \frac{1}{2}\frac{\int_{y\in[0,M]}\int_{x\in\mathbb{S}^{m-1}}|x_1|\,\mathbf{1}_{|x_1|>y}\,\mathrm{d}y\,\mathrm{d}x}{M\int_{x\in\mathbb{S}^{m-1}}\mathrm{d}x} = \frac{\int_{x\in\mathbb{S}^{m-1}}x_1^2\,\mathrm{d}x}{2M\int_{x\in\mathbb{S}^{m-1}}\mathrm{d}x}.$$

We continue as before with

$$\int_{x_1\in[-1,1]}|x_1|^2\,(1-x_1^2)^{\frac{m-2}{2}}\,\mathrm{d}x_1 = 2\int_{x\in[0,1]}x^2(1-x^2)^{\frac{m-2}{2}}\,\mathrm{d}x = \frac{2\pi\Gamma(\frac{m}{2})}{4\Gamma(\frac{m+3}{2})}.$$

This leads to

$$c_1 = \frac{\pi\Gamma(\frac{m}{2})}{4M\Gamma(\frac{m+3}{2})}\cdot\frac{\sqrt{\pi}\,\Gamma(\frac{m-1}{2})}{\Gamma(\frac{m}{2})} = \frac{\pi^{3/2}\Gamma(\frac{m-1}{2})}{4M\Gamma(\frac{m+3}{2})} = \frac{\pi^{3/2}}{M(m^2+4m+3)}.$$

This is the result stated in the lemma. □

10.C Median surrogate

Let us begin this section by proving Proposition 91. This result is actually the integration over $x \in \mathcal{X}$ of a pointwise result, so let us fix $x \in \mathcal{X}$. Consider a probability distribution $p \in \Delta_{\mathcal{Y}}$ over $\mathcal{Y}$, and its median $\Theta^* \subset \mathbb{R}^{\mathcal{Y}}$ defined as the minimizer of $\mathcal{R}_S(\theta) = \mathbb{E}_p[\|\theta - e_Y\|]$. We will to prove that $\cup_{\theta\in\Theta^*}\arg\max_{y\in\mathcal{Y}}\theta_y = \arg\max_{y\in\mathcal{Y}}p(y)$.

Let us begin by the inclusion $\arg\max_{y\in\mathcal{Y}}p(y) \subset \cup_{\theta\in\Theta^*}\arg\max_{y\in\mathcal{Y}}\theta_y$. To do so, consider $\theta \in \mathbb{R}^{\mathcal{Y}}$ and $\sigma \in \mathfrak{S}_{\mathcal{Y}}$ the transposition of two elements y and z in $\mathcal{Y}$. Denote by $\theta_\sigma \in \mathbb{R}^{\mathcal{Y}}$, the vector such that $(\theta_\sigma)_{y'} = \theta_{\sigma(y')}$ for any $y' \in \mathcal{Y}$. We have

$$\begin{aligned}
\mathcal{R}_S(\theta) - \mathcal{R}_S(\theta_\sigma) &= \sum_{y'\in\mathcal{Y}}p(y')\left(\left\|\theta - e_{y'}\right\| - \left\|\theta_\sigma - e_{y'}\right\|\right)\\
&= \sum_{y'\in\mathcal{Y}}p(y')\left(\sqrt{\sum_{z'\in\mathcal{Y}}\theta_{z'}^2 + (1-\theta_{y'})^2 - \theta_{y'}^2} - \sqrt{\sum_{z'\in\mathcal{Y}}\theta_{\sigma(z')}^2 + (1-\theta_{\sigma(y')})^2 - \theta_{\sigma(y')}^2}\right)\\
&= (p(y)-p(z))\left(\sqrt{\sum_{z'\in\mathcal{Y}}\theta_{z'}^2 + 1 - 2\theta_y} - \sqrt{\sum_{z'\in\mathcal{Y}}\theta_{z'}^2 + 1 - 2\theta_z}\right).
\end{aligned}$$

Because, for any $a \in \mathbb{R}_+$, the function $x \to \sqrt{a-2x}$ is increasing, if $p(y) > p(z)$, then to minimize $\mathcal{R}$, we should make sure that $\theta_y \geq \theta_z$i. As a consequence, because of symmetry, the modes of p do correspond to argmax of $(\theta_y^*)y \in \mathcal{Y}$ for some $\theta^* \in \Theta^*$.

Let us now prove the second inclusion. To do so, suppose that $p(1) > p(2)$, and let us show that $\theta_1^* > \theta_2^*$. Let us parametrize $\theta_1 = a + \varepsilon$ and $\theta_2 = a - \varepsilon$ for a given a, and show that $\varepsilon = 0$ is not optimal in order to minimize the risk $\mathcal{R}_S$ seen as a function of ε. To do so, we can use the Taylor expansion of $\sqrt{1+x} = 1 + x/2$. Hence, with $A = \sum_{y>2}(\theta_y^*)^2$, retaking the last derivations

$$\begin{aligned}
\mathcal{R}_S(\varepsilon) &= p(1)\sqrt{(a+\varepsilon)^2 + (a-\varepsilon)^2 + A + 1 - 2(a+\varepsilon)}\\
&\quad + p(2)\sqrt{(a+\varepsilon)^2 + (a-\varepsilon)^2 + A + 1 - 2(a-\varepsilon)}\\
&\quad + \sum_{y>2}p(y)\sqrt{(a+\varepsilon)^2 + (a-\varepsilon)^2 + A + 1 - 2\theta_y^*}\\
&= p(1)\sqrt{2a^2 + 2\varepsilon^2 + A + 1 - 2a - 2\varepsilon}\\
&\quad + p(2)\sqrt{2a^2 + 2\varepsilon^2 + A + 1 - 2a + 2\varepsilon} + c + o(\varepsilon)\\
&= \tilde{c} + \frac{\varepsilon}{\sqrt{2a^2 + A + 1 - 2a}}(p(2) - p(1)) + o(\varepsilon).
\end{aligned}$$

This shows that taking $\theta_1^* = \theta_2^*$, that is $\varepsilon = 0$, is not optimal, hence we have the second inclusion, which ends the proof. Note that we have proven a much stronger result, we have shown that (θ_y) and $p(y)$ are order in the exact same fashion (with respect to the strict comparison $p(y) > p(z) \Rightarrow \theta_y^* > \theta_z^*$ for any $\theta^* \in \Theta^*$).

10.C.1 Discussion around the median surrogate.

The median surrogate have some nice properties for a surrogate method, in particular it does not fully characterize the distribution $p(y)$ in the sense that there is no one-to-one mapping from p to θ^*. For example, when $\mathcal{Y} = \{1, 2, 3\}$ if $p(y = e_1), p(y = e_2), p(y = e_3) \propto (1, 1, 2\cos(\pi/6))$, then the geometric median correspond to $\theta^* = e_3$. This differs from smooth surrogates, such as logistic regression or least-squares, that implicitly learn the full distribution p, which should be seen as a waste of resources. Non-smooth surrogates tend to exhibit faster rates of convergence (in terms of decrease of the original risk as a function of the number of samples) than smooth surrogates when rates are derived through calibration inequalities (Nowak-Vila, 2021). It would be nice to derive generic calibration inequality for the median surrogate for multiclass, and see how to derive a median surrogate for more structured problems such as ranking problems.

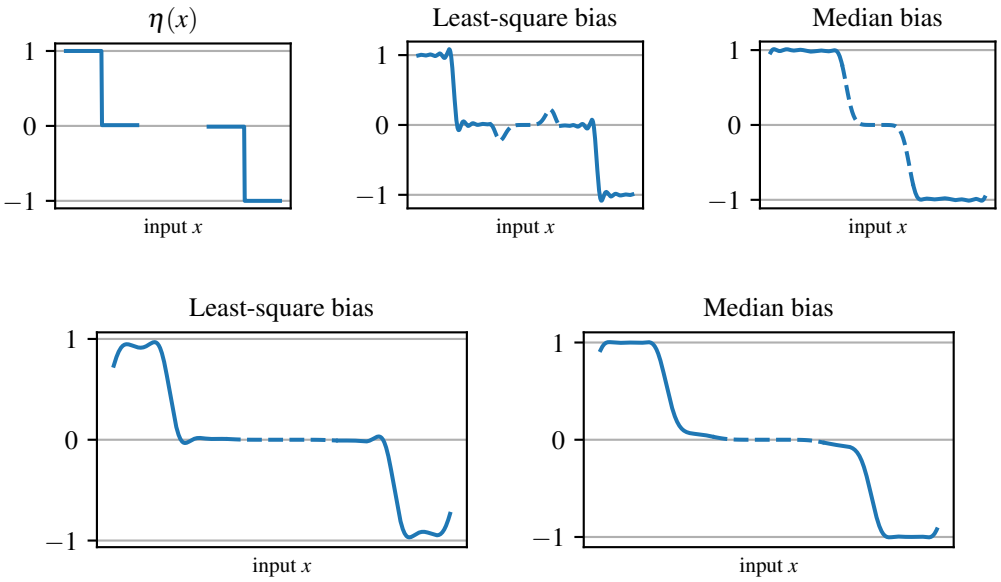

Figure 10.4: *Comparison of least-squares and absolute deviation with noise irregularity* for a classification problem specified by $\mathcal{X} = [0, 3]$, $\mathcal{Y} = \{-1, 1\}$ with X uniform on $[0, 1] \cup [2, 3]$ and $\eta(x) = \mathbb{E}\{Y \mid X = x\}$ specified on the left figure. The optimal classifier, with respect to the zero-one loss, $f^*(x) = \operatorname{sign} \eta$ takes value one on $[0, 1]$ and value minus one on $[2, 3]$. The regularized solution are defined as $\arg\min_g \mathbb{E}[\|\langle \varphi(X), \theta \rangle - Y\|^p] + \lambda \|\theta\|$ with $p = 2$ for least-squares (middle), and $p = 1$ for the median (right). They can be translated into classifiers with the decoding $f = \operatorname{sign} g$. In this figure, we choose φ implicitly through the Gaussian kernel $k(x, x') = \langle \varphi(x), \varphi(x') \rangle = \exp(-\|x - x'\|^2 / 2\sigma^2)$ with $\sigma = .1$ which explains the frequency of the observed oscillations, and choose $\lambda = 10^{-6}$ (top) and $\lambda = 10^{-2}$ (bottom). On the one hand, because the least-squares surrogate is trying to estimate η it suffers from its lack of regularity, leading to Gibbs phenomena that restricts it to be a perfect classifier. On the other hand, the absolute deviation is trying to approach the function f^* itself, and does not suffer from its lack of regularity. In this setting, if we approach the original classification problem by minimization of the surrogate empirical risks, and denote by g_n this minimizer and $f_n = \operatorname{sign} g_n$ its decoding, f_n obtained through median regression will converge exponentially fast toward f^*, while f_n obtained through least-squares will never converge to the solution f^*.

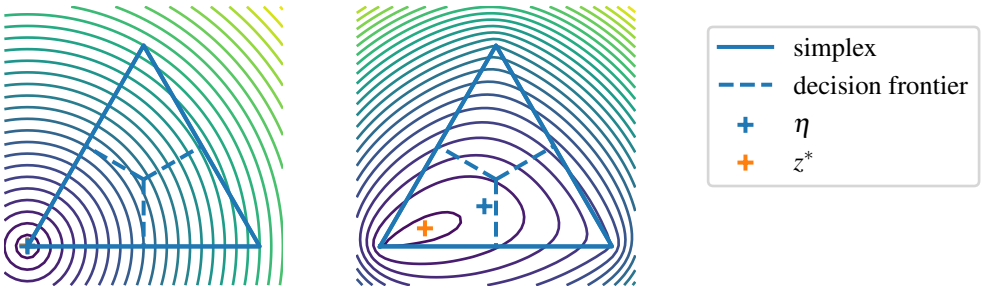

Figure 10.5: *Comparison of least-squares and median surrogate without context.* Consider a context-free classification problem that consists in estimating the mode of a distribution $p \in \Delta_{\mathcal{Y}}$, or equivalently the minimizer of the 0-1 loss. Such a problem can be visualized on the simplex Δ_m where $\mathcal{Y} = \{y_1, \cdots, y_m\} \simeq \{1, \cdots, m\}$ is mapped to the canonical basis $\{e_i\}_{i \in [m]} \in \mathbb{R}^m$. The figure illustrates the case $m = 3$. The least-squares and median surrogate methods can be understood as working in this simplex, estimating a quantity $z \in \Delta_{\mathcal{Y}}$, before performing the decoding $y(z) = \arg\max_y \langle z, e_y \rangle$. Such a decoding partitions the simplex in regions whose frontiers are represented in dashed blue on the figure. The distribution p is characterized on the simplex by $\eta = \mathbb{E}_{Y \sim p}[e_Y] = \arg\min \mathbb{E}_{Y \sim p}[\|z - e_Y\|^2]$. This quantity η is exactly the quantity estimated by the least-squares surrogate. The median surrogate searches the minimizer z^* of the quantity $\mathcal{E}(z) = \mathbb{E}_{Y \sim p}[\|z - e_Y\|]$, whose level lines are represented in solid on the figure. One of the main advantage of the median surrogate compared to the least-squares one is that z^* is always farther away from the boundary frontier than η, meaning that for a similar estimation error on this quantity, the error on the decoding, which corresponds to an estimate of the mode of p, will be much smaller for the median surrogate. The left figure represents the case $p = (1, 0, 0)$, the right figure the case $p = (.45, .35, .2)$.

10.D Classification with a min-max game

In this section, we prove and extend on Proposition 92. First of all, let us consider the average loss, for $(v_y) \in \mathbb{R}^{\mathcal{Y}}$ summing to one

$$\bar{L}(v, s) = 1 - \sum_{y \in s} v_y = \sum_{y \notin s} v_y.$$

Consider now this loss conditioned on the observation $\mathbf{1}_{y \in s}$, we have plenty of characterizations of L,

$$\begin{aligned} L(v, s; \mathbf{1}_{y \in s} - \mathbf{1}_{y \notin s}) &= \mathbf{1}_{y \in s} \bar{L}(v, s) + \mathbf{1}_{y \notin s} \bar{L}(v, \mathcal{Y} \setminus s) = \mathbf{1}_{y \in s} \sum_{y \notin s} v_y + \mathbf{1}_{y \notin s} \sum_{y \in S} v_y \\ &= \mathbf{1}_{y \in s} + (\mathbf{1}_{y \notin s} - \mathbf{1}_{y \in s}) \sum_{y \in s} v_y = \mathbf{1}_{y \notin s} + (\mathbf{1}_{y \in s} - \mathbf{1}_{y \notin s}) \sum_{y \notin s} v_y \\ &= \frac{1}{2} - \frac{1}{2}(\mathbf{1}_{y \in s} - \mathbf{1}_{y \notin s}) \left(\sum_{y \in s} v_y - \sum_{y \notin s} v_y \right) = \frac{1}{2} + \frac{1}{2}(\mathbf{1}_{y \in s} - \mathbf{1}_{y \notin s}) \left(1 - 2 \sum_{y \in s} v_y \right). \end{aligned}$$

Minimizing this loss or the loss $2L - 1$ as defined in Proposition 92 is equivalent.

10.D.1 Consistency

Let us consider the loss as defined in this proposition, we have the characterization

$$L(v, s; \mathbf{1}_{y \in s} - \mathbf{1}_{y \notin s}) = (\mathbf{1}_{y \in s} - \mathbf{1}_{y \notin s}) \left(\sum_{y \in s} v_y - \sum_{y \notin s} v_y \right).$$

Let us rewrite (10.8) based on this previous characterization of the loss, we have

$$\mathbb{E}_Y[L(v, s, \mathbf{1}_{Y \in s} - \mathbf{1}_{Y \notin s})] = -(\mathbb{P}_Y(Y \in s) - \mathbb{P}_Y(Y \notin s)) \left(\sum_{y \in s} v_y - \sum_{y \notin s} v_y \right).$$

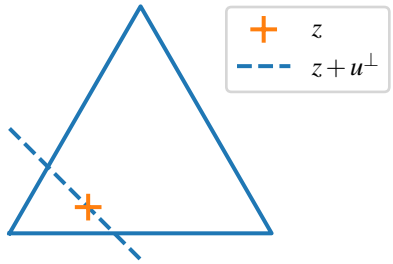

Figure 10.6: *Query strategy based on regression surrogate.* Retaking the simplex representation of Figure 10.5, the query strategy for classification approached with least-squares surrogate or median surrogate consists in looking at the current surrogate estimate z in the simplex $\Delta_{\mathcal{Y}}$, taking a random direction $u \in \mathbb{R}^{\mathcal{Y}}$ and querying $\text{sign}(\langle e_Y - z, u\rangle)$. We see that with three elements, when Y is deterministic, the optimal query strategy consists in considering $s = \{y\}$, while surrogate strategies, such as least-squares and median regression, that learn $z^* = e_y$, would only make such a query only two third of the time (which is the ratio of the solid angle of $[e_2, e_3]$ from e_1 divided by π). This shows that those surrogate strategies do not fully leverage the specific structure of the output.

Hence, without any context variable, the min-max game (10.8) can be rewritten as

$$\min_{v \in \Delta_{\mathcal{Y}}} \max_{\mu \in \Delta_{\mathcal{S}}} -\sum_{s \in \mathcal{S}} \mu_s (\mathbb{P}_Y(Y \in s) - \mathbb{P}_Y(Y \notin s)) \left(\sum_{y \in s} v_y - \sum_{y \notin s} v_y \right). \tag{10.30}$$

We will analyze this problem through the lens of a mix-actions zero-sum game. We know from von Neumann and Morgenstern (1944) that a solution to this min-max problem exists, and that one can switch the min-max to a max-min without modifying the value of the solution. Let us denote by (v^*, μ^*) the argument of a solution. To minimize the value of this game, the player v should play such that

$$\text{sign}(\sum_{y \in s} v^*_y - \sum_{y \notin s} v^*_y) = \text{sign}(\mathbb{P}(Y \in s) - \mathbb{P}(Y \notin s)) = \text{sign}(\sum_{y \in s} \mathbb{P}(Y = y) - \sum_{y \notin s} \mathbb{P}(Y = y)),$$

which allows this player to ensure a negative value to the game. Stated otherwise

$$\forall\, s \in \mathcal{S}, \qquad \mathbb{P}(Y \in s) > \frac{1}{2} \quad \Rightarrow \quad \sum_{y \in s} v^*_y \geq \frac{1}{2}. \tag{10.31}$$

As a consequence, if there exists any set such that $\mathbb{P}(Y \in s) = 1/2$, the best strategy of player μ is to play only those sets to ensure the value zero, and any v that satisfies (10.31) is optimal. It should be noted that (10.31) does not generally imply that $(v_y)_{y \in \mathcal{Y}}$ has the same ordering as $(\mathbb{P}(Y = y))_{y \in \mathcal{Y}}$.

When $\{y^*\} \in \mathcal{S}$ and $\mathbb{P}(Y = y^*) > 1/2$, if $v = \delta_{y^*}$, the prediction player is able to ensure a value of $\max_{s \in \mathcal{S}} -|2\,\mathbb{P}(Y \in s) - 1|$, which is maximized by the query player with $s = \{y^*\} \cup s'$ for any s' such that $\mathbb{P}(Y \in s') = 0$. Other strategies for v will only increase this value, hence $v^* = \delta_{y^*}$ which implies the first part of Proposition 92.

A counter example. While we hope that the solution (v^*, μ^*) does characterize the original solution y^*, it should be noted that v^* alone does not characterize y^*. Indeed, it is even possible to have v^* uniquely defined without having $y^* = \arg\max_{y \in \mathcal{Y}} v^*_y$. For example, consider the case where $\mathcal{Y} = \{1, 2, 3\}$ and $(\mathbb{P}(Y = i))_{i \in [3]} = (.4, .3, .3)$. By symmetry, the player μ only has to play on $\mathcal{S} = \{\{1\}, \{2\}, \{3\}\}$, which leads to the min-max game

$$\min_{v} \max_{\mu} \begin{pmatrix} \mu_{\{1\}} \\ \mu_{\{2\}} \\ \mu_{\{3\}} \end{pmatrix}^\top \begin{pmatrix} .2 & -.2 & -.2 \\ -.4 & .4 & -.4 \\ -.4 & -.4 & .4 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}.$$

The value of this game is $-.1$ and is achieved for $\mu^* = (.5, .25, .25)$, $v^* = (.25, .375, .375)$.

10.D.2 Optimization procedure

Let us rewrite the problem through the objective

$$\mathcal{E}(g, \mu) = \mathbb{E}_{(X,y)\sim\rho}\, \mathbb{E}_{S\sim\mu(x)}[L(g(X), S, \mathbf{1}_{Y\in S} - \mathbf{1}_{Y\notin S})].$$

We want to solve the min-max problem $\min_g \max_\mu \mathcal{E}(g, \mu)$. This problem can be solved efficiently based on the vector field point of view of gradient descent (Bubeck, 2015) if:

- we can parametrize the function $g : \mathcal{X} \to \Delta_\mathcal{Y}$ such that $\mathcal{E}$ is convex with respect to the parametrization of g;
- we can access unbiased stochastic gradients of $\mathcal{E}$ with respect to g that have a small second moment;
- we can parametrize the function $\mu : \mathcal{X} \to \Delta_\mathcal{S}$ such that $\mathcal{E}$ is concave with respect to the parametrization of μ;
- we can access unbiased stochastic gradients of $\mathcal{E}$ with respect to μ that have a small second moment.

The first two points are no problems, g can be parametrized with softmax regression, and since L is linear with respect to the scores, it will keep the problem convex. Moreover, to access a stochastic gradient of $\mathcal{E}$, one can sample $X_i \sim \rho_\mathcal{X}$ and $S_i \sim \mu(X_i)$ before querying $\mathbf{1}_{Y_i\in S_i}$ and computing the gradient of $L(g(X_i), S_i, \mathbf{1}_{Y_i\in S_i} - \mathbf{1}_{Y_i\notin S_i})$ with respect to the parametrization of g.

The third point is slightly harder to tackle. Since $\mathcal{E}$ is linear with respect to μ, one way to proceed is to find a linear parametrization of μ. In particular, one can take a family $(g_i)_{i\in[N]}$ of linearly independent functions from $\mathcal{X}$ to $\Delta_\mathcal{S}$ and search for g under the form $\sum_{i\in[N]} c_i g_i$ for (c_i) positive summing to one. To build such a family, one can eventually use "atom functions" and simple operations such as symmetry with respect to $\mathcal{Y}$ and $\mathcal{S}$, rescaling, translation, rotations with respect to $\mathcal{X}$. For example if $\mathcal{X}$ is a Banach space, one could define atom functions as, for $y_i \in \mathcal{Y}$

$$g_i : x \to \frac{\|x\|}{1+\|x\|}\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} e_s + \frac{1}{1+\|x\|} e_{\{y_i\}}.$$

Those functions could be rescaled and translated as $g_{\sigma,\tau,i}(x) = g_i(\sigma(x-\tau))$, in order to specify a family $(g_{\sigma,\tau,i})$ from few values for τ and σ.

The last point is the most difficult one. Without context variables, and with no-parametrization for μ, a naive unbiased gradient strategy for μ consists in asking random questions to update the full knowledge of $(\mathbb{P}(Y \in s))_{s\in\mathcal{S}}$. But such a strategy will be much worse than our median surrogate technique with queries $\mathbf{1}_{Y\in\{y\}}$ for y sampled uniformly at random in $\mathcal{Y}$. Eventually, one should go for a biased gradient strategy, while making sure to update μ coherently to avoid getting stalled on bad estimates as a result of biases.

10.E Experimental details

Our experiments are done in *Python*. We leverage the *C* implementation of high-level array instructions by Harris et al. (2020), as well as the visualization library of Hunter (2007). Randomness in experiments is controlled by choosing explicitly the seed of a pseudo-random number generator.

10.E.1 Comparison with fully supervised SGD

In this section, we investigate the difference between weakly and fully supervised SGD. According to Theorem 20, we only lost a constant factor of order $m^{3/2}$ in our rates compared to fully supervised (or plain) SGD. This behavior can be checked by adding the plain SGD curve on Figure 10.3. On the left side of Figure 10.8, we do observe that the risk of both Algorithm 2 and plain SGD decrease with same exponent with respect to number of iteration but with a different constant in front of the rates: that is we observe the same slopes on the logarithm scaled plot, but different intercepts. Going one step further to check the tightness of our bound, one can plot the intercept, or the error achieved by both Algorithm 2 and plain SGD as a function of the output space dimension m. The right side of Figure 10.8 shows evidence that this error grows as m^ε for some $\varepsilon \in [1, 3/2]$, which is coherent with our upper bound. Similarly to Figure 10.3, this figure was computed after cross validation to find the best scaling of the step sizes for each dimension m.

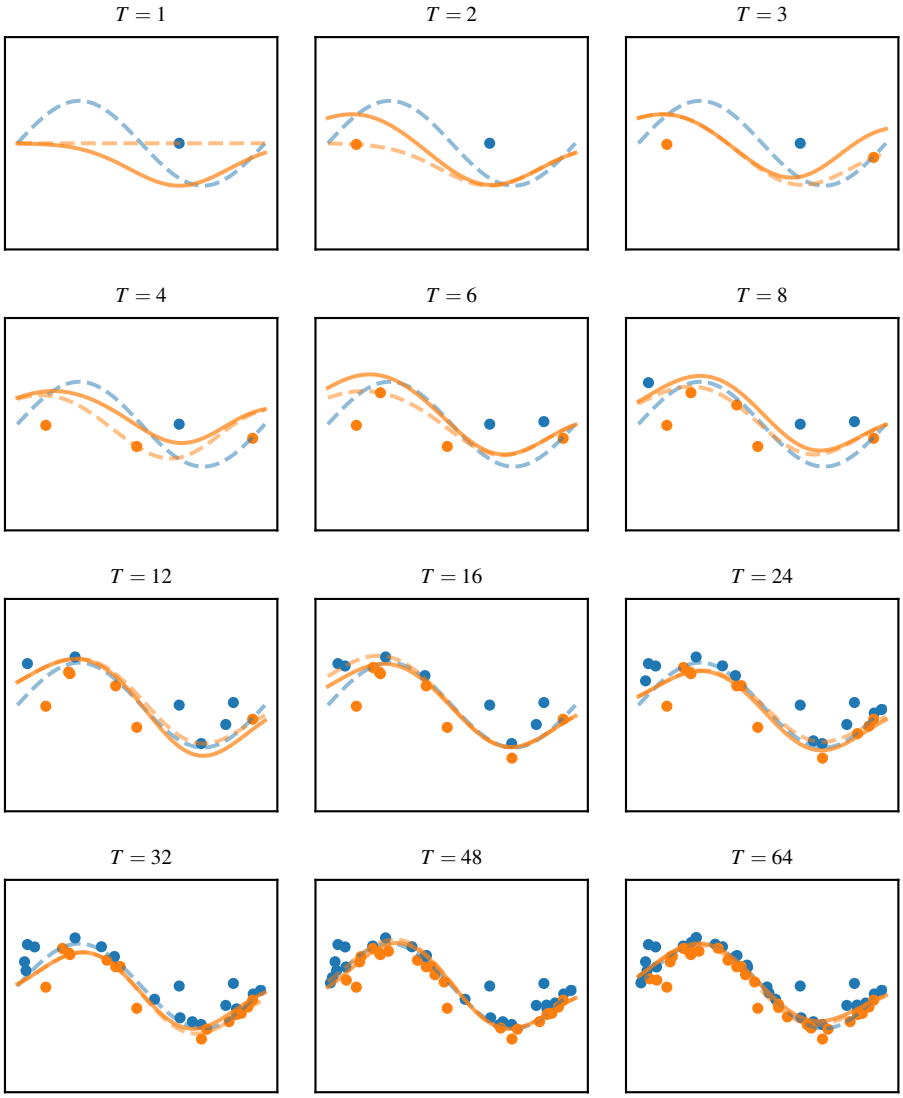

Figure 10.7: *Streaming history of the active strategy* to reconstruct the signal in dashed blue in the same setting as Figure 10.2. At any time t, a point X_t is given to us, our current estimate of θ_t plotted in dashed orange gives us $z = f_{\theta_t}(X_t)$, and we query $\operatorname{sign}(Y_t - z)$. Based on the answer to this query, we update θ_t to θ_{t+1} leading to the new estimate of the signal in solid orange. In this figure, we see that it might be useful for the practitioners in a streaming setting to reduce the bandwidth of φ as they advance in time.

10.E.2 Passive strategies for classification

A simple passive strategy for classification based on median surrogate consists in using the active strategy with coordinates sampling, that is u being uniform on $\{e_y\}_{y\in\mathcal{Y}}$, where $(e_y)_{y\in\mathcal{Y}}$ is the canonical basis of $\mathbb{R}^{\mathcal{Y}}$ used to define the simplex $\Delta_{\mathcal{Y}}$ as the convex hull of this basis. Querying $\mathbf{1}_{\langle g_\theta(x)-e_y, e_y\rangle>0}$ is formally equivalent to the query of $\mathbf{1}_{Y=y}$ when $g_\theta(x) \in \Delta_{\mathcal{Y}}$. This is the baseline we plot on Figure 10.3.

A more advanced passive baseline is provided by the infimum loss (Cour et al., 2011; Cabannes et al., 2020b). It consists in solving

$$\arg\min_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}_I(f) := \mathbb{E}_{(X,Y)\sim\rho}\, \mathbb{E}_S \left[L(f(X), S, \mathbf{1}_{Y\in S}) \right],$$

where S is a random subset of $\mathcal{Y}$ and L is defined from the original loss $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ as, for $z \in \mathcal{Y}$, $s \subset \mathcal{Y}$

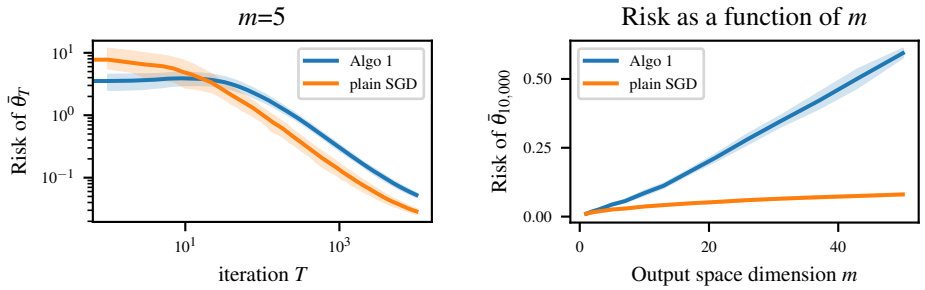

Figure 10.8: *Comparison of generalization errors of weakly and fully supervised SGD* as a function of the annotation budget T and output space dimension m. The setting is similar to Figure 10.3. We observe a transitory regime before convergence rates follows the behavior described by Theorem 20. The right side plots the error of both procedures after 10,000 iterations as a function of the output space dimension m between 1 and 50. The number of iteration ensures that, for all values of $m \in [50]$, the reported error is well characterized by our theory, in other terms that we have entered the regime described by Theorem 20.

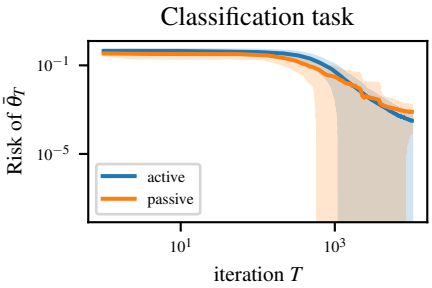

Figure 10.9: *Comparison with the infimum loss with better conditioned passive supervision* in a similar setting to Figure 10.3 yet with $m = 10$, $\varepsilon = 0$, that is X uniform on $\mathcal{X}$, and $\gamma_0 = 7.5$ for the active strategy and $\gamma_0 = 15$ for the passive strategy. We see no major differences between the active strategy based on the median surrogate and the passive strategy based on the median surrogate with the infimum loss. Note that the standard deviation is sometimes bigger than the average of the excess of risk, explaining the dive of the dark area on this logarithmic-scaled plot.

and $y \in \mathcal{Y}$,

$$L(z, s, \mathbf{1}_{y \in s}) = \begin{cases} \inf_{y' \in s} \ell(z, y') & \text{if } y \in s \\ \inf_{y' \notin s} \ell(z, y') & \text{otherwise.} \end{cases}$$

Random subsets S could be generated by making sure that the variable $(y \in S)_{y \in \mathcal{Y}}$ are independent balanced Bernoulli variables; and by removing the trivial sets $S = \emptyset$ and $S = \mathcal{Y}$ from the subsequent distribution. In order to optimize this risk in practice, one can use a parametric model and a surrogate differentiable loss together with stochastic gradient descent on the empirical risk. For classification with the 0-1 loss, we can reuse the surrogate introduced in Proposition 91 and minimize, assuming that we always observed $\mathbf{1}_{Y_i \in S_i} = 1$ for simplicity,

$$\hat{\mathcal{R}}_{I,S}(\theta) = \sum_{i=1}^{n} \inf_{y \in S_i} \left\| g_\theta(X_i) - e_y \right\|.$$

Stochastic gradients are then given by, assuming ties have no probability to happen,

$$\nabla_\theta \inf_{y \in S_t} \left\| g_\theta(X_t) - e_y \right\| = \left(\frac{g_\theta(X_t) - e_{y^*}}{\left\| g_\theta(X_t) - e_{y^*} \right\|} \right)^\top D g_\theta(X_t) \quad \text{with} \quad y^* := \arg\max_{y \in S_t} \left\langle g_\theta(X_t), e_y \right\rangle.$$

This gives a good passive baseline to compare our active strategy with. In our experiments with the Gaussian kernel, see Figure 10.9 for an example, we witness that this baseline is highly competitive. Although we find that it is slightly harder to properly tune the step size for SGD, and that the need to compute an argmax for each gradient slows-down the computations.

10.E.3 Real-world classification datasets

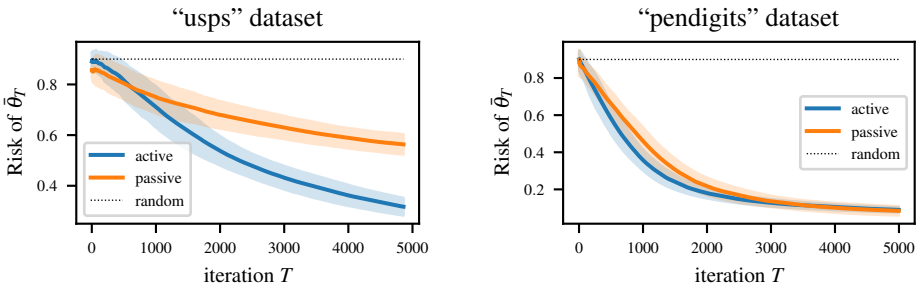

Figure 10.10: *Testing errors on two LIBSVM datasets* with a similar setting to Figure 10.9. Those empirical errors are reported after averaging over 100 different splits of the datasets. The step size parameter was optimized visually, which led to $\gamma_0 = 15$ for the active strategy on "USPS", $\gamma_0 = 60$ for the passive one, $\gamma_0 = 7.5$ for the active strategy on "pen digits", $\gamma_0 = 30$ for the passive one. The dotted line represents $\mathcal{R} = 1 - m^{-1}$ which is the performance of a random model.

In Figure 10.10, we compare the "well-conditioned" passive baseline with our active strategy on the real-world problems of LIBSVM (Chang and Lin, 2011). We choose the "USPS" and "pen digits" datasets as they contain $m = 10$ classes each with $n = 7291$ and $n = 7494$ samples respectively, with $d = 50$ and $d = 16$ features each. We have chosen those datasets as they present enough classes that leads to many different sets S to query, and they are made of the right number of samples to do some experiments on a laptop without the need for "advanced" computational techniques such as caching or low-rank approximation (Meanti et al., 2020). On Figure 10.10, we use the same linear model as for Figure 10.3, that is a Gaussian kernel. We choose the bandwidth to be $\sigma = d/5$, and we normalize the features beforehand to make sure that they are all centered with unit variance. We report error by taking two thirds of the samples for training and one third for testing, and averaging over one hundred different ways of splitting the datasets. We observe that the active strategy leads to important gains on the "USPS" dataset, yet is not that useful for the "pen digits" dataset. We have not dug in to understand those two different behaviors.

10.E.4 Real-world regression dataset & Nyström method

In this section, we provide two experiments on real-world datasets.

In order to deal with big regression datasets, it is useful to approximate the parameter space $\mathcal{Y} \otimes \mathcal{H}$ in Assumption 20 with a small dimensional space. To do so, let us remark that given samples $(X_i)_{i \leq n} \in \mathcal{X}^n$ for $n \in \mathbb{N}$, we know that our estimate f_{θ_n} can be represented as

$$f_{\theta_n}(\cdot) = \sum_{i \leq n} \sum_{j \leq m} a_{ij} \langle \varphi(x_i), \varphi(\cdot) \rangle e_j,$$

for some $(a_{ij}) \in \mathbb{R}^{p \times m}$ and where $(e_j)_{j \leq m}$ is the canonical basis of $\mathcal{Y} = \mathbb{R}^m$. For large datasets, that is when n is large, it is smart to approximate this representation through the parameterization

$$f_a(x) = \sum_{i \leq p} \sum_{j \leq m} a_{ij} k(x, x_i) e_j,$$

where $p \leq n$ is the rank of our approximation, and k is the kernel defined as $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$. Stated with words, we only use a small number p, instead of n, of vectors $\varphi(x_i)$ to parameterize f. This allows to only keep a matrix of size $p \times m$ in memory instead of $n \times m$, while not fundamentally changing the statistical guarantee of the method (Rudi et al., 2015). In this setting, the stochastic gradients are specified from the fact that

$$u^\top D_a f_a(x) = (u_j k(x, x_i))_{i,j} \in \mathbb{R}^{p \times m}.$$

In other terms, in order to update the parameter a with respect to the observation made at (x, u), we check how much each coordinate of a determines the value of $u^\top f_a(x)$.

In the following, we experiment with two real-world datasets. In order to learn the relation between inputs and outputs, we use a Gaussian kernel after normalizing input features so that each of them has zero mean and unit variance. To keep computational cost, we sample p random (Nyström) representers among the training inputs which are used to parameterize functions. To avoid overfitting, we add a small regularization to the empirical objective. It reads $\lambda \|\theta\|^2_{\mathcal{H}}$ with our notations and corresponds to the Hilbertian norm inherited from the reproducing kernel k of the function f_θ (Scholkopf and Smola, 2001).

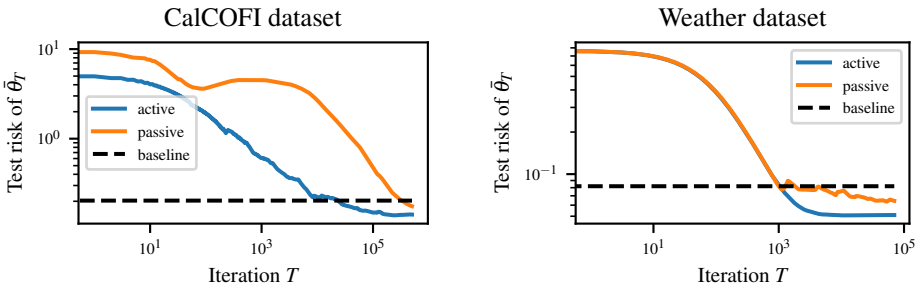

Figure 10.11: *Testing error on two real-world regression datasets.* On both datasets, a single pass was made through the data in a chronological fashion, and errors were computed from the 26,453 most recent data samples for the "Weather" dataset, and from a random sample of 10,000 samples among the 155,140 most recent samples for the "CalCOFI" dataset.

Our first experiment is based on the data collected by the California Cooperative Oceanic Fisheries Investigation between March 1949 and November 2016. It consists of more than 800,000 seawater samples including measurements of nutriments (set aside in our experiments) together with pressure, temperature, salinity, water density, dynamic height (providing five input parameters), as well as dissolved oxygen, and oxygen saturation (the two outputs we would like to predict). We assume that we can measure if any weighted sum of oxygen concentration and saturation is above a threshold by letting some population of bacteria evolves in the water sample and checking if it survives after a day. If the measurements are done on the day of the sample collection, this setting exactly fits in the streaming active labeling framework. After cleaning the dataset for missing values, the dataset contains 655,140 samples. The "CalCOFI" dataset results are reported on the left of Figure 10.11, parameters were chosen as $p = 100$, $\sigma = 10$, $\lambda = 10^{-6}$ and $\gamma_0 = 1$. For the passive strategy, random queries were chosen to follow a normal distribution with the same mean as the targets and one third of their standard deviation (*i.e.* we ask if the apparent temperature is lower than the usual one plus or minus a perturbation). The plotted baseline corresponds to linear regression performed over the entire dataset. It takes about 10,000 samples for our active strategy to be competitive with this baseline, and 200,000 samples for the passive one.

The second experiment makes use of data collected through the Dark Sky API (which is now part of Apple WeatherKit). It is made of 96,454 weather summaries between 2006 and 2016 in the city of Szeged, Hungary. Our task consists in computing the apparent temperature from real temperature, humidity, wind speed, wind bearing, visibility and pressure. The apparent temperature is an index that searches to quantify the subjective feeling of heat that humans perceive, it is expressed on the same scale as real temperature. One way to measure it would be to ask some humans if the outside is hotter or colder than a controlled room with a specific temperature and neutral meteorological conditions. Once again, this exactly fits into our streaming active labeling setting. The "Weather" dataset results are reported on the right of Figure 10.11. The baseline consists in predicting the apparent temperature as the real temperature. We observe a transitory regime where the first 1,000 samples seem to be used to calibrate the weights α. During this regime, our estimate is too bad for the active strategy to make smarter queries than the "random" ones that have been calibrated on temperature statistics. The main difference in the learning dynamic between the active and passive strategies is observed on the remaining 69,000 training samples. The parameters were the same as the "CalCOFI" dataset but for $\gamma_0 = 10^{-2}$.

Conclusion

In this thesis, we have approached weakly supervised learning through partial supervision. We have leveraged the algorithm of Ciliberto et al. (2016) to provide consistent estimators once we had formulated the problem through the infimum loss, which was the focus of Cabannes et al. (2020b). By characterizing implicit disambiguation due to the infimum loss, we have revisited and generalized the approach of Bach and Harchaoui (2007) to any type of discrete output problems in Cabannes et al. (2021b). Such a disambiguation strategy has the advantage of working with a smaller surrogate space (working on functions from $\mathcal{X}$ to $\mathbb{R}^{\mathcal{Y}}$ rather than from $\mathcal{X}$ to $\mathbb{R}^{2^{\mathcal{Y}}}$), hence reducing the variance of our estimates, and making approximation assumptions more comprehensible. Regarding optimization, our min-min formulations are hard to tackle in a generic fashion; while one can reuse the relaxation in Bach and Harchaoui (2007), our implementations were based on alternate minimization with good starting points; readers and practitioners eager to come up with different strategies might refer to literature on bilevel optimization for non-convex problems. In Cabannes et al. (2021a), advocating for the leverage of unlabeled data, we have brought up the approach of Zhu et al. (2003) to the realm of kernel methods; and to ensure computational efficiency, we have adapted the low-rank approximation of Rudi et al. (2015) and its proof to our specific setting with derivatives. We hope to see future works pushing forwards the usage of kernel methods with derivatives, could it be for sampling through Langevin dynamics (in the spirit of Pillaud-Vivien, 2020a) or for penalties inducing sparsity (in the spirit of Rosasco et al., 2013; Follain et al., 2022). Providing efficient and user-friendly open-source implementations would arguably foster the diffusion of kernels with derivatives in the research community.

In order to help the practitioner collecting data, we finally set ourselves before the data collection process, and introduced the "active labeling" problem. Following the observation that one does not need to access full information in order to build stochastic gradients, we provided a first solution to this problem based on stochastic gradient descent in Cabannes et al. (2022) that is naturally adapted to streaming settings. In a near future, we would like to focus on exploiting the discrete-output structure of classification problems more subtly than surrogate methods, notably with ideas steaming from the "bandit" literature. In particular, we will explore how entropic coding could help to tackle the active labeling problem. Beside this project, the "active" setup opens up new possibilities, *e.g.*, to detach ourselves from min-min formulations and try to come up with well-grounded min-max formulations that are easier to optimize, or, on a completely different level, to deal with privacy constraints.

While working on the proof of Cabannes et al. (2021b), we discovered that usual rates of convergence on discrete output learning problems are often suboptimal thanks to the work of Audibert and Tsybakov (2007). In substance, if you have L^∞-exponential concentration inequality, and margin conditions, then you can have up to exponential convergence rates. We studied the question in more detail in Cabannes et al. (2021c) and Cabannes and Vigogna (2022). Those derivations can still be pushed further, could it be to better understand the interplay between estimation error (a.k.a. variance) and approximation error (a.k.a. bias) without decoupling them, or by adapting the proof structure with concentration on other quantities than suprema and corresponding relative hardness conditions. Regarding statistical learning, a project of interest to pursue, yet which require other tools than the ones presented in the thesis, is to approach learning theory without thinking in terms of functions classes, but by thinking directly with measures, especially at the light of the disambiguation principles that we expressed directly in the space of measures. Interestingly, such an approach might help to exchange (if not unify) ideas between works on weakly supervised learning (*i.e.* missing output data) and works on missing (input) data.

Finally, while this thesis was more of theoretical nature, in order to maximize my "tangible impact" as a researcher, I would like to look at issues that are closer to "production" and focus more on real-world experiments. I hope that my future postdoc position at FAIR New York will provide me with such opportunities, for example by exposing myself to ongoing experimental works with self-supervised learning.

Bibliography

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *Conference on Operating Systems Design and Implementation*, 2016.

Nir Ailon. Active learning ranking from pairwise preferences with almost optimal query complexity. In *Advances in Neural Information Processing Systems*, 2011.

Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: ranking and clustering. In *Symposium on Theory of Computing*, 2005.

Ahmed El Alaoui, Xiang Cheng, Aaditya Ramdas, Martin Wainwright, and Michael Jordan. Asymptotic behavior of ℓ_p-based Laplacian regularization in semi-supervised learning. In *Conference on Learning Theory*, 2016.

Jean-Baptiste Alayrac. *Structured Learning from Videos and Language*. Phd thesis, Ecole Normale Supérieure, 2018.

Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Charalambos Aliprantis and Kim Border. *Infinite Dimensional Analysis: A Hitchhikers Guide*. Springer, 2006.

Martin Anthony and Peter Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. Technical Report 1907.02893, ArXiv, 2019.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Sanjeev Arora, László Babai, Jacques Stern, and Elizabeth Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences*, 54(2):317–331, 1997.

Kenneth Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346, 1950.

Jean-Yves Audibert and Alexander Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

Dmitry Babichev, Dmitrii Ostrovskii, and Francis Bach. Efficient primal-dual algorithms for large-scale multiclass classification. Technical Report 1902.03755, ArXiv, 2019.

Francis Bach. *Learning Theory from First Principles*. To appear at MIT press, 2023.

Francis Bach and Zaïd Harchaoui. DIFFRAC: a discriminative and flexible framework for clustering. In *Advances in Neural Information Processing Systems*, 2007.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems*, 2013.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Maching Learning*, 4(1):1–106, 2012.

Peter Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135 (3):311–334, 2006.

Peter Bartlett, Michael Jordan, and Jon McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Peter Bartlett, Dylan Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Gerald Beer. *Topologies on closed and closed convex sets*. Springer, 1993.

Sven Behnke. *Hierarchical Neural Networks for Image Interpretation*. Springer, 2003.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. Label propagation and quadratic criterion. In *Semi-Supervised Learning*. MIT Press, 2006.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Maching Learning Research*, 22(7):1–108, 2021.

Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity, 2019.

James Bennett and Stan Lanning. The netflix prize. In *Knowledge Discovery and Data Mining Cup and Workshop*, 2007.

Serguei Bernstein. On certain modifications of Chebyshev's inequality. *Doklady Akademii Nauk SSSR*, 17 (6):275–177, 1937.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2019.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer, 2015.

Lucien Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65(2):181–237, 1983.

Ingrid Blaschzyk and Ingo Steinwart. Improved classification rates under refined margin conditions. *Electronic Journal of Statistics*, 12(1):793–823, 2018.

Mathieu Blondel, André Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

Salomon Bochner. Monotone funktionen, stieltjessche integrale und harmonische analyse. *Mathematische Annalen*, 108(1):378–410, 1933.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, 2007.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Mark Braverman, Jieming Mao, and Yuval Peres. Sorted top-k in rounds. In *Conference on Learning Theory*, 2019.

Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. *Classification and Regression Trees*. Chapman & Hall, 1984.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Vivien Cabannes. Le futur du numérique sera-t-il incarné ? *Esprit*, 487:117–125, 2022.

Vivien Cabannes and Stefano Vigogna. A case of exponential convergence rates for SVM. Technical report, ArXiv, 2022.

Vivien Cabannes, Thomas Kerdreux, Louis Thiry, and Tina & Charly. Dialog on a canvas with a machine. In *NeurIPS workshop on Creativity*, 2019.

Vivien Cabannes, Thomas Kerdreux, and Louis Thiry. Diptychs of human and machine perception. In *NeurIPS workshop on Creativity*, 2020a.

Vivien Cabannes, Alessandro Rudi, and Francis Bach. Structured prediction with partial labelling through the infimum loss. In *International Conference on Machine Learning*, 2020b.

Vivien Cabannes, Loucas Pillaud-Vivien, Francis Bach, and Alessandro Rudi. Overcoming the curse of dimensionality with Laplacian regularization in semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2021a.

Vivien Cabannes, Alessandro Rudi, and Francis Bach. Disambiguation of weak supervision with exponential convergence rates. In *International Conference on Machine Learning*, 2021b.

Vivien Cabannes, Alessandro Rudi, and Francis Bach. Fast rates in structured prediction. In *Conference on Learning Theory*, 2021c.

Vivien Cabannes, Francis Bach, Vianney Perchet, and Alessandro Rudi. Active labeling: streaming stochastic gradients. Technical report, ArXiv, 2022.

Aylin Caliskan, Joanna Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Emmanuel Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2): 489–509, 2006.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *International Conference of Machine Learning*, 2007.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2006.

Vittorio Castelli and Thomas Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16(1):105–111, 1995.

René Erlín Castillo and Humberto Rafeiro. *An Introductory Course in Lebesgue Spaces*. Springer, 2016.

Nicolò Cesa-Bianchi, Claudio Gentile, and Luca Zaniboni. Incremental algorithms for hierarchical classification. *Journal of Machine Learning Research*, 7(2):31–54, 2006.

Nicolò Cesa-Bianchi, Tommaso Cesari, and Vianney Perchet. Dynamic pricing with finitely many unknown valuations. In *International Conference on Algorithmic Learning Theory*, 2019.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, 2014.

George Chen and Devavrat Shah. Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends in Machine Learning*, 10(5-6):337–588, 2018.

Lin Chen and Sheng Xu. Deep neural tangent kernel and Laplace kernel have the same RKHS. In *International Conference on Learning Representations*, 2021.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.

Clément Chevalier, Julien Bect, David Ginsbourger, Emmanuel Vázquez, Victor Picheny, and Yann Richet. Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4):455–465, 2014.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, 2020.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2019.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways. Technical Report 2204.02311, ArXiv, 2022.

Jesús Cid-Sueiro, Darío García-García, and Raúl Santos-Rodríguez. Consistency of losses for learning from weak labels. In *Machine Learning and Knowledge Discovery in Databases*, 2014.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems*, 2016.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A general framework for consistent structured prediction with implicit loss embeddings. *Journal of Machine Learning Research*, 21(98):1–67, 2020.

William Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

Maxime Cohen, Ilan Lobel, and Renato Paes Leme. Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943, 2020.

Ronald Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21 (1):5–30, 2006.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Council of European Union. Regulation (EU) 2016/679 of the European parliament (General Data Protection Regulation), 2016.

Timothée Cour, Benjamin Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12(42):1501–1535, 2011.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Thomas Cover and Joy Thomas. *Elements of Information Theory*. Wiley, 1991.

Mark Craven and Johan Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *International Conference on Intelligent Systems for Molecular Biology*, 1999.

Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3D human pose from structure and motion. In *European Conference on Computer Vision*, 2018.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Conference on Computer Vision and Pattern Recognition*, 2005.

Sanjoy Dasgupta. Two faces of active learning. *Theoretical Computer Science*, 412(19):1767–1781, 2011.

Allan Davis, Thomas Wiegers, Phoebe Roberts, Benjamin King, Jean Lay, Kelley Lennon-Hopkins, Daniela Sciaky, Robin Johnson, Heather Keating, Nigel Greene, Robert Hernandez, Kevin McConnell, Ahmed Enayetallah, and Carolyn Mattingly. A CTD-Pfizer collaboration: manual curation of 88,000 scientific articles text mined for drug-disease and drug-phenotype interactions. *Database (Oxford)*, 2012.

Philip Dawid and Allan Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

Arthur Dempster, Nan Laird, and Donald Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

Thierry Denoeux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):119–130, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

Luc Devroye and Terry Wagner. Distribution-free consistency results in nonparametric discrimination and regression function estimation. *The Annals of Statistics*, 8(2):231–239, 1980.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

Thomas Dietterich, Richard Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

AnHai Doan, Raghu Ramakrishnan, and Alon Halevy. Crowdsourcing systems on the world-wide web. *Communication of the ACM*, 54(4):86–96, 2011.

Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *International Conference on Computer Vision*, 2017.

David Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

John Duchi, Lester Mackey, and Michael Jordan. On the consistency of ranking algorithms. In *International Conference on Machine Learning*, 2010.

Richard Duda, Peter Hart, and David Stork. *Pattern Classification, 2nd Edition*. Wiley, 2000.

Richard Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Gabriel Dulac-Arnold, Neil Zeghidour, Marco Cuturi, Lucas Beyer, and Jean-Philippe Vert. Deep multiclass learning from label proportions. Technical Report 1905.12909, ArXiv, 2019.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, 2006.

Alan Edelman and Yuyang Wang. The GSVD: where are the ellipses?, matrix trigonometry, and more. *SIAM Journal on Matrix Analysis and Applications*, 41(4):1826–1856, 2020.

David Eriksson, Kun Dong, Eric Hans Lee, David Bindel, and Andrew Gordon Wilson. Scaling Gaussian process regression with derivatives. In *Advances in Neural Information Processing Systems*, 2018.

Robert Fano. *Transmission of Information: A Statistical Theory of Communications*. MIT press, 1968.

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential Experimental Design for Transductive Linear Bandits. In *Advances in Neural Information Processing Systems*, 2019.

Mathias Fink. Time reversed acoustics. *Physics Today*, 50(3):34–40, 1997.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.

Evelyn Fix and Joseph Hodges. Discriminatory analysis. Nonparametric discrimination: Consistency properties. Technical report, School of Aviation Medicine, Randolph Field, Texas, 1951.

Bertille Follain, Umut Şimşekli, and Francis Bach. Non-parametric subspace learning through trace norm penalty. Technical report, In preparation at INRIA, 2022.

Dimitris Fotakis, Alkis Kalavasis, Vasilis Kontonis, and Christos Tzamos. Efficient algorithms for learning from coarse labels. In *Conference on Learning Theory*, 2021.

Jerome Friedman. Flexible metric nearest neighbor classification. Technical report, Department of Statistics, Stanford University, 1994.

Rafael Frongillo and Bo Waggoner. Surrogate regret bounds for polyhedral losses. In *Advances in Neural Information Processing Systems*, 2021.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

Sachin Gangaputra and Donald Geman. A design principle for coarse-to-fine classification. In *Conference on Computer Vision and Pattern Recognition*, 2006.

Eva García-Martín, Crefeda Faviola Rodrigues, Graham Riley, and Håkan Grahn. Estimation of energy consumption in machine learning. *Journal of Parallel and Distributed Computing*, 134:75–88, 2019.

Nicolás García Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepčev. Error estimates for srdectritionl convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator. *Foundations of Computational Mathematics*, 20(4):827–887, 2019.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, 2016.

Josselin Garnier and George Papanicolaou. *Passive imaging with ambient noise*. Cambridge University Press, 2016.

Philippe Gautret, Jean-Christophe Lagier, Philippe Parola, Van Thuan Hoang, Line Meddeb, Morgane Mailhe, Barbara Doudier, Johan Courjon, Valérie Giordanengo, Vera Esteves Vieira, Hervé Tissot Dupont, Stéphane Honoréé, Philippe Colson, Eric Chabrière, Bernard La Scola, Jean-Marc Rolain, Philippe Brouqui, and Didier Raoult. Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. *International journal of antimicrobial agents*, 56(1):105949, 2020.

Timothy Gebhard, Niki Kilbertus, Ian Harry, and Bernhard Schölkopf. Convolutional neural networks: A magic bullet for gravitational-wave detection? *Physical Review D*, 100(6):063015, 2019.

Donald Geman and Bruno Jedynak. Shape recognition and twenty questions. Technical report, INRIA, 1993.

Claudio Gentile and Manfred Warmuth. Linear hinge loss and average margin. In *Advances in Neural Information Processing Systems*, 1999.

Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn & TensorFlow*. O'Reilly, 2017.

Edgar Gilbert. A comparison of signalling alphabets. *Bell System Technical Journal*, 31(3):504–522, 1952.

Mike Glennon and Michael Shirer. Investment in artificial intelligence solutions will accelerate as businesses seek insights, efficiency, and innovation, according to a new IDC spending guide. Technical report, International Data Corporation, 2021.

Gene Golub and Charles Van Loan. *Matrix computations (3rd edition)*. Johns Hopkins University Press, 1996.

Gene Golub and Charles Van Loan. *Matrix computations (4th edition)*. Johns Hopkins University Press, 2013.

Chen Gong, Tongliang Liu, Yuanyan Tang, Jian Yang, Jie Yang, and Dacheng Tao. A regularization approach for instance-based superset label learning. *IEEE Transactions on Cybernetics*, 48(3):967–978, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Yves Grandvalet. Logistic regression for partial labels. In *Information Processing and Management of Uncertainty*, 2002.

Romain Guillaume, Inés Couso, and Didier Dubois. Maximum likelihood with coarse data based on robust optimisation. In *International Symposium on Imprecise Probability*, 2017.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2-3):131–309, 2014.

Charles Harris, Jarrod Millman, Stéfan van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

James Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph Laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8(48):1325–1368, 2007.

Daniel Heitjan and Donald Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19(4):2244–2253, 1991.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Klaus-Uwe Höffgen and Hans Ulrich Simon. Robust trainability of single neurons. In *Computational Learning Theory*, 1992.

Alan Hoffman and Joseph Kruskal. Integral boundary points of convex polyhedra. In *50 Years of Integer Programming 1958-2008 - From the Early Years to the State-of-the-Art*. Springer, 2010.

Peter Huber. *Robust Statistics*. Wiley, 1981.

Eyke Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55:1519–1534, 2014.

Eyke Hüllermeier and Weiwei Cheng. Superset learning based on generalized loss minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2015.

Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 72(16):1897–1916, 2008.

John Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

IBM. *IBM ILOG CPLEX 12.7 User's Manual*. IBM ILOG CPLEX Division, 2017.

Il'dar Ibragimov and Rafail Khas'minskii. On the estimation of an infinite-dimensional parameter in gaussian white noise. *Doklady Akademii Nauk SSSR*, 236(5):1053–1055, 1977.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, 2019.

Tommi Jaakkola, Mark Diekhans, and David Haussle. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1-2):95–114, 2000.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Kevin Jamieson and Robert Nowak. Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems*, 2011.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems*, 2018.

Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In *Advances in Neural Information Processing Systems*, 2002.

Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*, 1998.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

Michael Jordan. Artificial Intelligence — the revolution hasn't happened yet. *Harvard Data Science Review*, 2019.

Armand Joulin, Francis Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2010.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon Kohl, Andrew Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition*, 2019.

Mina Karzand and Robert Nowak. Maximin active learning in overparameterized model classes. *IEEE Journal on Selected Areas in Information Theory*, 1(1):167–177, 2020.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the Asoociation for Computing Machinery*, 45(6):983–1006, 1998.

John Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959.

Maurice Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Stefan Klus, Feliks Nüske, and Boumediene Hamzi. Kernel-based approximation of the Koopman generator and Schrödinger operator. *Entropy*, 22(7):722, 2020.

Donald Knuth. The computer as master mind. *Journal of Recreational Mathematics*, 9(1):1–6, 1977.

Andrey Kolmogorov and Vladimir Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

Vladimir Koltchinskii and Olexandra Beznosova. Exponential convergence rates in classification. In *International Conference on Computational Learning Theory*, 2005.

Anna Korba, Alexandre Garcia, and Florence d'Alché-Buc. A structured prediction approach for label ranking. In *Advances in Neural Information Processing Systems*, 2018.

Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*, 2016.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Canadian Institute for Advanced Research, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, 2001.

Andrew Larkoski, Ian Moult, and Benjamin Nachman. Jet substructure at the large hadron collider: A review of recent advances in theory and machine learning. *Physics Reports*, 841:1–63, 2020.

Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time-series. In *The handbook of brain theory and neural networks*. MIT Press, 1995.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Marc Lelarge and Léo Miolane. Asymptotic Bayes risk for Gaussian mixture in a semi-supervised setting. In *International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2019.

Julian Lienen and Eyke Hüllermeier. From label smoothing to label relaxation. In *AAAI Conference on Artificial Intelligence*, 2021.

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3): 868–890, 2020.

Allen Liu, Renato Paes Leme, and Jon Schneider. Optimal contextual pricing and extensions. In *Symposium on Discrete Algorithms*, 2021.

Li-Ping Liu and Thomas Dietterich. Learnability of the superset label learning problem. In *International Conference on Machine Learning*, 2014.

Steve Lohr. A $1 million research bargain for Netflix, and maybe a model for other. *New York Times*, 2009.

Philip Long and Rocco Servedio. Consistency versus realizable h-consistency for multiclass classification. In *International Conference on Machine Learning*, 2013.

Xinghua Lou and Fred Hamprecht. Structured learning from partial annotations. In *International Conference on Machine Learning*, 2012.

David Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, 1999.

Jie Luo and Francesco Orabona. Learning from candidate labeling sets. In *Advances in Neural Information Processing Systems*, 2010.

Michael Lustig, David Donoho, Juan Santos, and John Pauly. Compressed sensing MRI. *Signal Processing Magazine*, 25(2):72–82, 2008.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Conference on Computer Vision and Pattern Recognition*, 2015.

Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. Convolutional kernel networks. In *Advances in Neural Information Processing Systems*, 2014.

Stéphane Mallat. Multifrequency channel decompositions of images and wavelet models. *Transactions on Acoustics, Speech, and Signal Processing*, 37(12):2091–2110, 1989.

Stéphane Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150203, 2016.

Enno Mammen and Alexander Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6): 1808–1829, 1999.

Gideon Mann and Andrew McCallum. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Journal of Machine Learning Research*, 11(32):955–984, 2010.

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Conference on Learning Theory*, 2019.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5): 2326–2366, 2006.

Andreas Maurer. A vector-contraction inequality for Rademacher complexities. In *Algorithmic Learning Theory*, 2016.

Warren McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, and Alessandro Rudi. Kernel methods through the roof: Handling billions of points efficiently. In *Advances in Neural Information Processing Systems*, 2020.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

James Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209:415–446, 1909.

Charles Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(95):2651–2667, 2006.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *International Conference on Computer Vision*, 2019.

Stanislav Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *International Joint Conference on Natural Language Processing*, 2009.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learnig*, 10(1-2): 1–141, 2017.

Stephen Muggleton and Luc de Raedt. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19-20:629–679, 1994.

Èlizbar Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Statistical analysis of semi-supervised learning: The limit of infinite unlabelled data. In *Advances in Neural Information Processing Systems*, 2009.

John Nelder and Robert Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

Nam Nguyen and Rich Caruana. Classification with partial labels. In *International Conference on Knowledge Discovery and Data Mining*, 2008.

Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, 2021.

Atsushi Nitanda and Taiji Suzuki. Stochastic gradient descent with exponential convergence rates of expected classification errors. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Alex Nowak-Vila. *Structured prediction with theoretical guarantees*. Phd thesis, Ecole Normale Supérieure, 2021.

Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. Sharp analysis of learning with discrete losses. In *Artificial Intelligence and Statistics*, 2019.

Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. A general theory for structured prediction with smooth convex surrogates. Technical Report 1902.01958, ArXiv, 2020.

Harry Nyquist. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644, 1928.

Sinno Pan and Qiang Yang. A survey on transfer learning. *Transactions on Knowledge and Data Engineering*, 22(10):1345–1459, 2010.

George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *International Conference on Computer Vision*, 2015.

Vardan Papyan, Yaniv Romano, and Michael Elad. Convolutional neural networks analyzed via convolutional sparse coding. *Journal of Machine Learning Research*, 18(83):1–52, 2017.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Andrzej Pelc. Searching games with errors - fifty years of coping with liars. *Theoretical Compututer Science*, 270(1):71–109, 2002.

Vianney Perchet and Marc Quincampoix. On a unified framework for approachability with full or partial monitoring. *Mathematics of Operations Research*, 40(3):596–610, 2015.

Loucas Pillaud-Vivien. Statistical estimation of the poincaré constant and application to sampling multimodal distributions. In *International Conference on Artificial Intelligence and Statistics*, 2020a.

Loucas Pillaud-Vivien. *Learning with Reproducing Kernel Hilbert Spaces: Stochastic Gradient Descent and Laplacian Estimation*. Phd thesis, Ecole Normale Supérieure, 2020b.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, 2018a.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Exponential convergence of testing error for stochastic gradient methods. In *Conference on Learning Theory*, 2018b.

Iosif Pinelis and Aleksandr Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory of Probability and Its Applications*, 30(1):143–148, 1986.

Bahar Qarabaqi and Mirek Riedewald. User-driven refinement of imprecise queries. In *International Conference on Data Engineering*, 2014.

Novi Quadrianto, Alexander Smola, Tibério Caetano, and Quoc V. Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10(82):2349–2374, 2009.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

Alexander Ratner, Stephen Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, 29(2):709–730, 2020.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(49):1369–1392, 2007.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

Lorenzo Rosasco, Silvia Villa, Sofia Mosci, Matteo Santoro, and Alessandro Verri. Nonparametric sparsity and regularization. *Journal of Machine Learning Research*, 14(16):1665–1714, 2013.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

Donald Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, 2015.

Robert Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

Bernhard Scholkopf and Alexander Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Laurent Schwartz. Sous-espaces Hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *Journal d'Analyse Mathématique*, 13(1):115–256, 1964.

Ramprasaath Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision*, 2017.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison, 2010.

Claude Shannon. Communication in the presence of noise. *Proceedings of the Institute of Radio Engineers*, 37(1):10–21, 1949.

William Sheppard. On the calculation of the most probable values of frequency constants, for data arranged according to equidistant division of a scale. *Proceedings of the London Mathematical Society*, s1-29(1): 353–380, 1897.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Emil Sidky and Xiaochuan Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine & Biology*, 53(17):4777–4807, 2008.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Herbert Simon. The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6): 467–482, 1962.

Eric Slud. Distribution inequalities for the binomial law. *Annals of Probability*, 5(3):404–412, 1977.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

Alexander Smola and Risi Kondor. Kernels and regularization on graphs. In *Conference on Computational Learning Theory*, 2003.

Clifford Spiegelman and Jerome Sacks. Consistent window estimation in nonparametric regression. *The Annals of Statistics*, 8(2):240–246, 1980.

Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under-to over-parametrization affects loss landscape and generalization. *Journal of Physics A: Mathematical and Theoretical*, 52(47):17, 2019.

Karthik Sridharan, Shai Shalev-shwartz, and Nathan Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, 2008.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2): 225–287, 2007.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer, 2008.

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 35(2):575–607, 2007.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

Charles Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977.

Charles Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6): 1348–1360, 1980.

Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhanand Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

James Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36, 1958.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(50):1453–1484, 2005.

Grigorios Tsoumakas, Eleftherios Spyromitros Xioufis, Jozef Vilcek, and Ioannis Vlahavas. MULAN: a java library for multi-label learning. *Journal of Machine Learning Research*, 12(71):2411–2414, 2011.

Alexander Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Alan Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

US Census Bureau. Income and poverty in the United States: 2020, 2021.

Leslie Valiant. Parallelism in comparison problems. *SIAM Journal on Computing*, 4(3):348–355, 1975.

Leslie Valiant. *Probably Approximately Correct*. Basic Books, 2013.

Jesper van Engelen and Holger Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2): 373–440, 2020.

Tim van Erven, Peter Grünwald, Nishant Mehta, Mark Reid, and Robert Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16(54):1793–1861, 2015.

Brendan van Rooyen and Robert Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2017.

Anke van Zuylen, Rajneesh Hegde, Kamal Jain, and David Williamson. Deterministic pivoting algorithms for constrained ranking and clustering problems. In *Symposium on Discrete Algorithms*, 2007.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

Rom Varshamov. Estimate of the number of signals in error correcting codes. *Doklady Akademii Nauk SSSR*, 117:739–741, 1957.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Jakob Verbeek and William Triggs. Scene Segmentation with CRFs Learned from Partially Labeled Images. In *Advances in Neural Information Processing Systems*, 2008.

Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *International Joint Conference on Artificial Intelligence*, 2019.

Stefano Vigogna, Giacomo Meanti, Ernesto De Vito, and Lorenzo Rosasco. Multiclass learning with margin: exponential rates with no bias-variance trade-off. In *International Conference on Machine Learning*, 2022.

Anatoliy Vitushkin. On Hilbert's thirteenth problem. *Proceedings of the USSR Academy of Sciences*, 95(4): 701–704, 1954.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Vladimir Vovk and Glenn Shafer. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008.

Abraham Wald. Statistical decision functions which minimize the maximum risk. *The Annals of Mathematics*, 46(2):265–280, 1945.

Mansfield Tracy Walsorth. *Twenty Questions: A Short Treatise on the Game*. Holt, 1882.

Dan Wang and Yi Shang. A new active labeling method for deep learning. In *International Joint Conference on Neural Networks*, 2014.

Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Conference on Computer Vision and Pattern Recognition*, 2017.

Geoffrey Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics*, 26(4):359–372, 1962.

Yair Weiss. Segmentation using eigenvectors: a unifying view. In *International Conference on Computer Vision*, 1999.

Edmund Whittaker. On the functions which are represented by the expansions of the interpolation theory. *Proceedings of the Royal Society of Edinburgh*, 35:181–194, 1915.

Harold Widom. Asymptotic behavior of the eigenvalues of certain integral equations. *Transactions of the American Mathematical Society*, 109(2), 1963.

Henry Wilbraham. On a certain periodic function. *The Cambridge and Dublin Mathematical Journal*, 3: 108–112, 1848.

Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, 2000.

Ronald Williams, Geoffrey Hinton, and David Rumelhart. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. In *Advances in Neural Information Processing Systems*, 2004.

Greg Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems*, 2021.

Yuhong Yang. Minimax nonparametric classification. I. Rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999.

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. Large-scale multi-label learning with missing labels. In *International Conference on Machine Learning*, 2014.

Vadim Vladimirovich Yurinskii. On an infinite-dimensional version of S. N. Bernstein's inequalities. *Theory of Probability and Its Applications*, 15(1):108–109, 1970.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Mingyuan Zhang and Shivani Agarwal. Bayes consistency vs. $\mathcal{H}$-consistency: The interplay between surrogate loss functions and the scoring function class. In *Advances in Neural Information Processing Systems*, 2020.

Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2019.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, 2003.

Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1):456–463, 2008.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference of Machine Learning*, 2003.

RÉSUMÉ

Les mathématiques appliquées et le calcul nourrissent beaucoup d'espoirs à la suite des succès récents de l'apprentissage supervisé. Dans l'industrie, beaucoup d'ingénieurs cherchent à remplacer leurs anciens paradigmes de pensée par l'apprentissage machine. Étonnamment, ces ingénieurs passent plus de temps à collecter, annoter et nettoyer des données qu'à raffiner des modèles. Ce phénomène motive la problématique de cette thèse: peut-on définir un cadre théorique plus général que l'apprentissage supervisé pour apprendre grâce à des données hétérogènes? Cette question est abordée via le concept de supervision faible, faisant l'hypothèse que le problème que posent les données est leur annotation. On modélise la supervision faible comme l'accès, pour une entrée donnée, non pas d'une sortie claire, mais d'un ensemble de sorties potentielles. On plaide pour l'adoption d'une perspective « optimiste » et l'apprentissage d'une fonction qui vérifie la plupart des observations. Cette perspective nous permet de définir un principe pour lever l'ambiguïté des informations faibles. On discute également de l'importance d'incorporer des techniques sans supervision d'appréhension des données d'entrée dans notre théorie, en particulier de compréhension de la variété sous-jacente via des techniques de diffusion, pour lesquelles on propose un algorithme réaliste afin d'éviter le fléau de la dimension, à l'inverse de ce qui existait jusqu'alors. Enfin, nous nous attaquons à la question de collecte active d'informations faibles, définissant le problème de « catalogage en ligne », où un intendant doit acquérir une maximum d'informations fiables sur ses données sous une contrainte de budget. Entre autres, nous tirons parti du fait que pour obtenir un gradient stochastique et effectuer une descente de gradient, il n'y a pas besoin de supervision totale.

MOTS CLÉS

Apprentissage statistique. Données faiblement supervisées. Acquisition active d'informations partielles.

ABSTRACT

Applied mathematics and machine computations have raised a lot of hope since the recent success of supervised learning. Many practitioners in industries have been trying to switch from their old paradigms to machine learning. Interestingly, those data scientists spend more time scrapping, annotating and cleaning data than fine-tuning models. This thesis is motivated by the following question: can we derive a more generic framework than the one of supervised learning in order to learn from clutter data? This question is approached through the lens of weakly supervised learning, assuming that the bottleneck of data collection lies in annotation. We model weak supervision as giving, rather than a unique target, a set of target candidates. We argue that one should look for an "optimistic" function that matches most of the observations. This allows us to derive a principle to disambiguate partial labels. We also discuss the advantage to incorporate unsupervised learning techniques into our framework, in particular manifold regularization approached through diffusion techniques, for which we derived a new algorithm that scales better with input dimension then the baseline method. Finally, we switch from passive to active weakly supervised learning, introducing the "active labeling" framework, in which a practitioner can query weak information about chosen data. Among others, we leverage the fact that one does not need full information to access stochastic gradients and perform stochastic gradient descent.

KEYWORDS

Statistical learning. Weakly supervised learning; partial supervision. Active labeling; weak queries.