Learning Theory for Kernel Bilevel Optimization

Fares El Khoury, Edouard Pauwels, Samuel Vaiter, Michael Arbel

Learning Theory for Kernel Bilevel Optimization

Fares El Khoury [1] **Edouard Pauwels** [2] **Samuel Vaiter** [3] **Michael Arbel** [1]

Abstract

Bilevel optimization has emerged as a technique for addressing a wide range of machine learning problems that involve an outer objective implicitly determined by the minimizer of an inner problem. In this paper, we investigate the generalization properties for kernel bilevel optimization problems where the inner objective is optimized over a Reproducing Kernel Hilbert Space. This setting enables rich function approximation while providing a foundation for rigorous theoretical analysis. In this context, we establish novel generalization error bounds for the bilevel problem under finite-sample approximation. Our approach adopts a functional perspective, inspired by (Petrulionytė et al., 2024), and leverages tools from empirical process theory and maximal inequalities for degenerate U-processes to derive uniform error bounds. These generalization error estimates allow to characterize the statistical accuracy of gradient-based methods applied to the empirical discretization of the bilevel problem.

1. Introduction

Optimization is a cornerstone of mathematical modeling, machine learning, and decision-making. Among the various frameworks, bilevel optimization stands out due to its hierarchical structure, where one optimization problem, called *outer-level*, is constrained by the solution of another problem, called *inner-level* (Candler & Norton, 1977). This framework was first introduced in the context of economic game theory by von Stackelberg (1934) to model leader-follower interactions, where one agent's decisions depend on the optimal response of another. Over time, bilevel optimization has naturally found applications in a broad spectrum of machine learning fields, including hyper-parameter tuning (Larsen et al., 1996; Bengio, 2000; Franceschi et al., 2018), meta-learning (Bertinetto et al., 2019; Pham et al., 2021), inverse problems (Holler et al., 2018), and reinforcement learning (Hong et al., 2023; Liu et al., 2023), making it a powerful and versatile tool in both theoretical and practical contexts.

The success of bilevel optimization in machine learning naturally raises fundamental questions about the generalization properties of models learned through these procedures, as the number of data samples increases. Several existing works have studied the generalization and convergence of bilevel algorithms under the assumption that the inner-level problem is strongly convex and that its parameters lie in a finite-dimensional space. These include analyses of the convergence of stochastic bilevel optimization algorithms (Arbel & Mairal, 2022a; Dagréou et al., 2022; Ghadimi & Wang, 2018; Ji et al., 2021) and approaches based on algorithmic stability (Bao et al., 2021; Zhang et al., 2024). The strong convexity assumption ensures the uniqueness of the inner-level solution, a key property for stability and convergence analysis in bilevel optimization. Moreover, restricting the inner-level parameters to a finite-dimensional space, instead of possibly richer infinite-dimensional spaces, as in kernel methods, circumvents additional complexities, where the parameter's dimension may grow with the sample size. This dependence on sample size, in non-parametric methods, poses additional challenges, as solutions at different sample sizes are not directly comparable. In contrast, in the finite-dimensional setting, generalization bounds can be derived by quantifying the convergence of finite-sample estimates of the inner-level solution toward the *population solution*, *i.e.*, the solution obtained in the limit of infinite samples, within the same parameter space.

Albeit convenient from a theoretical perspective, having both strong convexity and finite-dimensionality drastically limits expressiveness of the models, effectively restricting them to linear functions. Going beyond linear models requires either relaxing the strong convexity assumption to accommodate more expressive models, such as deep neural networks (Goodfellow et al., 2016), or considering non-parametric bilevel problems, where the inner-level variable lies in an expressive infinite-dimensional function space, such as a Reproducing Kernel Hilbert Space (RKHS) (Schölkopf & Smola, 2002). Early works in bilevel optimization for machine learning followed the latter approach,

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France [2]Toulouse School of Economics, Université Toulouse Capitole, 31080 Toulouse, France [3]CNRS & Université Côte d'Azur, Laboratoire J. A. Dieudonné, 06108 Nice, France. Correspondence to: Fares El Khoury <fares.el-khoury@inria.fr>.

developing bilevel methods for hyper-parameter selection in the context of kernel methods (Keerthi et al., 2006; Kunapuli et al., 2008). These works leverage the *representer theorem* (Schölkopf et al., 2001) to transform the infinite-dimensional problem into a finite-dimensional one with dimension depending on the sample size. However, they do not address how the sample size impacts generalization. Another line of research instead focuses on relaxing the strong convexity assumption, proposing new bilevel algorithms that can handle the loss of convexity (Arbel & Mairal, 2022b; Kwon et al., 2024; Shen & Chen, 2023). Nevertheless, non-convex bilevel optimization is a very hard problem in general (Liu et al., 2021; Arbel & Mairal, 2022b; Bolte et al., 2024), and obtaining strong generalization guarantees in this setting remains out of reach due to the lack of precise control over the inner-level solution. In all cases, learning theory for bilevel problems beyond the strongly convex parametric setting is essentially lacking.

In the present work, we take an initial step towards developing a learning theory that goes beyond the finite-dimensional setting. Specifically, we propose to study *Kernel Bilevel Optimization* (KBO) problems, where the inner objective $L_{in} : \mathbb{R}^d \times \mathcal{H} \to \mathbb{R}$ finds an optimal inner solution $h^\star_\omega$ in a RKHS $\mathcal{H}$ for a given parameter ω in $\mathbb{R}^d$, while the outer objective $L_{out} : \mathbb{R}^d \times \mathcal{H} \to \mathbb{R}$ optimizes the parameter ω over a closed subset $\mathcal{C}$ of $\mathbb{R}^d$, given the inner solution $h^\star_\omega$:

$$\begin{aligned} &\min_{\omega \in \mathcal{C}} \mathcal{F}(\omega) \coloneqq L_{out}(\omega, h^\star_\omega) \\ &\text{s.t.} \quad h^\star_\omega = \arg\min_{h \in \mathcal{H}} L_{in}(\omega, h). \end{aligned} \tag{KBO}$$

In particular, we focus on inner and outer level objectives that are expectation of point-wise losses, a common setting in learning theory. RKHS provides a natural framework to study learning theoretic arguments, and has been instrumental for many fruitful results in pattern recognition and machine learning. They allow to describe very expressive non-linear models with simple and stable algorithms, while enabling a rich statistical analysis and featuring adaptivity to the regularity of the population problem (Shawe-Taylor & Cristianini, 2004; Schölkopf & Smola, 2002; Hofmann et al., 2008). Our choice is also motivated by the relevance of kernel methods, even in the deep learning era. They remain competitive for some prediction problem, for example involving physics (Doumèche et al., 2024; Letizia et al., 2022). Furthermore, the mathematics of kernel methods are useful to describe the limiting behavior of deep network training for very large models (Jacot et al., 2018; Belkin et al., 2018). Indeed, in this limit, the problem becomes (strongly) convex in an infinite-dimensional functional space, simplifying the difficulties of non-convex model parameterization, a major bottleneck in the analysis of such models. This point of view was leveraged by Petrulionytė et al. (2024) who introduced functional bilevel optimization, and our setting is a special case for which the underlying function space is an RKHS. From a practical perspective, our setting is amenable to first-order methods using implicit differentiation techniques (Griewank & Faure, 2003; Bai et al., 2019; Blondel et al., 2022).

Contributions. We leverage empirical process theory and its extension to U-processes (Sherman, 1994) to derive uniform generalization bounds for the value function of (KBO), quantifying the discrepancy between $\mathcal{F}$ and its plug-in estimator $\widehat{\mathcal{F}}$ both in terms of their values and their gradients. This control in terms of gradients is crucial to study first-order optimization methods as the value function $\widehat{\mathcal{F}}$ is typically not convex, and iterative solution methods find approximate critical points ($\nabla\widehat{\mathcal{F}}$ small). Our result relies on an equivalence that we establish between the gradient $\nabla\widehat{\mathcal{F}}$ and a plug-in statistical estimate for $\nabla\mathcal{F}$ that is more amenable to a statistical analysis. We then use our uniform bounds to provide generalization guarantees for gradient descent and projected gradient descent applied to $\nabla\widehat{\mathcal{F}}$. Under specific assumptions, we show convergence rates for sub-optimality measures, depending on sample sizes and the number of algorithmic iterations. This illustrates the relevance of our generalization bounds on one of simplest bilevel optimization algorithms. For large number of algorithmic steps, gradient algorithms on the empirical (KBO) find approximate critical points of the population (KBO) up to a statistical error which we control.

Organization of the Paper. We start in Section 2 with a precise description of the (KBO) problem, application examples, and implicit differentiation in an RKHS. In Section 3, we describe the empirical (KBO) and state our first main result on the gradient of its value function. Our uniform generalization bound for (KBO) is described in Section 4, together with corollaries for gradient descent and projected gradient descent in Section 4.2. Finally, in Section 5, we discuss the strategy of the proof for the main result.

2. Kernel Bilevel Optimization

2.1. Problem Formulation

We consider the kernel bilevel optimization problem in (KBO) with an RKHS $\mathcal{H}$, which is a space of real-valued functions defined on an input space $\mathcal{X} \subset \mathbb{R}^p$ and associated with a reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We are interested, in particular, in (regularized) objectives expressed as expectations of point-wise loss functions, a formulation widely adopted in machine learning as it allows the loss functions to represent the average performance over some data distribution. Specifically, given two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ supported on $\mathcal{X} \times \mathcal{Y}$ for some target space

$\mathcal{Y} \subset \mathbb{R}^q$, we consider objectives of the form:

$$L_{out}(\omega, h) = \mathbb{E}_{\mathbb{Q}}\left[\ell_{out}(\omega, h(x), y)\right],$$
$$L_{in}(\omega, h) = \mathbb{E}_{\mathbb{P}}\left[\ell_{in}(\omega, h(x), y)\right] + \frac{\lambda}{2}\|h\|^2_{\mathcal{H}},$$

where ℓ_{in} and $\ell_{out} : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^q \to \mathbb{R}$ represent the inner and outer point-wise loss functions, $\lambda > 0$ is the regularization parameter which is fixed through this work, and $\|\cdot\|_{\mathcal{H}}$ denotes the norm in the RKHS $\mathcal{H}$. The regularization term in L_{in} is often used in practice to prevent overfitting, by penalizing overly complex models. In our setting, it ensures strong convexity of $h \mapsto L_{in}(\omega, h)$, under mild assumptions on ℓ_{in}, which will be critical to leverage functional implicit differentiation.

Assumptions. Through the paper, we will make the following four assumptions to derive generalization bounds while retaining a simple and modular presentation.

- **(A)** (Boundedness of K). There exists a positive constant $\kappa > 0$ such that $K(x, x) \leq \kappa$, for any $x \in \mathcal{X}$.
- **(B)** (Compactness of $\mathcal{Y}$). The subset $\mathcal{Y}$ of $\mathbb{R}^q$ is compact.
- **(C)** (Regularity of ℓ_{in} and ℓ_{out}). The functions ℓ_{in} and ℓ_{out} are of class C^3 jointly in their first two arguments (ω, v), and their derivatives are jointly continuous in (ω, v, y).
- **(D)** (Convexity of ℓ_{in} in its second argument). For any $(\omega, y) \in \mathbb{R}^d \times \mathbb{R}^q$, the map $v \mapsto \ell_{in}(\omega, v, y)$ is convex.

Assumption **(A)** on the kernel K holds for a wide class of kernels, such as the Gaussian and Matérn kernels (Wendland, 2004), both of which fulfill this assumption with $\kappa = 1$. Assumption **(B)** on the set $\mathcal{Y}$ is a mild assumption that holds in many practical cases, such as classification, where the set $\mathcal{Y}$ is finite, or when $\mathcal{Y} = [0, 1]^q$, where complex data, such as images, can be represented. Assumption **(C)** on the point-wise objectives is a mild regularity assumption. Assumptions **(A)** to **(C)** can be relaxed at the expense of weaker yet more technical assumptions, such as finite moments assumptions on $\mathbb{P}$ and $\mathbb{Q}$, and suitable polynomial growth of the kernel and some partial derivatives of ℓ_{in} and ℓ_{out}. It is also sufficient to require that Assumption **(C)** holds on $\mathcal{U} \times \mathbb{R} \times \mathbb{R}^q$ where $\mathcal{U}$ is an open neighborhood of $\mathcal{C}$ and that Assumption **(D)** holds for any $\omega \in \mathcal{C}$ and $y \in \mathcal{Y}$. We prefer to keep these stronger yet simpler assumptions for clarity. Finally, Assumption **(D)** is essential to ensure the existence and uniqueness of a smooth minimizer $h^\star_\omega$. It is a relatively weak assumption that was recently considered in (Petrulionytė et al., 2024) in the context of functional bilevel optimization and that holds in many cases of interest, as discussed in Section 2.2.

2.2. Examples of KBO in Machine Learning

To illustrate the relevance of (KBO), we consider two examples that highlight its applicability.

Hyper-Parameter Selection under Distribution Shift. In this application, the aim is to select the best hyper-parameters for a machine learning model, *e.g.*, regularization parameters, while accounting for distribution shift between the training and testing data, *i.e.*, when the training and test data distributions are different (Pedregosa, 2016; Franceschi et al., 2018). This can be viewed as an instance of (KBO) when using models in an RKHS. At the inner-level, the model is trained to minimize the regularized training squared error loss, where the hyper-parameter $\omega > 0$ denotes a weight for the data fitting term. At the outer-level, the task is to select the hyper-parameter ω that minimizes the model's performance on the distribution-shifted test. Both inner and outer objectives can thus be formulated as:

$$L_{out}(\omega, h^\star_\omega) = \mathbb{E}_{x,y}\left[\left|h^\star_\omega(x) - y\right|^2\right],$$
$$L_{in}(\omega, h) = \omega\mathbb{E}_{x,y}\left[\left|h(x) - y\right|^2\right] + \frac{1}{2}\|h\|^2_{\mathcal{H}}.$$

This formulation could be used for domain adaptation (Ben-David et al., 2006) or domain generalization (Wang et al., 2022) to choose hyper-parameters that perform well on the distribution-shifted test data.

Instrumental Variable Regression is a technique used to address endogeneity in statistical modeling by leveraging instruments to estimate causal relationships (Newey & Powell, 2003). The goal here is to estimate a function $t \mapsto f_\omega(t)$ parameterized by a vector ω, that satisfies $y = f_\omega(t) + \epsilon$, where $y \in \mathbb{R}$ is the observed outcome, t is a treatment, and ϵ is the error term. The key issue is that t is endogenous, which means that it is correlated with ϵ, *i.e.*, $\mathbb{E}[\epsilon \mid t] \neq 0$, making direct regression inconsistent. Indeed, such correlation leads to biased estimates of $f_\omega(t)$ because the assumption of exogeneity, *i.e.*, independence of t and ϵ, is violated. To resolve this, an instrumental variable x that is uncorrelated with ϵ, *i.e.*, $\mathbb{E}[\epsilon \mid x] = 0$, but correlated with t, can be used to recover the relationship between y and t, without being directly affected by the bias introduced by ϵ using the two stages least squares regression (Singh et al., 2019; Meunier et al., 2024). As shown in (Petrulionytė et al., 2024), this approach can be naturally expressed as a bilevel optimization problem of the form:

$$L_{out}(\omega, h^\star_\omega) = \mathbb{E}_{x,y}\left[\left|h^\star_\omega(x) - y\right|^2\right],$$
$$L_{in}(\omega, h) = \mathbb{E}_{x,t}\left[\left|h(x) - f_\omega(t)\right|^2\right] + \frac{\lambda}{2}\|h\|^2_{\mathcal{H}},$$

where h can be chosen to be in an RKHS to allow flexibility in the estimation while retaining uniqueness of the solution $h^\star_\omega$, a key property in bilevel optimization.

2.3. Implicit Differentiation in an RKHS

A stationarity measure in (KBO) is the gradient $\nabla\mathcal{F}(\omega)$ of the value function $\mathcal{F}$. Nonetheless, evaluating the gradient requires computing the Jacobian $\partial_\omega h^\star_\omega$, which can be viewed as a linear operator from $\mathcal{H}$ to $\mathbb{R}^d$. Indeed $h^\star_\omega$ depends implicitly on ω. A key ingredient for computing the Jacobian $\partial_\omega h^\star_\omega$ is the implicit function theorem (Ioffe & Tihomirov, 1979) which guarantees differentiability of the implicit function $\omega \mapsto h^\star_\omega$ and allows characterizing $\partial_\omega h^\star_\omega$ as the unique solution of a linear system of the form:

$$\partial^2_{\omega,h} L_{in}(\omega, h^\star_\omega) + \partial_\omega h^\star_\omega \partial^2_h L_{in}(\omega, h^\star_\omega) = 0, \qquad (1)$$

where $\partial^2_h L_{in}(\omega, h^\star_\omega)$ is a linear operator from $\mathcal{H}$ to itself representing the partial Hessian of L_{in} w.r.t. h, while $\partial^2_{\omega,h} L_{in}(\omega, h^\star_\omega)$ is an operator from $\mathcal{H}$ to $\mathbb{R}^d$ representing the cross derivatives of L_{in} w.r.t. to ω and h. Applying such result requires $h \mapsto L_{in}(\omega, h)$ to be Fréchet differentiable with gradient map $(\omega, h) \mapsto \partial_h L_{in}(\omega, h)$ jointly Fréchet differentiable and invertible Hessian operator. All these properties are satisfied in our setting under Assumptions **(A)** to **(D)** as shown in Propositions A.1 to A.3 of Appendix A.1. Furthermore, when the outer objective L_{out} is Fréchet differentiable, which is our case under our assumptions (Proposition A.1 of Appendix A.1), then by composition with $\omega \mapsto (\omega, h^\star_\omega)$, the map $\omega \mapsto \mathcal{F}(\omega)$ must also be differentiable with gradient obtained using the chain rule:

$$\nabla\mathcal{F}(\omega) = \partial_\omega L_{out}(\omega, h^\star_\omega) - \partial_\omega h^\star_\omega \partial_h L_{out}(\omega, h^\star_\omega).$$

The above expression for the gradient is intractable as it involves abstract operators. In Proposition 2.1 below, we derive an explicit expression for $\nabla\mathcal{F}(\omega)$ which exploits the particular structure of the objectives L_{in} and L_{out} as expectations of point-wise losses.

Proposition 2.1 (Expression of the total gradient)**.** *Under Assumptions (A) to (D), $\mathcal{F}$ is differentiable on $\mathbb{R}^d$, with gradient $\nabla\mathcal{F}(\omega)$, for any $\omega \in \mathbb{R}^d$, given by:*

$$\begin{aligned}\nabla\mathcal{F}(\omega) =& \mathbb{E}_{\mathbb{Q}}\left[\partial_\omega \ell_{out}(\omega, h^\star_\omega(x), y)\right] \\ &+ \mathbb{E}_{\mathbb{P}}\left[\partial^2_{\omega,v} \ell_{in}(\omega, h^\star_\omega(x), y) a^\star_\omega(x)\right],\end{aligned} \qquad (2)$$

where the adjoint function $a^\star_\omega \in \mathcal{H}$ *is the unique minimizer of a strongly-convex quadratic objective $a \mapsto L_{adj}(\omega, a)$ defined on $\mathcal{H}$ as:*

$$\begin{aligned}L_{adj}(\omega, a) \coloneqq& \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[\partial^2_v \ell_{in}(\omega, h^\star_\omega(x), y) a(x)^2\right] \\ &+ \mathbb{E}_{\mathbb{Q}}[\partial_v \ell_{out}(\omega, h^\star_\omega(x), y) a(x)] + \frac{\lambda}{2}\|a\|^2_{\mathcal{H}},\end{aligned} \qquad (3)$$

where $\partial_\omega \ell_{out}$ and $\partial_v \ell_{out}$ are the first-order partial derivatives of ℓ_{out} w.r.t. ω and v, while $\partial^2_{\omega,v}\ell_{in}$ and $\partial^2_v \ell_{in}$ denote the second-order partial derivatives of ℓ_{in} w.r.t. ω and v.

Proposition 2.1 is proved in Appendix A.1 and relies essentially on proving Bochner's integrability (Diestel & Uhl, 1977, Definition 1, Chapter 2) of some suitable operators on $\mathcal{H}$, and then applying Lebesgue's dominated convergence theorem for Bochner's integral (Diestel & Uhl, 1977, Theorem 3, Chapter 2) to exchange derivatives and expectations. The expression in Proposition 2.1 provides a natural way for approximating $\nabla\mathcal{F}(\omega)$ by estimating all expectations using finite sample averages, as we further discuss in Section 3.

3. Finite Sample Approximation of (KBO)

In this section, we consider an approximation to (KBO) problem when only a finite number of i.i.d. samples $(x_i, y_i)_{1\le i\le n}$ and $(\tilde{x}_j, \tilde{y}_j)_{1\le j\le m}$ from $\mathbb{P}$ and $\mathbb{Q}$ are available. This setting is ubiquitous in machine learning as it allows finding tractable approximate solutions to the original problem. As we are interested in approximately solving (KBO) using gradient methods, our focus here is to derive estimators for both the value function $\mathcal{F}(\omega)$ and its gradient $\nabla\mathcal{F}(\omega)$ for which the generalization properties will later be studied in Section 4.

In Section 3.1, we follow a commonly used approach of first deriving a plug-in estimator $\widehat{\mathcal{F}}$ of the value function, then considering its gradient $\nabla\widehat{\mathcal{F}}(\omega)$ as an approximation to $\nabla\mathcal{F}(\omega)$. Then, in Section 3.2, we show that such approximation is equivalent to a second estimator, more amenable to a statistical analysis, obtained by directly computing a plug-in estimator of $\nabla\mathcal{F}$ based on its expression in Equation (2). Figure 1 summarizes such equivalence.

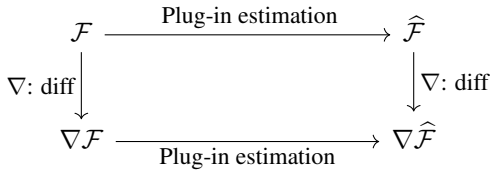

Figure 1. A commutative diagram illustrating that plug-in statistical estimation and differentiation can be interchanged for $\mathcal{F}$ and $\widehat{\mathcal{F}}$ resulting in a single gradient estimator.

3.1. Value Function: Plug-in Estimator and its Gradient

A natural approach for finding approximate solutions to (KBO) is to consider an approximate problem obtained after replacing the objectives L_{in} and L_{out} by their empirical approximations $\widehat{L}_{in}$ and $\widehat{L}_{out}$:

$$\widehat{L}_{out}(\omega, h) \coloneqq \frac{1}{m}\sum_{j=1}^{m} \ell_{out}(\omega, h(\tilde{x}_j), \tilde{y}_j)$$

$$\widehat{L}_{in}(\omega, h) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \ell_{in}(\omega, h(x_i), y_i) + \frac{\lambda}{2}\|h\|^2_{\mathcal{H}}.$$

A plug-in estimator $\omega \mapsto \widehat{\mathcal{F}}(\omega)$ is then obtained by first finding a solution $\hat{h}_\omega$ minimizing $h \mapsto \widehat{L}_{in}(\omega, h)$ that is meant to approximate the optimal inner solution $h^\star_\omega$, and subsequently plugging it into $\widehat{L}_{out}$. This procedure results in the following empirical version of (KBO):

$$\min_{\omega \in \mathcal{C}} \widehat{\mathcal{F}}(\omega) \coloneqq \widehat{L}_{out}\Big(\omega, \hat{h}_\omega\Big)$$
$$\text{s.t.} \quad \hat{h}_\omega = \arg\min_{h \in \mathcal{H}} \widehat{L}_{in}(\omega, h).$$

The inner problem still requires optimizing over a, potentially infinite-dimensional, RKHS. However, its finite sum structure allows equivalently expressing it as a finite-dimensional bilevel optimization, by application of the so called representer theorem (Schölkopf et al., 2001):

$$\min_{\omega \in \mathcal{C}} \widehat{\mathcal{F}}(\omega) \coloneqq \frac{1}{m} \sum_{j=1}^{m} \ell_{out}(\omega, \big(\overline{\mathbf{K}}\, \hat{\gamma}_\omega\big)_j, \tilde{y}_j) \qquad (\widehat{\text{KBO}})$$
$$\text{s.t. } \hat{\gamma}_\omega = \arg\min_{\gamma \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell_{in}(\omega, (\mathbf{K}\,\gamma)_i, y_i) + \frac{\lambda}{2} \gamma^\top \mathbf{K}\, \gamma\,.$$

In the above equation, $\mathbf{K} \in \mathbb{R}^{n \times n}$ and $\overline{\mathbf{K}} \in \mathbb{R}^{m \times n}$ are matrices containing the pairwise kernel similarities between the data points, *i.e.*, $\mathbf{K}_{ij} \coloneqq K(x_i, x_j)$ and $\overline{\mathbf{K}}_{ij} \coloneqq K(\tilde{x}_i, x_j)$, while γ is a parameter vector in $\mathbb{R}^n$ representing the inner-level variables. The optimal solution $\hat{\gamma}_\omega$ enables recovering the prediction function $\hat{h}_\omega$ by linearly combining kernel evaluations at inner-level samples, *i.e.*, $\hat{h}_\omega = \sum_{i=1}^n (\hat{\gamma}_\omega)_i K(x_i, \cdot)$.

The formulation in ($\widehat{\text{KBO}}$) allows deriving an expression for the gradient $\nabla\widehat{\mathcal{F}}(\omega)$ in terms of the Jacobian $\partial_\omega \hat{\gamma}_\omega$ by direct application of the chain rule. Unlike the Jacobian $\partial_\omega h^\star_\omega$ which requires solving an infinite-dimensional linear system given by Equation (1), the Jacobian of $\hat{\gamma}_\omega$ can be obtained as a solution of a finite-dimensional linear system by application of the implicit function theorem (see Proposition B.1 of Appendix B). Hence, ($\widehat{\text{KBO}}$) falls into a class of optimization problems for which a rich body of literature have proposed practical and scalable algorithms, leveraging the expression of $\nabla\widehat{\mathcal{F}}(\omega)$ (Ji et al., 2021; Arbel & Mairal, 2022a; Dagréou et al., 2022). Consequently, solving ($\widehat{\text{KBO}}$) provides a practical way for approximating the solution to the original population problem (KBO) as proposed in several prior works on bilevel optimization involving kernel methods (Keerthi et al., 2006; Kunapuli et al., 2008).

Despite its practical advantage, the above approach yields algorithms that are not directly amenable to a statistical analysis. The key challenge is to be able to control the approximation error between the true gradient $\nabla\mathcal{F}(\omega)$ and its approximation $\nabla\widehat{\mathcal{F}}(\omega)$ as the sample sizes n and m increase. Existing statistical analysis for bilevel optimization, such as (Bao et al., 2021; Zhang et al., 2024), considered objectives in the form of expectations/finite sums of pointwise losses, as we do here. However, they require both inner-level and outer-level parameters to belong to spaces of fixed dimensions that is independent of the sample sizes n and m. That is because these parameters are expected to converge towards some fixed vectors as $n, m \to +\infty$. Unfortunately, these results are not applicable in our case since the inner-level parameter γ has a dimension that increases with the sample size n ($\gamma \in \mathbb{R}^n$) and is not expected to converge towards any well-defined object. Next, we provide an equivalent expression for $\nabla\widehat{\mathcal{F}}(\omega)$ that will be crucial in our statistical analysis in Section 4.

3.2. Plug-in Estimator of the Total Gradient

We consider now an, a priori, different approach for approximating the total gradient $\nabla\mathcal{F}(\omega)$ based on direct plug-in estimation from Equation (2) and show that it recovers the previously introduced estimator $\nabla\widehat{\mathcal{F}}(\omega)$. Such approach consists in replacing all expectations in Equation (2) by empirical averages, then replacing $h^\star_\omega$ and $a^\star_\omega$ by finite sample estimates $\hat{h}_\omega$ and $\hat{a}_\omega$:

$$\begin{aligned}\widehat{\nabla\mathcal{F}}(\omega) =& \frac{1}{m} \sum_{j=1}^{m} \partial_\omega \ell_{out}(\omega, \hat{h}_\omega(\tilde{x}_j), \tilde{y}_j) \\ &+ \frac{1}{n} \sum_{i=1}^{n} \partial^2_{\omega,v} \ell_{in}(\omega, \hat{h}_\omega(x_i), y_i) \hat{a}_\omega(x_i).\end{aligned} \tag{4}$$

Just as in Section 3.1, $h^\star_\omega$ can be estimated by $\hat{h}_\omega$, the minimizer of the empirical objective $h \mapsto \widehat{L}_{in}(\omega, h)$. Similarly, $a^\star_\omega$ can be estimated by minimizing an empirical version $a \mapsto \widehat{L}_{adj}(\omega, a)$ of the adjoint objective L_{adj} given in Equation (3):

$$\begin{aligned}\widehat{L}_{adj}(\omega, a) =& \frac{1}{2n} \sum_{i=1}^{n} \partial^2_v \ell_{in}(\omega, \hat{h}_\omega(x_i), y_i) a^2(x_i) \\ &+ \frac{1}{m} \sum_{j=1}^{m} \partial_v \ell_{out}(\omega, \hat{h}_\omega(\tilde{x}_j), \tilde{y}_j) a(\tilde{x}_j) + \frac{\lambda}{2} \|a\|^2_{\mathcal{H}}.\end{aligned} \tag{5}$$

Both functions $\hat{h}_\omega$ and $\hat{a}_\omega$ can be expressed as linear combinations of kernel evaluations with some given parameters that increase with the sample size n, (see Proposition B.1 for $\hat{h}_\omega$ and Lemma B.2 for $\hat{a}_\omega$, both in Appendix B). However, these parameters are never required for expressing the plug-in estimator $\widehat{\nabla\mathcal{F}}(\omega)$ in Equation (4), since only the function values of $\hat{h}_\omega$ and $\hat{a}_\omega$ are needed for computing it. This property is precisely what makes $\widehat{\nabla\mathcal{F}}(\omega)$ suitable for a statistical analysis, since its estimation error depends on the errors of $\hat{h}_\omega$ and $\hat{a}_\omega$ which always belong to the same space $\mathcal{H}$, regardless of the sample size, and are expected to approach their population counter-parts. This is unlike the expression of $\nabla\widehat{\mathcal{F}}(\omega)$ obtained by implicit differentiation

and which suggests controlling the behavior of the vector $\hat{\gamma}_\omega$.

The next proposition demonstrates that, surprisingly, both estimators $\nabla\widehat{\mathcal{F}}(\omega)$ and $\widehat{\nabla\mathcal{F}}(\omega)$ are precisely equal.

Proposition 3.1. *Under Assumptions (A) to (D), the gradient $\nabla\widehat{\mathcal{F}}(\omega)$ of the plug-in estimator $\widehat{\mathcal{F}}(\omega)$ of $\mathcal{F}(\omega)$ defined in (KBO) is equal to the plug-in estimator $\widehat{\nabla\mathcal{F}}(\omega)$ of the total gradient $\nabla\mathcal{F}(\omega)$ introduced in Equation (4).*

Proposition 3.1 is proved in Appendix B and relies on an application of the representer theorem (Schölkopf et al., 2001) to provide explicit expressions for both estimators in terms of γ_ω, kernel matrices $\mathbf{K}$ and $\overline{\mathbf{K}}$ and partial derivatives of the point-wise objectives ℓ_{in} and ℓ_{out}. Both expressions are then shown to be equal using optimality conditions on the parameters defining $\hat{a}_\omega$. The result in Proposition 3.1 precisely says that the operations of differentiation and plug-in estimation commute in the case of (KBO). Such a commutativity property does not necessarily hold anymore if one considers spaces other than an RKHS as discussed in (Petrulionytė et al., 2024, Appendix F). Next, we leverage the expression of the plug-in estimator $\widehat{\nabla\mathcal{F}}(\omega)$ to provide generalization bounds.

4. Generalization Bounds for (KBO)

In this section, we present the main result of the present work: a maximal inequality controlling how both value function $\mathcal{F}$ and its gradient $\nabla\mathcal{F}$ are well approximated by their empirical counter-parts uniformly over a compact subset Ω of $\mathbb{R}^d$. In Section 4.1, we state the main result and discuss its implications. Then we present the general proof strategy, in Section 5, and discuss possible alternative approaches.

4.1. Maximal Inequalities for (KBO)

The following theorem provides finite sample bounds on the uniform approximation errors on the objective and its gradient in expectation over both inner and upper-level samples.

Theorem 4.1 (Maximal inequalities)**.** *Fix any compact subset Ω of $\mathbb{R}^d$. Under Assumptions (A) to (D), the following maximal inequalities hold:*

$$\mathbb{E}\left[\sup_{\omega\in\Omega}\left|\mathcal{F}(\omega)-\widehat{\mathcal{F}}(\omega)\right|\right]\leq C\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right),$$
$$\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\nabla\mathcal{F}(\omega)-\widehat{\nabla\mathcal{F}}(\omega)\right\|\right]\leq C\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right),$$

where the expectation is taken over the finite samples and C is a constant that depends only Ω, dimension d, regularization parameter λ, κ and local upper-bounds on ℓ_{in}, ℓ_{out} and their partial derivatives over suitable compact set.

Theorem 4.1 states that the estimation error can be decomposed into two contributions each resulting from finite sample approximation of L_{in} and L_{out} with a *parametric rate* of $\frac{1}{\sqrt{n}}$ and $\frac{1}{\sqrt{m}}$ up to a constant factor C. We provide a detailed expression for the constant in Theorem D.7 of Appendix D. The restriction to compact subsets Ω instead of the whole space $\mathbb{R}^d$ allows controlling the complexity of some function classes indexed by the parameter ω. Without additional assumptions on the objectives, we obtain a constant C that grow with the diameter of the subset Ω.

To illustrate the implications of Theorem 4.1, we provide two convergence results for bilevel gradient methods in Section 4.2 and defer the discussion on the general proof strategy to Section 5 with a full proof provided in Appendix D.

4.2. Applications to Empirical Bilevel Gradient Methods

A typical strategy to solve (KBO) is to obtain empirical samples and solve (KBO) using a bilevel optimization algorithm. Note that this is a finite-dimensional problem, and the lower-level is typically strongly convex. Our results allow to provide statistical guarantees for such approaches. We start with the simplest possible gradient algorithm for the unconstrained problem, $\mathcal{C}=\mathbb{R}^d$, given $\eta>0$:

$$\omega_{t+1}=\omega_t-\eta\nabla\widehat{\mathcal{F}}(\omega_t)\qquad t\geq 0. \tag{6}$$

The algorithm requires access to the (almost surely) strongly convex inner-level solution and its derivative which can be obtained using implicit differentiation.

Corollary 4.2 (Generalization for bilevel gradient descent)**.** *Consider Assumptions (A) to (D) and a fixed $\lambda>0$. Assume furthermore that $\mathbf{K}$ in (KBO) is almost surely definite and that there is $c>0$ such that $\inf_{\omega,v,y}\ell_{out}(\omega,v,y)-c\|\omega\|^2>-\infty$. Fix $\omega_0\in\mathbb{R}^d$ and assume that ω_t is given by (6) for all $t\geq 0$. Then there are $\bar{\eta}>0$ and a constant $\bar{c}>0$ such that for any $0<\eta<\bar{\eta}$, and for any $t>0$,*

$$\mathbb{E}\Big[\min_{i=0,\dots,t}\|\nabla\mathcal{F}(\omega_i)\|\Big]\leq\bar{c}\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{t+1}}\right),$$
$$\mathbb{E}\Big[\limsup_{i\to\infty}\|\nabla\mathcal{F}(\omega_i)\|\Big]\leq\bar{c}\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right).$$

Proof. Consider that $\inf_{\omega,v,y}\ell_{out}(\omega,v,y)-c\|\omega\|^2\geq 0$, which entails $\ell_{out}(\omega,v,y)\geq c\|\omega\|^2$ for all v,y. Using Proposition C.1 and setting $B=\sup_{y\in\mathcal{Y}}|\partial_v\ell_{in}(\omega_0,0,y)|$, we have almost surely

$$\widehat{\mathcal{F}}(\omega_0)\leq\max_{|v|\leq B\kappa/\lambda,y\in\mathcal{Y}}\ell_{out}(\omega_0,v,y):=\bar{\ell}.$$

Therefore, for any ω such that $\widehat{\mathcal{F}}(\omega)\leq\widehat{\mathcal{F}}(\omega_0)$, we have $\|\omega\|^2\leq\bar{\ell}/c$. Set Ω the ball of radius $\sqrt{\bar{\ell}/c}$ and center 0. Using the fact that $\widehat{\nabla\mathcal{F}}=\nabla\widehat{\mathcal{F}}$ in Proposition 3.1 and the representation in (4), it is clear from Proposition C.2 and the C^3 assumption that $\nabla\widehat{\mathcal{F}}$ is Lipschitz on Ω with a

deterministic constant L. It follows from standard analysis of the gradient algorithm for nonconvex $\widehat{\mathcal{F}}$ with Lipschitz gradient (see, *e.g.*, (Beck, 2014, Theorems 4.25, 4.26)) that setting $\bar{\eta} = 1/L$, almost surely,

- $\widehat{\mathcal{F}}(\omega_t) \leq \widehat{\mathcal{F}}(\omega_0)$ and $\omega_t \in \Omega$ for all $t \geq 0$.
- $\nabla\widehat{\mathcal{F}}(\omega_t) \to 0$ as $t \to \infty$.
- $\min_{i=0,\dots,t} \left\|\nabla\widehat{\mathcal{F}}(\omega_i)\right\| \leq \bar{c}/\sqrt{t+1}$ for all $t \geq 0$, where $\bar{c}$ is a deterministic constant.

The corollary then follows by combining Proposition 3.1 and the uniform bound in Theorem 4.1. □

Remark 4.3. The additional assumption on ℓ_{out} is a device to ensure a priori almost sure boundedness of the sequence. It is rather mild as it can be enforced by a small perturbation of the form $(\omega, v, y) \mapsto \ell_{out}(\omega, v, y) + c\|\omega\|^2$ assuming $\ell_{out} \geq 0$, which is typical in applications. Any other device ensuring a priori boundedness could be considered. The assumption on $\mathbf{K}$ is satisfied almost surely for most kernels.

Now, considering the constrained problem and assuming $\mathcal{C}$ is convex compact, the projected gradient descent initialized at $\omega_0 \in \mathcal{C}$, given $\eta > 0$, iterates the following recursion:

$$\omega_{t+1} = \Pi_{\mathcal{C}}(\omega_t - \eta\nabla\widehat{\mathcal{F}}(\omega_t))$$

for all $t \geq 0$, where $\Pi_{\mathcal{C}}$ denotes the orthogonal projection on $\mathcal{C}$. The algorithmic requirements are the same as the gradient algorithm, with the addition of the orthogonal projection, which is a cheap operation for basic sets such as balls. For constrained optimization, the optimality condition should take the constraints into account. We consider the gradient mapping, see (Beck, 2017, Section 10.3),

$$\widehat{G}_\eta \colon \omega \mapsto \frac{1}{\eta}\left(\omega - \Pi_{\mathcal{C}}(\omega - \eta\nabla\widehat{\mathcal{F}}(\omega))\right),$$
$$G_\eta \colon \omega \mapsto \frac{1}{\eta}\left(\omega - \Pi_{\mathcal{C}}(\omega - \eta\nabla\mathcal{F}(\omega))\right).$$

This captures stationarity for the recursion, and any local minimum of $\mathcal{F}$ on $\mathcal{C}$ satisfies $G_\eta = 0$ for all $\eta > 0$.

Corollary 4.4 (Generalization for bilevel projected gradient descent)**.** *Consider Assumptions (A) to (D) and a fixed $\lambda > 0$. Assume furthermore that $\mathbf{K}$ in (KBO) is almost surely definite and $\mathcal{C}$ is convex compact. Assume that $\omega_{t+1} = \omega_t - \eta\nabla\widehat{\mathcal{F}}(\omega_t)$ for all $t \geq 0$. Then there are $\bar{\eta} > 0$ and a constant $\bar{c} > 0$ such that for any $0 < \eta < \bar{\eta}$, and for any $t > 0$,*

$$\mathbb{E}\Big[\min_{i=0,\dots,t} \|G_\eta(\omega_i)\|\Big] \leq \bar{c}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{t+1}}\right),$$
$$\mathbb{E}\Big[\limsup_{i\to\infty} \|G_\eta(\omega_i)\|\Big] \leq \bar{c}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right).$$

Proof. We choose $\Omega = \mathcal{C}$. All iterates obviously remain in Ω. Similarly as in the proof of Corollary 4.2, $\nabla\widehat{\mathcal{F}}$ is Lipschitz and $\mathcal{F}$ is bounded on Ω with deterministic constants. It then follows using classical analysis of nonconvex projected gradient algorithm (see, *e.g.*, (Beck, 2017, Theorem 10.15)), that for small η, almost surely $\min_{i=0,\dots,t} \left\|\widehat{G}_\eta(\omega_i)\right\| \leq \bar{c}/\sqrt{t+1}$ for a determinstic $\bar{c} > 0$, and that $\widehat{G}_\eta(\omega_i) \to 0$ as $i \to \infty$. Using the fact that the orthogonal projection is 1-Lipschitz, (see, *e.g.*, (Beck, 2017, Theorem 6.42)), we have for all $\omega \in \Omega$

$$\|\widehat{G}_\eta(\omega) - G_\eta(\omega)\| \leq \|\nabla\widehat{\mathcal{F}}(\omega) - \nabla\mathcal{F}(\omega)\|.$$

The result follows by combining Proposition 3.1 and the uniform bound in Theorem 4.1. □

5. General Proof Strategy for Theorem 4.1

The main strategy behind the proof of Theorem 4.1 in Appendix D consists in three steps: (step 1) obtaining a point-wise error decomposition of the errors into manageable error terms that holds almost surely for any $\omega \in \Omega$, then applying maximal inequalities to suitable empirical processes (step 2) and some degenerate U-processes (step 3) to control each of these terms. The final error bounds are obtained by combining all these bounds as shown in Appendix D.3.

Step 1: Point-wise error decomposition. A main challenge in controlling the errors in Theorem 4.1 is the non-linear dependence of both estimators $\widehat{\mathcal{F}}(\omega)$ and $\widehat{\nabla\mathcal{F}}(\omega)$ on the empirical distributions, as they are obtained via a plug-in procedure. We address this challenge by breaking down the error into components based on the discrepancies between expected values and their empirical counterparts of individual point-wise losses and their derivatives, evaluated at the optimal solution $h^\star_\omega$. Specifically, we denote by δ^{out}_ω and δ^{in}_ω the errors on the objectives defined as:

$$\delta^{out}_\omega := \left|L_{out}(\omega, h^\star_\omega) - \widehat{L}_{out}(\omega, h^\star_\omega)\right|,$$
$$\delta^{in}_\omega := \left|L_{in}(\omega, h^\star_\omega) - \widehat{L}_{in}(\omega, h^\star_\omega)\right|.$$

Additionally, we quantify the errors between the partial derivatives of these objectives and their empirical counterparts. To simplify our proof outline, we slightly abuse notation by denoting $\partial_h\delta^{out}_\omega$, $\partial_h\delta^{in}_\omega$, $\partial_\omega\delta^{out}_\omega$, $\partial^2_{\omega,h}\delta^{in}_\omega$ and $\partial^2_h\delta^{in}_\omega$ to refer to these errors in terms of partial derivatives. For instance, $\partial_h\delta^{out}_\omega$ is defined as $\left\|\partial_h L_{out}(\omega, h^\star_\omega) - \partial_h\widehat{L}_{out}(\omega, h^\star_\omega)\right\|_{\mathcal{H}}$, with similar definitions for the other error terms that can be found in Appendix D.1. The next proposition formalizes the above discussions.

Proposition 5.1 (Approximation bounds)**.** *Let Ω be an arbitrary compact subset of $\mathbb{R}^d$. Under Assumptions (A) to (D),*

the following holds almost surely, for any $\omega \in \Omega$:

$$\left|\mathcal{F}(\omega) - \widehat{\mathcal{F}}(\omega)\right| \leq C\big(\delta_\omega^{out} + \partial_h \delta_\omega^{in}\big),$$
$$\left\|\nabla\mathcal{F}(\omega) - \widehat{\nabla\mathcal{F}}(\omega)\right\| \leq C(\partial_\omega \delta_\omega^{out} + \partial_h \delta_\omega^{out} + \partial_h^2 \delta_\omega^{in} + \partial_{\omega,h}^2 \delta_\omega^{in} + \partial_h \delta_\omega^{in}),$$

where C is a positive constant that depends only Ω, dimension d, regularization parameter λ, κ and local upper-bounds on ℓ_{in}, ℓ_{out} and their partial derivatives over suitable compact set.

A more detailed version of this proposition, including the exact constants, along with its proof are provided in Proposition D.4 of Appendix D.1. The error terms arising in the above decomposition are more amenable to a statistical analysis using empirical process theory as we discuss next.

Step 2: Maximal inequalities for empirical processes. Some of the error terms, namely δ_ω^{out} and $\partial_\omega \delta_\omega^{out}$, can be controlled directly using empirical process theory. For instance, δ_ω^{out} is associated to the family of random functions $\sqrt{m}\Big(L_{out}(\omega, h_\omega^\star) - \widehat{L}_{out}(\omega, h_\omega^\star)\Big)_{\omega\in\Omega}$, which defines an empirical process, a scaled and centered empirical average of real-valued functions indexed by the parameter ω. Thus, provided that suitable estimates of the complexity of the class are available (as measured by its packing number in Proposition E.1 of Appendix E), which is easy to obtain in our setting, we show in Proposition D.5 of Appendix D.2 that a maximal inequality of the form below follows from classical results on empirical processes:

$$\mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega} \overbrace{\left|L_{out}(\omega, h_\omega^\star) - \widehat{L}_{out}(\omega, h_\omega^\star)\right|}^{\delta_\omega^{out}}\right] \leq \frac{C}{\sqrt{m}}.$$

Step 3: Maximal inequalities for degenerate U-processes. Unfortunately, the approach in step 2 cannot be readily used for the remaining error terms involving partial derivatives w.r.t. h ($\partial_h \delta_\omega^{out}$, $\partial_h \delta_\omega^{in}$, $\partial_{\omega,h}^2 \delta_\omega^{in}$ and $\partial_h^2 \delta_\omega^{in}$). These error terms are associated to processes that are not real-valued anymore but take values in an infinite-dimensional space. While the recent work in (Park & Muandet, 2023) develops an empirical process theory for functions taking values in a vector space, the provided complexity estimates would result in unfavorable dependence on the sample size. Instead, we leverage the structure of the RKHS to control these errors using maximal inequalities for suitable degenerate U-processes of order 2 indexed by the parameter ω and for which such inequalities were provided in the seminal works of Sherman (1994); Nolan & Pollard (1987). U-processes of order 2 are generalization of empirical processes that involve empirical averages of real-valued functions which depend on pairs of samples, instead of a single one as in empirical processes. These functions arise, in our case, when taking the square of the error term $\partial_h \delta_\omega^{out}$, for instance, and exploiting the reproducing property in the RKHS. This approach, presented in Proposition D.6 of Appendix D.2, allows us to obtain maximal inequalities of the form:

$$\mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega} \partial_h \delta_\omega^{out}\right] \leq \mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega} \left(\partial_h \delta_\omega^{out}\right)^2\right]^{\frac{1}{2}} \leq \frac{C}{\sqrt{m}}.$$

Combining the maximal inequalities from step 2 and 3 with the error decomposition from step 1 allows to obtain the result of Theorem 4.1.

Discussion. Alternative approaches to U-processes could be used to derive generalization bounds, although these would result in degraded sample dependence. Specifically, one could employ a variational formulation of the RKHS norm appearing in some of the error terms, such as $\partial_h \delta_\omega^{out}$, to express them as the error of some real-valued empirical process to which standard results could be applied. However, this comes at the cost of considering processes indexed not only by the finite-dimensional parameter ω, but also by functions in a unit RKHS ball, in contrast with our approach based on U-processes indexed by finite-dimensional vectors. Consequently, these families have much larger complexities as measured by their covering/packing numbers (Yang et al., 2020, Lemma D.2), which directly impacts the generalization rate. Our proposed approach bypasses this challenge by using real-valued U-processes indexed by finite-dimensional parameters, at the expense of employing a more general empirical process theory for degenerate U-processes (Sherman, 1994).

6. Conclusion and Perspectives

In this work, we established the first statistical generalization bounds for (KBO). This paper is a first step in providing generalization results for bilevel gradient-based methods in a non-parametric setting. We showcased the applicability of these bounds on the simplest algorithms and expect that our results can be readily extended to state-of-the-art methods such as those discussed in (Arbel & Mairal, 2022b; Ghadimi & Wang, 2018; Chen et al., 2021; Dagréou et al., 2022). The results concerning U-processes, however, are derived under the restrictive assumption of i.i.d. data. Relaxing such assumption to handle Markovian data is a promising direction for future work, as it would allow considering reinforcement learning applications where bilevel optimization methods have shown promising results (Nikishin et al., 2022).

Acknowledgements

This work was supported by the ANR project BONSAI (grant ANR-23-CE23-0012-01). EP thanks AI Interdisci-

plinary Institute ANITI funding, through the French "Investments for the Future – PIA3" program under the grant agreement ANR-19-PI3A0004, Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant numbers FA8655-22-1-7012. EP thanks TSE-P, acknowledges support from ANR Chess, grant ANR-17-EURE-0010, ANR Regulia, support from IUF. EP and SV acknowledge support from ANR MAD. SV thanks PEPR PDE-AI (ANR-23-PEIA-0004) and the chair 3IA Côte d'Azur BOGL.

References

Arbel, M. and Mairal, J. Amortized implicit differentiation for stochastic bilevel optimization. *International Conference on Learning Representations (ICLR)*, 2022a.

Arbel, M. and Mairal, J. Non-convex bilevel games with critical point selection maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b.

Bai, S., Kolter, J. Z., and Koltun, V. Deep equilibrium models. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Bao, F., Wu, G., Li, C., Zhu, J., and Zhang, B. Stability and generalization of bilevel programming in hyperparameter optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:4529–4541, 2021.

Beck, A. *Introduction to nonlinear optimization: Theory, algorithms, and applications with MATLAB*. SIAM, 2014.

Beck, A. *First-order methods in optimization*. SIAM, 2017.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning (ICML)*, pp. 541–549. PMLR, 2018.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems (NIPS)*, 19, 2006.

Bengio, Y. Gradient-based optimization of hyperparameters. *Neural Computation*, 12(8):1889–1900, 2000. ISSN 0899-7667, 1530-888X. doi: 10.1162/089976600300015187.

Bertinetto, L., Henriques, J. F., Torr, P., and Vedaldi, A. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=HyxnZh0ct7.

Blondel, M., Berthet, Q., Cuturi, M., Frostig, R., Hoyer, S., Llinares-López, F., Pedregosa, F., and Vert, J.-P. Efficient and modular implicit differentiation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 5230–5242, 2022.

Bolte, J., Lê, T., Édouard Pauwels, and Vaiter, S. Geometric and computational hardness of bilevel programming. *Preprint*, 2024. URL https://arxiv.org/abs/2407.12372.

Candler, W. and Norton, R. *Multi-Level Programming and Development Policy*, volume 1. World Bank, 1977.

Chen, T., Sun, Y., and Yin, W. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:25294–25307, 2021.

Dagréou, M., Ablin, P., Vaiter, S., and Moreau, T. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Diestel, J. and Uhl, J. J. *Vector Measures*, volume 15 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1977. ISBN 978-0-8218-1515-1.

Doumèche, N., Bach, F., Biau, G., and Boyer, C. Physics-informed kernel learning. *arXiv preprint arXiv:2409.13786*, 2024.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning (ICML)*, pp. 1568–1577. PMLR, 2018. URL https://proceedings.mlr.press/v80/franceschi18a.html.

Ghadimi, S. and Wang, M. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Griewank, A. and Faure, C. Piggyback differentiation and optimization. In *Large-scale PDE-constrained optimization*, pp. 148–164. Springer, 2003.

Hofmann, T., Schölkopf, B., and Smola, A. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

Holler, G., Kunisch, K., and Barnard, R. C. A bilevel approach for parameter learning in inverse problems. *Inverse Problems*, 34(11):115012, 2018. doi: 10.1088/1361-6420/aae473.

Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Ioffe, A. D. and Tihomirov, V. M. *Theory of Extremal Problems*. Series: Studies in Mathematics and its Applications 6. Elsevier, 1979.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 31, 2018.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning (ICML)*, pp. 4882–4892. PMLR, 2021.

Keerthi, S., Sindhwani, V., and Chapelle, O. An Efficient Method for Gradient-Based Adaptation of Hyperparameters in SVM Models. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

Kosorok, M. R. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer, New York, 2008. URL https://doi.org/10.1007/978-0-387-74978-5.

Kunapuli, G., Bennett, K. P., Hu, J., and Pang, J.-S. Classification model selection via bilevel programming. *Optimization Methods & Software*, 23(4):475–489, 2008.

Kwon, J., Kwon, D., Wright, S., and Nowak, R. D. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. In *International Conference on Learning Representations (ICLR)*, 2024.

Larsen, J., Hansen, L., Svarer, C., and Ohlsson, M. Design and regularization of neural networks: The optimal use of a validation set. In *Neural Networks for Signal Processing VI. Proceedings of the 1996 IEEE Signal Processing Society Workshop*, pp. 62–71, Kyoto, Japan, 1996. IEEE. doi: 10.1109/NNSP.1996.548300.

Letizia, M., Losapio, G., Rando, M., Grosso, G., Wulzer, A., Pierini, M., Zanetti, M., and Rosasco, L. Learning new physics efficiently with nonparametric methods. *The European Physical Journal C*, 82(10):879, 2022.

Liu, R., Liu, Y., Zeng, S., and Zhang, J. Towards gradient-based bilevel optimization with non-convex followers and beyond. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Liu, R., Liu, X., Zeng, S., Zhang, J., and Zhang, Y. Value-function-based sequential minimization for bi-level optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Meunier, D., Li, Z., Christensen, T., and Gretton, A. Nonparametric Instrumental Regression via Kernel Methods is Minimax Optimal. *arXiv preprint arXiv:2411.19653*, 2024.

Newey, W. K. and Powell, J. L. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5): 1565–1578, 2003.

Nikishin, E., Abachi, R., Agarwal, R., and Bacon, P.-L. Control-oriented model-based reinforcement learning with implicit differentiation. *AAAI Conference on Artificial Intelligence*, 2022.

Nolan, D. and Pollard, D. U-processes: rates of convergence. *The Annals of Statistics*, pp. 780–799, 1987.

Park, J. and Muandet, K. Towards empirical process theory for vector-valued functions: Metric entropy of smooth function classes. In *International Conference on Algorithmic Learning Theory*, pp. 1216–1260. PMLR, 2023.

Pedregosa, F. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning (ICML)*, pp. 737–746. PMLR, 2016.

Petrulionytė, I., Mairal, J., and Arbel, M. Functional Bilevel Optimization for Machine Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Pham, Q., Liu, C., Sahoo, D., and Steven, H. Contextual transformation networks for online continual learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Rudin, W. *Real and Complex Analysis*. McGraw-Hill, New York, 3rd edition, 1987. ISBN 978-0070542341.

Schölkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. The MIT Press, 2002.

Schölkopf, B., Herbrich, R., and Smola, A. J. A Generalized Representer Theorem. In Helmbold, D. and Williamson, B. (eds.), *Computational Learning Theory*, pp. 416–426, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-44581-4.

Shawe-Taylor, J. and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004.

Shen, H. and Chen, T. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning (ICML)*, pp. 30992–31015. PMLR, 2023.

Sherman, R. P. Maximal Inequalities for Degenerate U-Processes with Applications to Optimization Estimators. *The Annals of Statistics*, 22(1):439 – 459, 1994. URL https://doi.org/10.1214/aos/1176325377.

Singh, R., Sahani, M., and Gretton, A. Kernel Instrumental Variable Regression. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Van der Vaart, A. W. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

von Stackelberg, H. *Marktform und Gleichgewicht*. Die Handelsblatt-Bibliothek "Klassiker der Nationalökonomie". J. Springer, 1934.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Philip, S. Y. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.

Wendland, H. *Scattered Data Approximation*, volume 17. Cambridge University Press, 2004.

Yang, Z., Jin, C., Wang, Z., Wang, M., and Jordan, M. I. On Function Approximation in Reinforcement Learning: Optimism in the Face of Large State Spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Zhang, X., Chen, H., Gu, B., Gong, T., and Zheng, F. Fine-grained analysis of stability and generalization for stochastic bilevel optimization. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5508–5516, 2024.

Appendices

Roadmap. We begin by presenting and establishing regularity properties of the objective functions in Appendix A. In Appendix B, we introduce the gradient estimators. Appendix C is dedicated to proving the boundedness and Lipschitz continuity of $h^\star_\omega$ and $\hat{h}_\omega$, along with local boundedness and Lipschitz properties of ℓ_{in}, ℓ_{out} and their derivatives. The generalization results are provided in Appendix D. In Appendix E, we establish maximal inequalities for bounded and Lipschitz families of functions. Differentiability properties of the objectives are studied in Appendix F. Finally, Appendix G contains auxiliary technical lemmas used throughout the proofs.

Notations. $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$, $\|\cdot\|_{\mathcal{H}}$ denotes the norm in the RKHS $\mathcal{H}$, $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm, and $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert-Schmidt norm. $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ denotes the inner product on $\mathcal{H}$, and $\langle\cdot,\cdot\rangle_{\mathrm{HS}}$ denotes the Hilbert-Schmidt inner product. $K(x,\cdot)$ denotes the feature map, for any $x \in \mathcal{X}$. For any two normed spaces E and F, $\mathcal{L}(E,F)$ denotes the space of continuous linear operators from E to F. For any two probability distributions $\mathcal{P}$ and $\mathcal{Q}$, $\mathcal{P}\otimes\mathcal{Q}$ denotes the product measure of $\mathcal{P}$ and $\mathcal{Q}$. Given two Hilbert spaces $(H_1, \langle\cdot,\cdot\rangle_{H_1})$ and $(H_2, \langle\cdot,\cdot\rangle_{H_2})$, the tensor product of $u \in H_1$ and $v \in H_2$, denoted by $u \otimes v$, is an operator from H_2 to H_1 defined, for any $e \in H_2$, as $(u\otimes v)\, e = u \langle v, e\rangle_{H_2}$. For any $v_1,\dots,v_n \in \mathbb{R}$, $\mathbf{diag}(v_1,\dots,v_n) \in \mathbb{R}^{n\times n}$ denotes a diagonal matrix of size $n\times n$, where the diagonal entries are $v_1,\dots,v_n$ and all the off-diagonal entries are 0. $\mathbb{1}_m$ denotes a vector of size m where all entries are 1. $\mathbb{1}_{n\times n}$ denotes an $n\times n$ matrix where all entries are 1. For any vector space V over $\mathbb{R}$, Id_V denotes the identity operator on V. Given a compact set $\mathcal{K}$, $\mathrm{diam}(\mathcal{K})$ denotes its diameter. $v^\top$ denotes the transpose of either a vector or a matrix, depending on the context.

A. Regularity and Differentiability Results

A.1. Regularity of the objectives

The following propositions establish differentiability of considered objectives. We defer their proof to Appendix F.

Proposition A.1 (Differentiability of L_{in} and L_{out})**.** *Under Assumptions (A) to (C), for any $(\omega,h) \in \mathbb{R}^d \times \mathcal{H}$, the functions L_{in}, L_{out} admit finite values at (ω,h), are jointly differentiable in (ω,h), with gradient given by:*

$$\begin{aligned}
\partial_\omega L_{out}(\omega,h) &= \mathbb{E}_{\mathbb{Q}}\left[\partial_\omega \ell_{out}(\omega,h(x),y)\right] \in \mathbb{R}^d, & \partial_h L_{out}(\omega,h) &= \mathbb{E}_{\mathbb{Q}}\left[\partial_v \ell_{out}(\omega,h(x),y) K(x,\cdot)\right] \in \mathcal{H},\\
\partial_\omega L_{in}(\omega,h) &= \mathbb{E}_{\mathbb{P}}\left[\partial_\omega \ell_{in}(\omega,h(x),y)\right] \in \mathbb{R}^d, & \partial_h L_{in}(\omega,h) &= \mathbb{E}_{\mathbb{P}}\left[\partial_v \ell_{in}(\omega,h(x),y) K(x,\cdot)\right] + \lambda h \in \mathcal{H}.
\end{aligned}$$

Similarly, the empirical estimates $\widehat{L}_{in}$ and $\widehat{L}_{out}$ admit finite values, and are differentiable with gradients admitting similar expressions as above with $\mathbb{P}$ and $\mathbb{Q}$ replaced by their empirical estimates $\hat{\mathbb{P}}_n$ and $\hat{\mathbb{Q}}_m$.

Proposition A.2 (Differentiability of $\partial_h L_{in}$)**.** *Under Assumptions (A) to (C), for any $(\omega,h) \in \mathbb{R}^d \times \mathcal{H}$, the functions $(\omega,h) \mapsto \partial_h L_{in}(\omega,h)$ is differentiable with partial derivatives given by:*

$$\begin{aligned}
\partial^2_{\omega,h} L_{in}(\omega,h) &= \mathbb{E}_{\mathbb{P}}\left[\partial^2_{\omega,v} \ell_{in}(\omega,h(x),y) K(x,\cdot)\right] \in \mathcal{L}(\mathcal{H},\mathbb{R}^d),\\
\partial^2_h L_{in}(\omega,h) &= \mathbb{E}_{\mathbb{P}}\left[\partial^2_v \ell_{in}(\omega,h(x),y) K(x,\cdot)\otimes K(x,\cdot)\right] + \lambda\,\mathrm{Id}_{\mathcal{H}} \in \mathcal{L}(\mathcal{H},\mathcal{H}).
\end{aligned}$$

Moreover, for any $\omega \in \mathbb{R}^d$ and $h \in \mathcal{H}$, the operators $\partial^2_{\omega,h} L_{in}(\omega,h)$ and $\partial^2_h L_{in}(\omega,h) - \lambda\,\mathrm{Id}_{\mathcal{H}}$ are Hilbert-Schmidt, i.e., bounded operators with finite Hilbert-Schmidt norm. The same conclusions hold for the empirical estimate $(\omega,h) \mapsto \partial_h \widehat{L}_{in}(\omega,h)$ with partial derivatives admitting similar expressions as above with $\mathbb{P}$ replaced by its empirical estimate $\hat{\mathbb{P}}_n$.

Proposition A.3 (Strong convexity of the inner objective in its second variable and invertibility of the Hessians)**.** *Under Assumptions (A) to (D), $h \mapsto L_{in}(\omega,h)$ and $h \mapsto \widehat{L}_{in}(\omega,h)$ are λ-strongly convex for any $\omega \in \mathbb{R}^d$. Moreover, for any $\omega \in \mathbb{R}^d$ and $h \in \mathcal{H}$, the Hessian operators $\partial^2_h L_{in}(\omega,h)$ and $\partial^2_h \widehat{L}_{in}(\omega,h)$ are invertible with their operator norm bounded by $\frac{1}{\lambda}$.*

Proof. By Assumption (D), we know that $v \mapsto \ell_{in}(\omega,v,y)$ is convex for any $\omega \in \mathbb{R}^d$ and $y \in \mathcal{Y}$. Moreover, by Proposition A.1, $(x,y) \mapsto \ell_{in}(\omega,h(x),y)$ is integrable for any $\omega \in \mathbb{R}^d$ and $h \in \mathcal{H}$. Consequently, by integration, we directly deduce that $h \mapsto \mathbb{E}_{\mathbb{P}}[\ell_{in}(\omega,h(x),y)]$ is convex for any $\omega \in \mathbb{R}^d$. Finally, $h \mapsto L_{in}(\omega,h) \coloneqq \mathbb{E}_{\mathbb{P}}[\ell_{in}(\omega,h(x),y)] + \frac{\lambda}{2}\|h\|^2_{\mathcal{H}}$ must be λ-strongly convex, for any $\omega \in \mathbb{R}^d$, as a sum of a convex function and a λ-strongly convex function. Similarly, we deduce that $h \mapsto \widehat{L}_{in}(\omega,h)$ is λ-strongly convex, for any $\omega \in \mathbb{R}^d$. Invertibility follows from the expression of the Hessian operator in Proposition A.2 □

A.2. Differentiability of the value function

Proposition A.4 (Total functional gradient $\nabla\mathcal{F}$)**.** *Assume Assumptions (A), (C) and (D) hold. For any $\omega \in \mathbb{R}^d$, the total functional gradient $\nabla\mathcal{F}(\omega)$ satisfies:*

$$\nabla\mathcal{F}(\omega) = \partial_\omega L_{out}(\omega, h^\star_\omega) + \partial^2_{\omega,h} L_{in}(\omega, h^\star_\omega) a^\star_\omega \in \mathbb{R}^d, \tag{7}$$

where $a^\star_\omega$ is the unique minimizer of the following quadratic objective:

$$L_{adj}(\omega, a) \coloneqq \frac{1}{2}\langle a, H_\omega a\rangle_{\mathcal{H}} + \langle a, d_\omega\rangle_{\mathcal{H}}, \quad \text{for any } a \in \mathcal{H}, \tag{8}$$

with $H_\omega \coloneqq \partial^2_h L_{in}(\omega, h^\star_\omega) : \mathcal{H} \to \mathcal{H}$ being the Hessian operator and $d_\omega \coloneqq \partial_h L_{out}(\omega, h^\star_\omega) \in \mathcal{H}$.

Proof. By applying Propositions A.1 and A.3, we know that $h \mapsto L_{in}(\omega, h)$ has finite values, is λ-strongly convex and Fréchet differentiable. Moreover, by Proposition A.2, $\partial_h L_{in}$ is Fréchet differentiable on $\mathbb{R}^d \times \mathcal{H}$, and, a fortiori, Hadamard differentiable. Therefore, by the functional implicit differentiation theorem (Ioffe & Tihomirov, 1979), we deduce that the map $\omega \mapsto h^\star_\omega$ is uniquely defined and is Fréchet differentiable with Jacobian $\partial_\omega h^\star_\omega$ solving the following linear system for any $\omega \in \mathbb{R}^d$:

$$\partial^2_{\omega,h} L_{in}(\omega, h^\star_\omega) + \partial_\omega h^\star_\omega \partial^2_h L_{in}(\omega, h^\star_\omega) = 0.$$

Using that $\partial^2_h L_{in}(\omega, h^\star_\omega)$ is invertible by Proposition A.3, we can express $\partial_\omega h^\star_\omega$ as:

$$\partial_\omega h^\star_\omega = -\partial^2_{\omega,h} L_{in}(\omega, h^\star_\omega) \left(\partial^2_h L_{in}(\omega, h^\star_\omega)\right)^{-1}.$$

Furthermore, L_{out} is jointly Fréchet differentiable by application of Proposition A.1, so that $\omega \mapsto \mathcal{F}(\omega)$ is also differentiable by composition of the functions $(\omega, h) \mapsto L_{out}(\omega, h)$ and $\omega \mapsto (\omega, h^\star_\omega)$. For a given $\omega \in \mathbb{R}^d$, the gradient of $\mathcal{F}$ is then given by the chain rule:

$$\nabla\mathcal{F}(\omega) = \partial_\omega L_{out}(\omega, h^\star_\omega) + \partial_\omega h^\star_\omega \partial_h L_{out}(\omega, h^\star_\omega). \tag{9}$$

Substituting the expression of $\partial_\omega h^\star_\omega$ into Equation (9) yields:

$$\nabla\mathcal{F}(\omega) = \partial_\omega L_{out}(\omega, h^\star_\omega) - \partial^2_{\omega,h} L_{in}(\omega, h^\star_\omega) \left(\partial^2_h L_{in}(\omega, h^\star_\omega)\right)^{-1} \partial_h L_{out}(\omega, h^\star_\omega).$$

To conclude, it suffices to notice that the function $a^\star_\omega$ appearing in Equation (7) must be equal to $-H^{-1}_\omega d_\omega$. Indeed, $a^\star_\omega$ is defined as the minimizer of the quadratic objective $L_{adj}(\omega, a)$ in Equation (8) which is strongly convex since the Hessian operator is lower-bounded by $\lambda \operatorname{Id}_{\mathcal{H}}$. Consequently, the minimizer $a^\star_\omega$ exists and is uniquely characterized by the optimality condition:

$$H_\omega a^\star_\omega + d_\omega = 0.$$

The above equation is a linear system in $\mathcal{H}$ whose solution is given by $a^\star_\omega \coloneqq -H^{-1}_\omega d_\omega$. □

B. Gradient Estimators

Proposition B.1 (Expression of $\nabla\widehat{\mathcal{F}}(\omega)$ by implicit differentiation)**.** *Under Assumptions (A) to (D), for any $\omega \in \mathbb{R}^d$, the gradient $\nabla\widehat{\mathcal{F}}(\omega)$ of the discretized functional bilevel optimization problem (KBO) is given by:*

$$\nabla\widehat{\mathcal{F}}(\omega) = \frac{1}{m}\mathbf{D}^{\mathbf{out}}_{\omega}\,\mathbb{1}_m - \frac{1}{m}\mathbf{D}^{\mathbf{in}}_{\omega,\mathbf{v}}\,\mathbf{M}^{-1}\,\mathbf{u} \in \mathbb{R}^d,$$

where $\mathbf{K}$ and $\overline{\mathbf{K}}$ are the Gram matrices in $\mathbb{R}^{n\times n}$ and $\mathbb{R}^{m\times n}$ with entries given by $\mathbf{K}_{ij} \coloneqq K(x_i, x_j)$ and $\mathbf{K}_{ij} \coloneqq K(\tilde{x}_j, x_j)$, and $\mathbf{M}$, $\mathbf{u}$, $\mathbf{D}^{\mathbf{out}}_{\mathbf{v}}$, $\mathbf{D}^{\mathbf{in}}_{\mathbf{v},\mathbf{v}}$, $\mathbf{D}^{\mathbf{out}}_{\omega}$ and $\mathbf{D}^{\mathbf{in}}_{\omega,\mathbf{v}}$ are defined as:

$$\mathbf{M} \coloneqq \mathbf{K}\,\mathbf{D}^{\mathbf{in}}_{\mathbf{v},\mathbf{v}} + n\lambda\mathbb{1}_{n\times n} \in \mathbb{R}^{n\times n}, \qquad \mathbf{u} \coloneqq \overline{\mathbf{K}}^\top \mathbf{D}^{\mathbf{out}}_{\mathbf{v}} \in \mathbb{R}^n,$$

$$\mathbf{D}^{\mathbf{out}}_{\mathbf{v}} \coloneqq \left(\partial_v \ell_{out}\Big(\omega, \hat{h}_\omega(\tilde{x}_j), \tilde{y}_j\Big)\right)_{1\le j\le m} \in \mathbb{R}^m, \qquad \mathbf{D}^{\mathbf{out}}_{\omega} \coloneqq \Big(\partial_\omega \ell_{out}(\omega, \hat{h}_\omega(\tilde{x}_j), \tilde{y}_j)\Big)_{1\le j\le m} \in \mathbb{R}^{d\times m},$$

$$\mathbf{D}^{\mathbf{in}}_{\mathbf{v},\mathbf{v}} \coloneqq \mathbf{diag}\left(\Big(\partial^2_v \ell_{in}(\omega, \hat{h}_\omega(x_i), y_i)\Big)_{1\le i\le n}\right) \in \mathbb{R}^{n\times n}, \qquad \mathbf{D}^{\mathbf{in}}_{\omega,\mathbf{v}} \coloneqq \Big(\partial^2_{\omega,v} \ell_{in}(\omega, \hat{h}_\omega(x_i), y_i)\Big)_{1\le i\le n} \in \mathbb{R}^{d\times n}.$$

Proof. Let $\omega \in \mathbb{R}^d$. Recall the expression of $\widehat{\mathcal{F}}(\omega)$:

$$\widehat{\mathcal{F}}(\omega) \coloneqq \frac{1}{m}\sum_{j=1}^{m} \ell_{out}\Big(\omega, \hat{h}_\omega(\tilde{x}_j), \tilde{y}_j\Big)$$
$$\text{s.t.} \quad \hat{h}_\omega = \operatorname*{arg\,min}_{h\in\mathcal{H}} \widehat{L}_{in}(\omega, h) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \ell_{in}(\omega, h(x_i), y_i) + \frac{\lambda}{2}\|h\|_{\mathcal{H}}^2.$$

By the representer theorem, it is easy to see that $\hat{h}_\omega$ must be a linear combination of $K(x_1, \cdot), \dots, K(x_n, \cdot)$:

$$\hat{h}_\omega = \sum_{i=1}^{n} (\hat{\gamma}_\omega)_i K(x_i, \cdot). \tag{10}$$

Hence, finding $\hat{h}_\omega$ amounts to minimizing $\widehat{L}_{in}(\omega, h)$ over the span of $(K(x_1, \cdot), \dots, K(x_n, \cdot))$, *i.e.*, over functions h^γ of the form $h^\gamma = \sum_{i=1}^n \gamma_i K(x_i, \cdot)$ for $\gamma \in \mathbb{R}^n$. Restricting the objective to such functions results in the following inner optimization problem which is finite-dimensional:

$$\hat{\gamma}_\omega \coloneqq \operatorname*{arg\,min}_{\gamma\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} \ell_{in}(\omega, (\mathbf{K}\,\gamma)_i, y_i) + \frac{\lambda}{2}\gamma^\top \mathbf{K}\,\gamma,$$

where we used that $(h^\gamma(x_i))_{1\le i\le n} = \mathbf{K}\,\gamma$ and $\|h^\gamma\|_{\mathcal{H}}^2 = \gamma^\top \mathbf{K}\,\gamma$. Similarly, using that $(h^\gamma(\tilde{x}_j))_{1\le j\le m} = \overline{\mathbf{K}}\,\gamma$, we can express $\widehat{\mathcal{F}}(\omega)$ as follows:

$$\widehat{\mathcal{F}}(\omega) = \frac{1}{m}\sum_{j=1}^{m} \ell_{out}\big(\omega, (\overline{\mathbf{K}}\,\hat{\gamma}_\omega)_j, \tilde{y}_j\big).$$

Differentiating the above expression w.r.t. ω and applying the chain rule result in:

$$\nabla\widehat{\mathcal{F}}(\omega) = \frac{1}{m}\sum_{j=1}^{m} \partial_\omega \ell_{out}\big(\omega, (\overline{\mathbf{K}}\,\hat{\gamma}_\omega)_j, \tilde{y}_j\big) + \frac{1}{m}\sum_{j=1}^{m} (\partial_\omega \hat{\gamma}_\omega\, \overline{\mathbf{K}}^\top)_j \partial_v \ell_{out}\big(\omega, (\overline{\mathbf{K}}\,\hat{\gamma}_\omega)_j, \tilde{y}_j\big),$$

where $\partial_\omega \hat{\gamma}_\omega$ denotes the Jacobian of $\hat{\gamma}_\omega$. We can further express the above equation in matrix form to get:

$$\nabla\widehat{\mathcal{F}}(\omega) = \frac{1}{m}\,\mathbf{D}_\omega^{\mathbf{out}}\,\mathbb{1}_m + \frac{1}{m}\partial_\omega\hat{\gamma}_\omega\,\overline{\mathbf{K}}^\top\,\mathbf{D}_{\mathbf{v}}^{\mathbf{out}}\,. \tag{11}$$

Moreover, an application of the implicit function theorem[1] allows to directly express the Jacobian $\partial_\omega\hat{\gamma}_\omega$ as a solution of the following linear system obtained by differentiating the optimality condition for $\hat{\gamma}_\omega$ w.r.t. ω:

$$\mathbf{D}_{\omega,\mathbf{v}}^{\mathbf{in}}\,\mathbf{K} + (\partial_\omega\hat{\gamma}_\omega)\underbrace{\big(\mathbf{K}\,\mathbf{D}_{\mathbf{v},\mathbf{v}}^{\mathbf{in}} + n\lambda\mathbb{1}_{n\times n}\big)}_{\mathbf{M}}\mathbf{K} = 0.$$

A solution of the form $\partial_\omega\hat{\gamma}_\omega = -\,\mathbf{D}_{\omega,\mathbf{v}}^{\mathbf{in}}\,\mathbf{M}^{-1}$ always exists by invertibility of the matrix $\mathbf{M}$. The result follows after replacing $\partial_\omega\hat{\gamma}_\omega$ by $-\,\mathbf{D}_{\omega,\mathbf{v}}^{\mathbf{in}}\,\mathbf{M}^{-1}$ in Equation (11). □

Lemma B.2 (Estimator of the total functional gradient)**.** *Let $\omega \in \mathbb{R}^d$. Consider the following functional estimator:*

$$\widehat{\nabla\mathcal{F}}(\omega) = \frac{1}{m}\sum_{j=1}^{m} \partial_\omega \ell_{out}(\omega, \hat{h}_\omega(\tilde{x}_j), \tilde{y}_j) + \frac{1}{n}\sum_{i=1}^{n} \partial^2_{\omega,v}\ell_{in}(\omega, \hat{h}_\omega(x_i), y_i)\hat{a}_\omega(x_i).$$

Then, under Assumptions (A) to (D), $\widehat{\nabla\mathcal{F}}(\omega)$ admits the following expression:

$$\widehat{\nabla\mathcal{F}}(\omega) = \frac{1}{m}\,\mathbf{D}_\omega^{\mathbf{out}}\,\mathbb{1}_m + \frac{1}{n}\,\mathbf{D}_{\omega,\mathbf{v}}^{\mathbf{in}}\begin{bmatrix}\mathbf{K} & \mathbf{u}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\alpha}}_\omega \\ \hat{\beta}_\omega\end{bmatrix} \in \mathbb{R}^d,$$

[1] In the case where the matrix $\mathbf{K}$ is non-invertible, one needs to restrict γ to the orthogonal complement of the null space of $\mathbf{K}$. Such a restriction is valid since the resulting solution $\hat{h}_\omega$ will not depend on the component belonging to the null space of $\mathbf{K}$.

where $\mathbf{D}_{\mathbf{v}}^{\mathbf{out}}, \mathbf{D}_{\mathbf{v},\mathbf{v}}^{\mathbf{in}}, \mathbf{D}_{\omega}^{\mathbf{out}}, \mathbf{D}_{\omega,\mathbf{v}}^{\mathbf{in}}$ *are the same matrices given in Proposition B.1, while,* $\hat{\boldsymbol{\alpha}}_\omega \in \mathbb{R}^n$ *and* $\hat{\beta}_\omega \in \mathbb{R}$ *are solutions to the linear system:*

$$\begin{bmatrix} \mathbf{M}\,\mathbf{K} & \mathbf{M}\,\mathbf{u} \\ \mathbf{u}^\top \mathbf{M} & p \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}}_\omega \\ \hat{\beta}_\omega \end{bmatrix} = -\frac{n}{m} \begin{bmatrix} \mathbf{u} \\ v \end{bmatrix}, \tag{12}$$

where the vector $\mathbf{u}$ *and matrix* $\mathbf{M}$ *are the same as in Proposition B.1, while* p *and* v *are non-negative scalars.*

Proof. Let $\omega \in \mathbb{R}^d$. We start by providing an expression of $\hat{a}_\omega$ as a linear combination of the kernel evaluated at the inner training points x_i, *i.e.*, $K(x_i, \cdot)$, and some element $\xi \in \mathcal{H}$ that we will characterize shortly. From it, we will obtain the expression of $\widehat{\nabla \mathcal{F}}(\omega)$.

Expression of $\hat{a}_\omega$. Recall that $\hat{a}_\omega$ is the unique minimizer of $\widehat{L}_{adj}$ in Equation (5), which admits, for any $a \in \mathcal{H}$, the following simple expression by the reproducing property:

$$\widehat{L}_{adj}(\omega, a) = \frac{1}{2n} \sum_{i=1}^n \partial_v^2 \ell_{in}(\omega, \hat{h}_\omega(x_i), y_i) a^2(x_i) + \frac{1}{m} \left\langle a, \overbrace{\sum_{j=1}^m \partial_v \ell_{out}(\omega, \hat{h}_\omega(\tilde{x}_j), \tilde{y}_j) K(\tilde{x}_j, \cdot)}^{\xi} \right\rangle_{\mathcal{H}} + \frac{\lambda}{2} \|a\|_{\mathcal{H}}^2.$$

Hence, by application of the representer theorem, it follows that $\hat{a}_\omega$ admits an expression of the form:

$$\hat{a}_\omega = \sum_{i=1}^n (\hat{\boldsymbol{\alpha}}_\omega)_i K(x_i, \cdot) + \hat{\beta}_\omega \xi. \tag{13}$$

Therefore, it is possible to recover $\hat{a}_\omega$ by minimizing $a \mapsto L_{adj}(\omega, a)$ over the span of $(\xi, K(x_1, \cdot), \ldots, K(x_n, \cdot))$. Hence, to find the optimal coefficients $\hat{\boldsymbol{\alpha}}_\omega \coloneqq ((\hat{\boldsymbol{\alpha}}_\omega)_i)_{1 \le i \le n}$ and $\hat{\beta}_\omega$ we first need to express the objective L_{adj} in terms of the coefficients $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$ for a given $a^{\boldsymbol{\alpha},\beta} \in \mathcal{H}$ of the form $a^{\boldsymbol{\alpha},\beta} = \sum_{i=1}^n (\boldsymbol{\alpha})_i K(x_i, \cdot) + \beta \xi$. To this end, note that the vector $(\xi(x_1), \ldots, \xi(x_n))$ is exactly equal to $\mathbf{u} = \overline{\mathbf{K}}^\top \mathbf{D}_{\mathbf{v}}^{\mathbf{out}}$ as defined in Proposition B.1. Moreover, using the reproducing property, we directly have:

$$(a^{\boldsymbol{\alpha},\beta}(x_i))_{1 \le i \le n} = \begin{bmatrix} \mathbf{K} & \mathbf{u} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix}, \qquad \langle a^{\boldsymbol{\alpha},\beta}, \xi \rangle_{\mathcal{H}} = \begin{bmatrix} \mathbf{u}^\top & \|\xi\|_{\mathcal{H}}^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix}, \qquad \left\|a^{\boldsymbol{\alpha},\beta}\right\|_{\mathcal{H}}^2 = \begin{bmatrix} \boldsymbol{\alpha}^\top & \beta \end{bmatrix} \begin{bmatrix} \mathbf{K} & \mathbf{u} \\ \mathbf{u}^\top & \|\xi\|_{\mathcal{H}}^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix}.$$

We can therefore express the objective $\widehat{L}_{adj}$ as follows:

$$\widehat{L}_{adj}(\omega, a^{\boldsymbol{\alpha},\beta}) = \frac{1}{2n} \begin{bmatrix} \boldsymbol{\alpha}^\top & \beta \end{bmatrix} \begin{bmatrix} \mathbf{K} \\ \mathbf{u}^\top \end{bmatrix} \mathbf{D}_{\mathbf{v},\mathbf{v}}^{\mathbf{in}} \begin{bmatrix} \mathbf{K} & \mathbf{u} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix} + \frac{1}{m} \begin{bmatrix} \mathbf{u}^\top & \|\xi\|_{\mathcal{H}}^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix} + \frac{\lambda}{2} \begin{bmatrix} \boldsymbol{\alpha}^\top & \beta \end{bmatrix} \begin{bmatrix} \mathbf{K} & \mathbf{u} \\ \mathbf{u}^\top & \|\xi\|_{\mathcal{H}}^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix}.$$

Hence, the optimal coefficients $\hat{\boldsymbol{\alpha}}_\omega \coloneqq ((\hat{\boldsymbol{\alpha}}_\omega)_i)_{1 \le i \le n}$ and $\hat{\beta}_\omega$ are those minimizing the above quadratic form and are characterized by the following optimality condition:

$$\begin{bmatrix} \overbrace{(\mathbf{K}\,\mathbf{D}_{\mathbf{v},\mathbf{v}}^{\mathbf{in}} + n\lambda \mathbb{1}_{n\times n})\,\mathbf{K}}^{\mathbf{M}} & \overbrace{(\mathbf{K}\,\mathbf{D}_{\mathbf{v},\mathbf{v}}^{\mathbf{in}} + n\lambda \mathbb{1}_{n\times n})\,\mathbf{u}}^{\mathbf{M}} \\ \mathbf{u}^\top \underbrace{(\mathbf{K}\,\mathbf{D}_{\mathbf{v},\mathbf{v}}^{\mathbf{in}} + n\lambda \mathbb{1}_{n\times n})}_{\mathbf{M}} & \underbrace{\mathbf{u}^\top \mathbf{D}_{\mathbf{v},\mathbf{v}}^{\mathbf{in}}\,\mathbf{u} + n\lambda \|\xi\|_{\mathcal{H}}^2}_{p \ge 0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}}_\omega \\ \hat{\beta}_\omega \end{bmatrix} = -\frac{n}{m} \begin{bmatrix} \mathbf{u} \\ \underbrace{\|\xi\|_{\mathcal{H}}^2}_{v \ge 0} \end{bmatrix}.$$

Expression of $\widehat{\nabla \mathcal{F}}(\omega)$. The result follows directly after expressing $\widehat{\nabla \mathcal{F}}(\omega)$ in vector form using the notations $\mathbf{D}_{\omega}^{\mathbf{out}}$ and $\mathbf{D}_{\omega,\mathbf{v}}^{\mathbf{in}}$ from Proposition B.1 and recalling that $(\hat{a}_\omega(x_i))_{1 \le i \le n} = \begin{bmatrix} \mathbf{K} & \mathbf{u} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}}_\omega \\ \hat{\beta}_\omega \end{bmatrix}$. □

Proof of Proposition 3.1. Let $\omega \in \mathbb{R}^d$. Define

$$\widehat{\nabla \mathcal{F}}(\omega) = \frac{1}{m} \sum_{j=1}^m \partial_\omega \ell_{out}(\omega, \hat{h}_\omega(\tilde{x}_j), \tilde{y}_j) + \frac{1}{n} \sum_{i=1}^n \partial_{\omega,v}^2 \ell_{in}(\omega, \hat{h}_\omega(x_i), y_i) \hat{a}_\omega(x_i),$$

where $\hat{h}_\omega$ and $\hat{a}_\omega$ are given by Equations (10) and (13). We will show that $\widehat{\nabla \mathcal{F}}(\omega) = \nabla \widehat{\mathcal{F}}(\omega)$. By Lemma B.2 and Proposition B.1, we know that $\nabla \widehat{\mathcal{F}}(\omega)$ and $\widehat{\nabla \mathcal{F}}(\omega)$ admit the following expressions:

$$\nabla \widehat{\mathcal{F}}(\omega) = \frac{1}{m} \mathbf{D}^{\mathbf{out}}_{\omega} \mathbb{1}_m - \frac{1}{m} \mathbf{D}^{\mathbf{in}}_{\omega,\mathbf{v}} \mathbf{M}^{-1} \mathbf{u}$$
$$\widehat{\nabla \mathcal{F}}(\omega) = \frac{1}{m} \mathbf{D}^{\mathbf{out}}_{\omega} \mathbb{1}_m + \frac{1}{n} \mathbf{D}^{\mathbf{in}}_{\omega,\mathbf{v}} \begin{bmatrix} \mathbf{K} & \mathbf{u} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}}_\omega \\ \hat{\beta}_\omega \end{bmatrix}.$$

Taking the difference of the two estimators yields:

$$\begin{aligned} \widehat{\nabla \mathcal{F}}(\omega) - \nabla \widehat{\mathcal{F}}(\omega) &= \frac{1}{m} \mathbf{D}^{\mathbf{in}}_{\omega,\mathbf{v}} \left(\mathbf{M}^{-1} \mathbf{u} + \frac{m}{n} \begin{bmatrix} \mathbf{K} & \mathbf{u} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}}_\omega \\ \hat{\beta}_\omega \end{bmatrix} \right) \\ &= \frac{1}{m} \mathbf{D}^{\mathbf{in}}_{\omega,\mathbf{v}} \mathbf{M}^{-1} \underbrace{\left(\mathbf{u} + \frac{m}{n} \mathbf{M} \begin{bmatrix} \mathbf{K} & \mathbf{u} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}}_\omega \\ \hat{\beta}_\omega \end{bmatrix} \right)}_{=0}, \end{aligned}$$

where the term $\mathbf{u} + \frac{m}{n}(\mathbf{M}\,\mathbf{K}\,\hat{\boldsymbol{\alpha}}_\omega + \hat{\beta}_\omega\,\mathbf{M}\,\mathbf{u})$ is equal to 0 by definition of $\hat{\boldsymbol{\alpha}}_\omega$ and $\hat{\beta}_\omega$ as solutions of the linear system (12) of Lemma B.2. □

C. Preliminary Results

In this section, Ω is an arbitrary compact subset of $\mathbb{R}^d$ with $\text{hull}(\Omega)$ denoting its convex hull, which is also compact. We also consider an arbitrary fixed positive value Λ such that $\lambda \leq \Lambda$ as this would allow us to simplify the dependence of the boundedness and Lipschitz constants on λ.

C.1. Boundedness and Lipschitz continuity of $h^\star_\omega$ and $\hat{h}_\omega$

Proposition C.1 (Boundedness of $h^\star_\omega$ and $\hat{h}_\omega$)**.** *Under Assumptions (A) to (D), the functions $\omega \mapsto \|h^\star_\omega\|_{\mathcal{H}}$ and $\omega \mapsto \|\hat{h}_\omega\|_{\mathcal{H}}$ are bounded over* $\text{hull}(\Omega)$ *by $\frac{B\sqrt{\kappa}}{\lambda}$, where $B \coloneqq \sup_{\omega \in \text{hull}(\Omega), y \in \mathcal{Y}} |\partial_v \ell_{in}(\omega, 0, y)| > 0$. Moreover, for all $\omega \in \text{hull}(\Omega)$ and $x \in \mathcal{X}$, $h^\star_\omega(x)$ and $\hat{h}_\omega(x)$ take value in the compact interval $\mathcal{V} \coloneqq \left[-\frac{B\kappa}{\lambda}, \frac{B\kappa}{\lambda}\right] \subset \mathbb{R}$.*

Proof. **Boundedness of $\|h^\star_\omega\|_{\mathcal{H}}$ and $\left\|\hat{h}_\omega\right\|_{\mathcal{H}}$.** Let $\omega \in \text{hull}(\Omega)$. Using Lemma G.2, we know, for any $h \in \mathcal{H}$, that:

$$\|h - h^\star_\omega\|_{\mathcal{H}} \leq \frac{1}{\lambda} \|\partial_h L_{in}(\omega, h)\|_{\mathcal{H}}.$$

This is particularly valid for $h = 0$. Thus,

$$\|h^\star_\omega\|_{\mathcal{H}} \leq \frac{1}{\lambda} \|\partial_h L_{in}(\omega, 0)\|_{\mathcal{H}}.$$

Using the expression of the partial derivative $\partial_h L_{in}$ established in Proposition A.1, we obtain:

$$\|h^\star_\omega\|_{\mathcal{H}} \leq \frac{1}{\lambda} \left\| \mathbb{E}_{\mathbb{P}} \left[\partial_v \ell_{in}(\omega, 0, y) K(x, \cdot) \right] \right\|_{\mathcal{H}}.$$

By Assumption (A), K is bounded by κ. Hence, Jensen's inequality yields:

$$\|h^\star_\omega\|_{\mathcal{H}} \leq \frac{1}{\lambda} \mathbb{E}_{\mathbb{P}} \Big[|\partial_v \ell_{in}(\omega, 0, y)| \, \|K(x, \cdot)\|_{\mathcal{H}} \Big] \leq \frac{\sqrt{\kappa}}{\lambda} \mathbb{E}_{\mathbb{P}} \Big[|\partial_v \ell_{in}(\omega, 0, y)| \Big].$$

By Assumption (B), Ω and $\mathcal{Y}$ are compact, which implies that $\text{hull}(\Omega) \times \mathcal{Y}$ is compact. From Assumption (C), we know that the function $(\omega, y) \mapsto \partial_v \ell_{in}(\omega, 0, y)$ is continuous. Given that every continuous function on a compact space is bounded, we obtain:

$$\|h^\star_\omega\|_{\mathcal{H}} \leq \frac{B\sqrt{\kappa}}{\lambda} < +\infty, \quad \text{where} \quad B \coloneqq \sup_{\omega \in \text{hull}(\Omega), y \in \mathcal{Y}} |\partial_v \ell_{in}(\omega, 0, y)| > 0.$$

To prove that $\left\|\hat{h}_\omega\right\|_{\mathcal{H}} \leq \frac{B\sqrt{\kappa}}{\lambda}$, we follow a similar approach to that of $\|h^\star_\omega\|_{\mathcal{H}} \leq \frac{B\sqrt{\kappa}}{\lambda}$. More precisely, we investigate the case where the expectation is with respect to the empirical estimate $\hat{\mathbb{P}}_n$ of $\mathbb{P}$.

$h^\star_\omega(x)$ and $\hat{h}_\omega(x)$ belong to $\mathcal{V}$. Let $\omega \in \text{hull}(\Omega)$ and $x \in \mathcal{X}$. By the reproducing property, the Cauchy-Schwarz inequality, and Assumption (A), we have:

$$|h^\star_\omega(x)| \leq \sqrt{\kappa}\,\|h^\star_\omega\| \quad \text{and} \quad \left|\hat{h}_\omega(x)\right| \leq \sqrt{\kappa}\left\|\hat{h}_\omega\right\|.$$

Using the bound on $\|h^\star_\omega\|_{\mathcal{H}}$ and $\left\|\hat{h}_\omega\right\|_{\mathcal{H}}$ already proved in the first part of this proof, we get:

$$|h^\star_\omega(x)| \leq \frac{B\kappa}{\lambda} \quad \text{and} \quad \left|\hat{h}_\omega(x)\right| \leq \frac{B\kappa}{\lambda}.$$

This concludes the proof. □

Proposition C.2 (Lipschitz continuity of $\omega \mapsto h^\star_\omega$)**.** *Under Assumptions (A) to (D), the function $\omega \mapsto h^\star_\omega$ is $\frac{L\sqrt{\kappa}}{\lambda}$-Lipschitz continuous on* hull(Ω)*, where $L \coloneqq \sup_{\omega\in\text{hull}(\Omega),v\in\mathcal{V},y\in\mathcal{Y}} \left\|\partial^2_{\omega,v}\ell_{in}(\omega,v,y)\right\| > 0$, and $\mathcal{V}$ is the compact interval introduced in Proposition C.1.*

Proof. To prove this proposition, we adopt the strategy of finding an upper bound for the Jacobian, which serves as the Lipschitz constant.

Let $\omega \in \text{hull}(\Omega)$. Using Propositions A.1 and A.3, we know that $h \mapsto L_{in}(\omega, h)$ is λ-strongly convex and Fréchet differentiable. Also, by Proposition A.2, $\partial_h L_{in}$ is Fréchet differentiable on $\mathbb{R}^d \times \mathcal{H}$, and, a fortiori, Hadamard differentiable. Then, by the functional implicit differentiation theorem (Petrulionytė et al., 2024, Theorem 2.1), the Jacobian $\partial_\omega h^\star_\omega : \mathcal{H} \to \mathbb{R}^d$ can be expressed as:

$$\partial_\omega h^\star_\omega = -\partial^2_{\omega,h} L_{in}(\omega, h^\star_\omega)\left(\partial^2_h L_{in}(\omega, h^\star_\omega)\right)^{-1}.$$

We have:

$$\begin{aligned}
\|\partial_\omega h^\star_\omega\|_{\text{op}} &\leq \left\|\partial^2_{\omega,h} L_{in}(\omega, h^\star_\omega)\right\|_{\text{op}} \left\|\left(\partial^2_h L_{in}(\omega, h^\star_\omega)\right)^{-1}\right\|_{\text{op}} \\
&\leq \frac{\left\|\partial^2_{\omega,h} L_{in}(\omega, h^\star_\omega)\right\|_{\text{op}}}{\lambda} \\
&= \frac{\left\|\mathbb{E}_{\mathbb{P}}\left[\partial^2_{\omega,v}\ell_{in}(\omega, h^\star_\omega(x), y)K(x,\cdot)\right]\right\|_{\text{op}}}{\lambda} \\
&\leq \frac{\mathbb{E}_{\mathbb{P}}\left[\left\|\partial^2_{\omega,v}\ell_{in}(\omega, h^\star_\omega(x), y)\right\|\|K(x,\cdot)\|_{\mathcal{H}}\right]}{\lambda} \\
&\leq \frac{\sqrt{\kappa}\mathbb{E}_{\mathbb{P}}\left[\left\|\partial^2_{\omega,v}\ell_{in}(\omega, h^\star_\omega(x), y)\right\|\right]}{\lambda},
\end{aligned} \tag{14}$$

where the first line uses the sub-multiplicative property of the operator norm $\|\cdot\|_{\text{op}}$, the second line stems from the fact that $h \mapsto L_{in}(\omega, h)$ is λ-strongly convex, for any $\omega \in \mathbb{R}^d$, as proved in Proposition A.3, the third line follows from Proposition A.2, the fourth line uses Jensen's inequality, and the last line is a direct consequence of the boundedness of K by κ (Assumption (A)). According to Proposition C.1, $h^\star_\omega(x) \in \mathcal{V} \coloneqq \left[-\frac{B\kappa}{\lambda}, \frac{B\kappa}{\lambda}\right]$, which is a compact interval of $\mathbb{R}$, where $B \coloneqq \sup_{\omega\in\text{hull}(\Omega),y\in\mathcal{Y}} |\partial_v \ell_{in}(\omega, 0, y)| > 0$. By Assumption (B), Ω and $\mathcal{Y}$ are compact sets, hence hull$(\Omega) \times \mathcal{V} \times \mathcal{Y}$ is compact. Besides, by Assumption (C), $(\omega, v, y) \mapsto \partial_v \ell_{in}(\omega, v, y)$ is continuous over the domain hull$(\Omega) \times \mathcal{V} \times \mathcal{Y}$. Since every continuous function on a compact set is bounded, this leads to:

$$\mathbb{E}_{\mathbb{P}}\left[\left\|\partial^2_{\omega,v}\ell_{in}(\omega, h^\star_\omega(x), y)\right\|\right] \leq L \coloneqq \sup_{\omega\in\text{hull}(\Omega),v\in\mathcal{V},y\in\mathcal{Y}} \left\|\partial^2_{\omega,v}\ell_{in}(\omega, v, y)\right\| < +\infty.$$

Substituting this bound into Equation (14) means that $\frac{L\sqrt{\kappa}}{\lambda}$ is an upper-bound on $\|\partial_\omega h^\star_\omega\|_{\text{op}}$. Thus, the result follows as desired. □

C.2. Local boundedness and Lipschitz properties of ℓ_{in}, ℓ_{out}, and their derivatives

Proposition C.3 (Local boundedness). *Under Assumptions (A) to (D), the functions* $(\omega, x, y) \mapsto \ell_{out}(\omega, h^\star_\omega(x), y)$, $(\omega, x, y) \mapsto \partial_\omega \ell_{out}(\omega, h^\star_\omega(x), y)$, *and* $(\omega, x, y) \mapsto \partial_v \ell_{out}(\omega, h^\star_\omega(x), y)$ *are bounded over* $\mathrm{hull}(\Omega) \times \mathcal{X} \times \mathcal{Y}$ *by some positive constant* M_{out}. *Similarly, the functions* $(\omega, x, y) \mapsto \partial_v \ell_{in}(\omega, h^\star_\omega(x), y)$, $(\omega, x, y) \mapsto \partial^2_v \ell_{in}(\omega, h^\star_\omega(x), y)$, *and* $(\omega, x, y) \mapsto \partial^2_{\omega,v} \ell_{in}(\omega, h^\star_\omega(x), y)$ *are bounded over* $\mathrm{hull}(\Omega) \times \mathcal{X} \times \mathcal{Y}$ *by some positive constant* M_{in}. *The constants* M_{out} *and* M_{in} *are defined as:*

$$M_{out} \coloneqq \sup_{\omega \in \mathrm{hull}(\Omega), v \in \mathcal{V}, y \in \mathcal{Y}} \max\left(|\ell_{out}(\omega, v, y)|, \|\partial_\omega \ell_{out}(\omega, v, y)\|, |\partial_v \ell_{out}(\omega, v, y)|\right) > 0,$$
$$M_{in} \coloneqq \sup_{\omega \in \mathrm{hull}(\Omega), v \in \mathcal{V}, y \in \mathcal{Y}} \max\left(|\partial_v \ell_{in}(\omega, v, y)|, \left|\partial^2_v \ell_{in}(\omega, v, y)\right|, \left\|\partial^2_{\omega,v} \ell_{in}(\omega, v, y)\right\|\right) > 0,$$

where $\mathcal{V} \subset \mathbb{R}$ *is the compact interval defined in Proposition C.1.*

Proof. By Proposition C.1, we have that $h^\star_\omega(x) \in \mathcal{V} \coloneqq \left[-\frac{B\kappa}{\lambda}, \frac{B\kappa}{\lambda}\right] \subset \mathbb{R}$, for any $x \in \mathcal{X}$. From Assumption (C), we know that ℓ_{in}, ℓ_{out}, and their partial derivatives are all continuous on $\mathrm{hull}(\Omega) \times \mathcal{V} \times \mathcal{Y}$. Also, $\mathcal{Y}$ is compact by Assumption (B). Thus, $\mathrm{hull}(\Omega) \times \mathcal{V} \times \mathcal{Y}$ is compact. As every continuous function defined over a compact space is bounded, we obtain that:

$$\sup_{\omega \in \mathrm{hull}(\Omega), x \in \mathcal{X}, y \in \mathcal{Y}} |\ell_{out}(\omega, h^\star_\omega(x), y)| \leq \sup_{\omega \in \mathrm{hull}(\Omega), v \in \mathcal{V}, y \in \mathcal{Y}} |\ell_{out}(\omega, v, y)| < +\infty,$$
$$\sup_{\omega \in \mathrm{hull}(\Omega), x \in \mathcal{X}, y \in \mathcal{Y}} \|\partial_\bullet \ell_\circ(\omega, h^\star_\omega(x), y)\| \leq \sup_{\omega \in \mathrm{hull}(\Omega), v \in \mathcal{V}, y \in \mathcal{Y}} \|\partial_\bullet \ell_\circ(\omega, v, y)\| < +\infty,$$

where $\bullet \in \{\{v\}, \{w\}, \{w, v\}\}$ and $\circ \in \{in, out\}$. This implies the desired result. □

Proposition C.4 (Local Lipschitz continuity). *Under Assumptions (A) to (D), there exists a positive constant* Lip_{out} *so that for any* (x, y) *in* $\mathcal{X} \times \mathcal{Y}$, *the functions* $\omega \mapsto \ell_{out}(\omega, h^\star_\omega(x), y)$, $\omega \mapsto \partial_\omega \ell_{out}(\omega, h^\star_\omega(x), y)$, *and* $\omega \mapsto \partial_v \ell_{out}(\omega, h^\star_\omega(x), y)$ *are locally* $\frac{\mathrm{Lip}_{out}}{\lambda}$*-Lipschitz continuous over* $\mathrm{hull}(\Omega)$. *Similarly, there exists a positive constant* Lip_{in} *so that for any* (x, y) *in* $\mathcal{X} \times \mathcal{Y}$, *the functions* $\omega \mapsto \partial_v \ell_{in}(\omega, h^\star_\omega(x), y)$, $\omega \mapsto \partial^2_v \ell_{in}(\omega, h^\star_\omega(x), y)$, *and* $\omega \mapsto \partial^2_{\omega,v} \ell_{in}(\omega, h^\star_\omega(x), y)$ *are locally* $\frac{\mathrm{Lip}_{in}}{\lambda}$*-Lipschitz continuous* $\mathrm{hull}(\Omega)$. *The constants* Lip_{out} *and* Lip_{in} *are defined, for any* $0 < \lambda \leq \Lambda$, *as:*

$$\mathrm{Lip}_{out} \coloneqq (\Lambda + M_{in}\kappa) \max\left(M_{out}, \bar{M}_{out}\right) > 0$$
$$\mathrm{Lip}_{in} \coloneqq (\Lambda + M_{in}\kappa) \max\left(M_{in}, \bar{M}_{in}\right) > 0,$$

where:

$$\bar{M}_{out} \coloneqq \sup_{\omega \in \mathrm{hull}(\Omega), v \in \mathcal{V}, y \in \mathcal{Y}} \max\left(\left\|\partial^2_\omega \ell_{out}(\omega, v, y)\right\|_{\mathrm{op}}, \left\|\partial^2_{\omega,v} \ell_{out}(\omega, v, y)\right\|, \left|\partial^2_v \ell_{out}(\omega, v, y)\right|\right) > 0,$$
$$\bar{M}_{in} \coloneqq \sup_{\omega \in \mathrm{hull}(\Omega), v \in \mathcal{V}, y \in \mathcal{Y}} \max\left(\left\|\partial_\omega \partial^2_v \ell_{in}(\omega, v, y)\right\|, \left|\partial^3_v \ell_{in}(\omega, v, y)\right|, \left\|\partial_\omega \partial^2_{\omega,v} \ell_{in}(\omega, v, y)\right\|\right) > 0,$$

with M_{in} *and* M_{out} *being the positive constants defined in Proposition C.3, and* $\mathcal{V} \subset \mathbb{R}$ *is the compact interval defined in Proposition C.1.*

Proof. For any $(\omega, x, y) \in \mathrm{hull}(\Omega) \times \mathcal{X} \times \mathcal{Y}$, we have:

$$\begin{aligned}
\|\nabla_\omega \ell_{out}(\omega, h^\star_\omega(x), y)\| &= \|\partial_\omega \ell_{out}(\omega, h^\star_\omega(x), y) + \partial_v \ell_{out}(\omega, h^\star_\omega(x), y) \partial_\omega h^\star_\omega(x)\| \\
&\leq \|\partial_\omega \ell_{out}(\omega, h^\star_\omega(x), y)\| + |\partial_v \ell_{out}(\omega, h^\star_\omega(x), y)| \|\partial_\omega h^\star_\omega\|_{\mathrm{op}} \|K(x, \cdot)\|_{\mathcal{H}} \\
&\leq M_{out}\left(1 + \frac{M_{in}\kappa}{\lambda}\right) \\
&\leq \frac{M_{out}\left(\Lambda + M_{in}\kappa\right)}{\lambda},
\end{aligned}$$

where the first line uses the chain rule, the second line applies the triangle inequality and the reproducing property of the RKHS $\mathcal{H}$, the third line follows from Proposition C.3 to bound the derivatives of ℓ_{out}, from Proposition C.2, which states that the function $\omega \mapsto h^\star_\omega$ is $\frac{L\sqrt{\kappa}}{\lambda}$-Lipschitz continuous with $L \coloneqq \sup_{\omega \in \mathrm{hull}(\Omega), v \in \mathcal{V}, y \in \mathcal{Y}} \left\|\partial^2_{\omega,v} \ell_{in}(\omega, v, y)\right\| < M_{in}$,

to bound $\|\partial_\omega h^\star_\omega\|_{\mathrm{op}}$, and from Assumption (A) to bound $\|K(x,\cdot)\|_{\mathcal{H}}$, and the last line is a direct consequence of $0 < \lambda \leq \Lambda$. In a similar way, we obtain:

$$\|\nabla_\omega \partial_\omega \ell_{out}(\omega, h^\star_\omega(x), y)\|_{\mathrm{op}} \leq \frac{\bar{M}_{out}\,(\Lambda + M_{in}\kappa)}{\lambda}, \quad \|\nabla_\omega \partial_v \ell_{out}(\omega, h^\star_\omega(x), y)\| \leq \frac{\bar{M}_{out}\,(\Lambda + M_{in}\kappa)}{\lambda},$$
$$\|\nabla_\omega \partial_v \ell_{in}(\omega, h^\star_\omega(x), y)\| \leq \frac{M_{in}\,(\Lambda + M_{in}\kappa)}{\lambda}, \quad \left\|\nabla_\omega \partial_v^2 \ell_{in}(\omega, h^\star_\omega(x), y)\right\| \leq \frac{\bar{M}_{in}\,(\Lambda + M_{in}\kappa)}{\lambda},$$
$$\left\|\nabla_\omega \partial^2_{\omega,v} \ell_{in}(\omega, h^\star_\omega(x), y)\right\|_{\mathrm{op}} \leq \frac{\bar{M}_{in}\,(\Lambda + M_{in}\kappa)}{\lambda}.$$

Combining all these bounds concludes the proof. □

D. Generalization Properties

As before, let Ω be an arbitrary compact subset of $\mathbb{R}^d$.

D.1. Point-wise estimates

We present a point-wise upper-bound on the value error $\left|\mathcal{F}(\omega) - \widehat{\mathcal{F}}(\omega)\right|$ and gradient error $\left\|\nabla\mathcal{F}(\omega) - \widehat{\nabla\mathcal{F}}(\omega)\right\|$. To this end, we introduce the following notation for the error between the inner and outer objectives and their empirical approximations evaluated at the optimal inner solution $h^\star_\omega$:

$$\delta^{out}_\omega \coloneqq \left|L_{out}(\omega, h^\star_\omega) - \widehat{L}_{out}(\omega, h^\star_\omega)\right|, \qquad \delta^{in}_\omega \coloneqq \left|L_{in}(\omega, h^\star_\omega) - \widehat{L}_{in}(\omega, h^\star_\omega)\right|.$$

By abuse of notation, we introduce the following error between partial derivatives of L_{in} and $\widehat{L}_{in}$ (resp. L_{out} and $\widehat{L}_{out}$), evaluated at $(\omega, h^\star_\omega)$, *i.e.*,

$$\partial_h \delta^{out}_\omega \coloneqq \left\|\partial_h L_{out}(\omega, h^\star_\omega) - \partial_h \widehat{L}_{out}(\omega, h^\star_\omega)\right\|_{\mathcal{H}}, \qquad \partial_\omega \delta^{out}_\omega \coloneqq \left\|\partial_\omega L_{out}(\omega, h^\star_\omega) - \partial_\omega \widehat{L}_{out}(\omega, h^\star_\omega)\right\|,$$
$$\partial_h \delta^{in}_\omega \coloneqq \left\|\partial_h L_{in}(\omega, h^\star_\omega) - \partial_h \widehat{L}_{in}(\omega, h^\star_\omega)\right\|_{\mathcal{H}}, \qquad \partial_\omega \delta^{in}_\omega \coloneqq \left\|\partial_\omega L_{in}(\omega, h^\star_\omega) - \partial_\omega \widehat{L}_{in}(\omega, h^\star_\omega)\right\|,$$
$$\partial_h^2 \delta^{in}_\omega \coloneqq \left\|\partial_h^2 L_{in}(\omega, h^\star_\omega) - \partial_h^2 \widehat{L}_{in}(\omega, h^\star_\omega)\right\|_{\mathrm{op}}, \qquad \partial^2_{\omega,h} \delta^{in}_\omega \coloneqq \left\|\partial^2_{\omega,h} L_{in}(\omega, h^\star_\omega) - \partial^2_{\omega,h} \widehat{L}_{in}(\omega, h^\star_\omega)\right\|_{\mathrm{op}}.$$

Proposition D.1. *Under Assumptions (A) to (D), the following holds for any $\omega \in \Omega$:*

$$\left\|h^\star_\omega - \hat{h}_\omega\right\|_{\mathcal{H}} \leq \frac{1}{\lambda}\left\|\partial_h \widehat{L}_{in}(\omega, h^\star_\omega)\right\|_{\mathcal{H}} = \frac{1}{\lambda}\partial_h \delta^{in}_\omega.$$

Proof. Let $\omega \in \Omega$. The function $h \mapsto \widehat{L}_{in}(\omega, h)$ is λ-strongly convex and Fréchet differentiable by Propositions A.1 and A.3. Moreover, $\hat{h}_\omega$ is the minimizer of $h \mapsto \widehat{L}_{in}(\omega, h)$ by definition. Therefore, using Lemma G.2, we obtain a control on the distance in $\mathcal{H}$ to the optimum $\hat{h}_\omega$ of $h \mapsto \widehat{L}_{in}(\omega, h)$ in terms of the gradient $\partial_h \widehat{L}_{in}(\omega, h)$:

$$\left\|h - \hat{h}_\omega\right\|_{\mathcal{H}} \leq \frac{1}{\lambda}\left\|\partial_h \widehat{L}_{in}(\omega, h)\right\|_{\mathcal{H}}, \qquad \forall h \in \mathcal{H}.$$

In particular, choosing $h = h^\star_\omega$ yields the first inequality. The fact that $\left\|\partial_h \widehat{L}_{in}(\omega, h^\star_\omega)\right\|_{\mathcal{H}} = \partial_h \delta^{in}_\omega$ follows from the optimality of $h^\star_\omega$ which implies that $\partial_h L_{in}(\omega, h^\star_\omega) = 0$. □

Proposition D.2. *Under Assumptions (A) to (D), the following inequalities hold for any $\omega \in \Omega$:*

$$
\begin{aligned}
E_\omega^{out} &:= \left|\widehat{L}_{out}(\omega, h_\omega^\star) - \widehat{L}_{out}(\omega, \hat{h}_\omega)\right| \leq C_{out}\left\|h_\omega^\star - \hat{h}_\omega\right\|_{\mathcal{H}}, \\
\partial_h E_\omega^{out} &:= \left\|\partial_h \widehat{L}_{out}(\omega, h_\omega^\star) - \partial_h \widehat{L}_{out}(\omega, \hat{h}_\omega)\right\|_{\mathcal{H}} \leq C_{out}\left\|h_\omega^\star - \hat{h}_\omega\right\|_{\mathcal{H}}, \\
\partial_\omega E_\omega^{out} &:= \left\|\partial_\omega \widehat{L}_{out}(\omega, h_\omega^\star) - \partial_\omega \widehat{L}_{out}(\omega, \hat{h}_\omega)\right\| \leq C_{out}\left\|h_\omega^\star - \hat{h}_\omega\right\|_{\mathcal{H}}, \\
\partial_h^2 E_\omega^{in} &:= \left\|\partial_h^2 \widehat{L}_{in}(\omega, h_\omega^\star) - \partial_h^2 \widehat{L}_{in}(\omega, \hat{h}_\omega)\right\|_{\mathrm{op}} \leq C_{in}\left\|h_\omega^\star - \hat{h}_\omega\right\|_{\mathcal{H}}, \\
\partial_{\omega,h}^2 E_\omega^{in} &:= \left\|\partial_{\omega,h}^2 \widehat{L}_{in}(\omega, h_\omega^\star) - \partial_{\omega,h}^2 \widehat{L}_{in}(\omega, \hat{h}_\omega)\right\|_{\mathrm{op}} \leq C_{in}\left\|h_\omega^\star - \hat{h}_\omega\right\|_{\mathcal{H}}.
\end{aligned}
$$

The positive constants C_{out} and C_{in} are defined as:

$$
\begin{aligned}
C_{out} &:= \max\left(M_{out}\sqrt{\kappa}, \bar{M}_{out}\kappa, \bar{M}_{out}\sqrt{\kappa}\right) > 0, \\
C_{in} &:= \max\left(\bar{M}_{in}\kappa\sqrt{\kappa}, \bar{M}_{in}\kappa, M_{in}\sqrt{d\kappa}\right) > 0,
\end{aligned}
$$

where M_{out}, $\bar{M}_{out}$, and $\bar{M}_{in}$ are the positive constants defined in Propositions C.3 and C.4.

Proof. **Lipschitz continuity of some functions of interest.** Let $\omega \in \Omega$. According to Proposition C.1, both $h_\omega^\star(x)$ and $\hat{h}_\omega(x)$ lie in the compact interval $\mathcal{V} := \left[-\frac{B\kappa}{\lambda}, \frac{B\kappa}{\lambda}\right] \subset \mathbb{R}$, for any $x \in \mathcal{X}$, where $B := \sup_{\omega \in \mathrm{hull}(\Omega), y \in \mathcal{Y}} |\partial_v \ell_{in}(\omega, 0, y)| > 0$. By Assumption (B), $\mathcal{Y}$ is a compact set. Hence $\Omega \times \mathcal{V} \times \mathcal{Y}$ is a compact set as well. Furthermore, by Assumption (C), $(\omega, v, y) \mapsto \ell_{in}(\omega, v, y)$, $(\omega, v, y) \mapsto \ell_{out}(\omega, v, y)$, and their derivatives are all continuous over the compact domain $\Omega \times \mathcal{V} \times \mathcal{Y}$. Therefore, these functions and their derivatives are bounded on this domain. In particular, this also holds when v takes the specific values $h_\omega^\star(x)$ or $\hat{h}_\omega(x)$. Let $\bar{v}$ be either $h_\omega^\star(x)$ or $\hat{h}_\omega(x)$, for any $x \in \mathcal{X}$. For any $\omega \in \Omega$, and $y \in \mathcal{Y}$, we have:

$$
\begin{aligned}
|\partial_v \ell_{out}(\omega, \bar{v}, y)| &\leq \sup_{\omega \in \Omega, v \in \mathcal{V}, y \in \mathcal{Y}} |\partial_v \ell_{out}(\omega, v, y)| \leq M_{out} < +\infty, \\
\left|\partial_v^2 \ell_{out}(\omega, \bar{v}, y)\right| &\leq \sup_{\omega \in \Omega, v \in \mathcal{V}, y \in \mathcal{Y}} \left|\partial_v^2 \ell_{out}(\omega, v, y)\right| \leq \bar{M}_{out} < +\infty, \\
\left\|\partial_{\omega,v}^2 \ell_{out}(\omega, \bar{v}, y)\right\| &\leq \sup_{\omega \in \Omega, v \in \mathcal{V}, y \in \mathcal{Y}} \left\|\partial_{\omega,v}^2 \ell_{out}(\omega, v, y)\right\| \leq \bar{M}_{out} < +\infty, \\
\left|\partial_v^3 \ell_{in}(\omega, \bar{v}, y)\right| &\leq \sup_{\omega \in \Omega, v \in \mathcal{V}, y \in \mathcal{Y}} \left|\partial_v^3 \ell_{in}(\omega, v, y)\right| \leq \bar{M}_{in} < +\infty, \\
\left\|\partial_v \partial_{\omega,v}^2 \ell_{in}(\omega, \bar{v}, y)\right\|_{\mathrm{op}} &\leq \sup_{\omega \in \Omega, v \in \mathcal{V}, y \in \mathcal{Y}} \left\|\partial_\omega \partial_v^2 \ell_{in}(\omega, v, y)\right\| \leq \bar{M}_{in} < +\infty
\end{aligned}
$$

This means that $v \in \mathcal{V} \mapsto \ell_{out}(\omega, v, y)$, $v \in \mathcal{V} \mapsto \partial_v \ell_{out}(\omega, v, y)$, $v \in \mathcal{V} \mapsto \partial_\omega \ell_{out}(\omega, v, y)$, $v \in \mathcal{V} \mapsto \partial_v^2 \ell_{in}(\omega, v, y)$, and $v \in \mathcal{V} \mapsto \partial_{\omega,v}^2 \ell_{in}(\omega, v, y)$ are Lipschitz continuous, with Lipschitz constants M_{out}, $\bar{M}_{out}$, $\bar{M}_{out}$, $\bar{M}_{in}$, and $\bar{M}_{in}$, respectively, for any $\omega \in \Omega$ and $y \in \mathcal{Y}$.

Upper-bounds. We have:

$$
\begin{aligned}
E_\omega^{out} := \left|\widehat{L}_{out}(\omega, h_\omega^\star) - \widehat{L}_{out}(\omega, \hat{h}_\omega)\right| &= \left|\frac{1}{m}\sum_{j=1}^m \ell_{out}(\omega, h_\omega^\star(\tilde{x}_j), \tilde{y}_j) - \frac{1}{m}\sum_{j=1}^m \ell_{out}(\omega, \hat{h}_\omega(\tilde{x}_j), \tilde{y}_j)\right| \\
&\leq \frac{1}{m}\sum_{j=1}^m \left|\ell_{out}(\omega, h_\omega^\star(\tilde{x}_j), \tilde{y}_j) - \ell_{out}(\omega, \hat{h}_\omega(\tilde{x}_j), \tilde{y}_j)\right| \\
&\leq \frac{M_{out}}{m}\sum_{j=1}^m \left|h_\omega^\star(\tilde{x}_j) - \hat{h}_\omega(\tilde{x}_j)\right| \\
&\leq M_{out}\sqrt{\kappa}\left\|h_\omega^\star - \hat{h}_\omega\right\|_{\mathcal{H}},
\end{aligned}
$$

where the first line uses the definition of $(\omega, h) \mapsto \widehat{L}_{out}(\omega, h)$, the second line applies the triangle inequality, the third line leverages the fact that $v \mapsto \ell_{out}(\omega, v, y)$ is M_{out}-Lipschitz continuous, for any $\omega \in \Omega$ and $y \in \mathcal{Y}$, and the last line follows from the reproducing property of the RKHS $\mathcal{H}$, Cauchy-Schwarz's inequality, and Assumption (A) to bound $\|K(x, \cdot)\|_{\mathcal{H}}$ by $\sqrt{\kappa}$. Similarly, we obtain:

$$\partial_h E_\omega^{out} \leq \bar{M}_{out}\kappa \left\| h_\omega^\star - \hat{h}_\omega \right\|_{\mathcal{H}}, \quad \partial_\omega E_\omega^{out} \leq \bar{M}_{out}\sqrt{\kappa} \left\| h_\omega^\star - \hat{h}_\omega \right\|_{\mathcal{H}}, \quad \partial_h^2 E_\omega^{in} \leq \bar{M}_{in}\kappa\sqrt{\kappa} \left\| h_\omega^\star - \hat{h}_\omega \right\|_{\mathcal{H}},$$
$$\partial_{\omega,h}^2 E_\omega^{in} \leq \bar{M}_{in}\kappa \left\| h_\omega^\star - \hat{h}_\omega \right\|_{\mathcal{H}}.$$

Combining all the bounds finishes the proof. □

Proposition D.3. *Under Assumptions (A) to (D), the following inequalities hold for any $\omega \in \Omega$:*

$$\|\partial_h L_{out}(\omega, h_\omega^\star)\|_{\mathcal{H}} \leq C_{out}, \quad \left\|\partial_{\omega,h}^2 L_{in}(\omega, h_\omega^\star)\right\|_{\text{op}} \leq C_{in}, \quad \left\|\partial_{\omega,h}^2 \widehat{L}_{in}(\omega, \hat{h}_\omega)\right\|_{\text{op}} \leq C_{in},$$

where C_{out} and C_{in} are the positive constants defined in Proposition D.2.

Proof. Let $\omega \in \Omega$.

Upper-bound on $\|\partial_h L_{out}(\omega, h_\omega^\star)\|_{\mathcal{H}}$. We have:

$$\begin{aligned}
\|\partial_h L_{out}(\omega, h_\omega^\star)\|_{\mathcal{H}} &= \left\| \mathbb{E}_{\mathbb{Q}} \left[\partial_v \ell_{out}(\omega, h_\omega^\star(x), y) K(x, \cdot) \right] \right\|_{\mathcal{H}} \\
&\leq \mathbb{E}_{\mathbb{Q}} \Big[\left| \partial_v \ell_{out}(\omega, h_\omega^\star(x), y) \right| \|K(x, \cdot)\|_{\mathcal{H}} \Big] \\
&\leq \sqrt{\kappa} \mathbb{E}_{\mathbb{Q}} \Big[\left| \partial_v \ell_{out}(\omega, h_\omega^\star(x), y) \right| \Big],
\end{aligned}$$

where the first line follows from Proposition A.1, the second line results from the triangle inequality, and the last line uses Assumption (A) to bound $\|K(x, \cdot)\|_{\mathcal{H}}$ by $\sqrt{\kappa}$. Furthermore, we know by Proposition C.1 that $(\omega, h_\omega^\star(x), y)$ belongs to the compact subset $\Omega \times \mathcal{V} \times \mathcal{Y}$ and by Proposition C.3 that $\partial_v \ell_{out}(\omega, h_\omega^\star(x), y)$ is bounded by a constant M_{out} on $\text{hull}(\Omega) \times \mathcal{V} \times \mathcal{Y}$. Hence, it follows that:

$$\|\partial_h L_{out}(\omega, h_\omega^\star)\|_{\mathcal{H}} \leq \sqrt{\kappa} M_{out} \leq C_{out},$$

where C_{out} is defined in Proposition D.2.

Upper-boud on $\left\|\partial_{\omega,h}^2 L_{in}(\omega, h_\omega^\star)\right\|_{\text{op}}$. According to Proposition A.2, $\partial_{\omega,h}^2 L_{in}(\omega, h_\omega^\star)$ is a Hilbert-Schmidt operator, which points to:

$$\left\|\partial_{\omega,h}^2 L_{in}(\omega, h_\omega^\star)\right\|_{\text{op}} \leq \left\|\partial_{\omega,h}^2 L_{in}(\omega, h_\omega^\star)\right\|_{\text{HS}} = \sqrt{\sum_{l=1}^{d} \left\|\partial_{\omega_k,h}^2 L_{in}(\omega, h_\omega^\star)\right\|_{\mathcal{H}}^2}. \tag{15}$$

This means that to find an upper-bound on $\left\|\partial_{\omega,h}^2 L_{in}(\omega, h_\omega^\star)\right\|_{\text{op}}$, it suffices to establish an upper-bound on $\left\|\partial_{\omega_l,h}^2 L_{in}(\omega, h_\omega^\star)\right\|_{\mathcal{H}}^2$ for any $l \in \{1, \dots, d\}$. For a fixed $l \in \{1, \dots, d\}$, we have:

$$\begin{aligned}
\left\|\partial_{\omega_l,h}^2 L_{in}(\omega, h_\omega^\star)\right\|_{\mathcal{H}}^2 &= \left\| \mathbb{E}_{\mathbb{P}} \left[\partial_{\omega_l,v}^2 \ell_{in}(\omega, h_\omega^\star(x), y) K(x, \cdot) \right] \right\|_{\mathcal{H}}^2 \\
&\leq \mathbb{E}_{\mathbb{P}} \left[\left| \partial_{\omega_l,v}^2 \ell_{in}(\omega, h_\omega^\star(x), y) \right|^2 \|K(x, \cdot)\|_{\mathcal{H}}^2 \right] \\
&\leq \mathbb{E}_{\mathbb{P}} \left[\left\| \partial_{\omega,v}^2 \ell_{in}(\omega, h_\omega^\star(x), y) \right\|^2 \right] \kappa,
\end{aligned}$$

where the first line follows from Proposition A.2, the second line is a consequence of Jensen's inequality applied on the convex function $\|\cdot\|^2$, and the last line applies Assumption (A) to bound $\|K(x, \cdot)\|_{\mathcal{H}}^2$ by κ. Incorporating this upper-bound into Equation (15) yields:

$$\left\|\partial_{\omega,h}^2 L_{in}(\omega, h_\omega^\star)\right\|_{\text{op}} \leq \sqrt{\mathbb{E}_{\mathbb{P}} \left[\left\| \partial_{\omega,v}^2 \ell_{in}(\omega, h_\omega^\star(x), y) \right\|^2 \right] d\kappa} \leq M_{in}\sqrt{d\kappa} \leq C_{out},$$

where we used Proposition C.3 to bound $\partial^2_{\omega,v}\ell_{in}(\omega, h^\star_\omega(x), y)$ by the constant M_{in}.

Upper-bound on $\left\|\partial^2_{\omega,h}\widehat{L}_{in}(\omega,\hat{h}_\omega)\right\|_{\text{op}}$. The derivation of this upper bound follows the same steps as the previous one, with the only differences being the use of $\widehat{L}_{in}$ instead of L_{in}, and $\hat{h}_\omega$ instead of $h^\star_\omega$.

Note that in the last step of each of the three upper-bounds, we used the fact that the functions we are dealing with are continuous (by Assumption **(C)**) on $\Omega \times \mathcal{V} \times \mathcal{Y}$, which is compact because Ω and $\mathcal{Y}$ are compact by Assumption **(B)** and $\mathcal{V}$ is a compact interval of $\mathbb{R}$ defined in Proposition C.1. Hence, those functions are bounded. □

Proposition D.4 (Approximation bounds)**.** *Under Assumptions **(A)** to **(D)**, the following holds for any $\omega \in \Omega$:*

$$\left|\mathcal{F}(\omega) - \widehat{\mathcal{F}}(\omega)\right| \leq \delta^{out}_\omega + \frac{C_{out}}{\lambda}\partial_h\delta^{in}_\omega,$$
$$\left\|\nabla\mathcal{F}(\omega) - \widehat{\nabla\mathcal{F}}(\omega)\right\| \leq \partial_\omega\delta^{out}_\omega + \frac{C_{in}}{\lambda}\partial_h\delta^{out}_\omega + \frac{C_{out}C_{in}}{\lambda^2}\partial^2_h\delta^{in}_\omega + \frac{C_{out}}{\lambda}\partial^2_{\omega,h}\delta^{in}_\omega + \frac{C_{out}}{\lambda}\left(1 + 2\frac{C_{in}}{\lambda} + \frac{C^2_{in}}{\lambda^2}\right)\partial_h\delta^{in}_\omega,$$

where the constants C_{in} and C_{out} are given in Proposition D.2.

Proof. In all what follows, we fix a value for ω in Ω. We start by controlling the value function, then its gradient.

Control on the value function. By the triangle inequality, we have:

$$\left|\mathcal{F}(\omega) - \widehat{\mathcal{F}}(\omega)\right| \leq \underbrace{\left|L_{out}(\omega, h^\star_\omega) - \widehat{L}_{out}(\omega, h^\star_\omega)\right|}_{\delta^{out}_\omega} + \underbrace{\left|\widehat{L}_{out}(\omega, h^\star_\omega) - \widehat{L}_{out}(\omega, \hat{h}_\omega)\right|}_{E^{out}_\omega}, \tag{16}$$

According to Proposition D.2, the error term E^{out}_ω is controlled by the norm of the difference $h^\star_\omega - \hat{h}_\omega$, *i.e.*, $E^{out}_\omega \leq C_{out}\left\|h^\star_\omega - \hat{h}_\omega\right\|_{\mathcal{H}}$. Moreover, by Proposition D.1, we know that $\left\|h^\star_\omega - \hat{h}_\omega\right\|_{\mathcal{H}} \leq \frac{1}{\lambda}\partial_h\delta^{in}_\omega$. Therefore, combining both bounds yields: $E^{out}_\omega \leq \frac{C_{out}}{\lambda}\partial_h\delta^{in}_\omega$. The upper-bound on value function follows by substituting the previous inequality into Equation (16).

Control on the gradient. By Proposition A.4, we have the following expression for the total gradient $\nabla\mathcal{F}$:

$$\nabla\mathcal{F}(\omega) = \partial_\omega L_{out}(\omega, h^\star_\omega) - \partial^2_{\omega,h}L_{in}(\omega, h^\star_\omega)\left(\partial^2_h L_{in}(\omega, h^\star_\omega)\right)^{-1}\partial_h L_{out}(\omega, h^\star_\omega).$$

Similarly, the gradient estimator $\widehat{\nabla\mathcal{F}}$ is defined by replacing L_{out} and L_{in} by their empirical versions $\widehat{L}_{out}$ and $\widehat{L}_{in}$, and $h^\star_\omega$ by $\hat{h}_\omega \coloneqq \arg\min_{h\in\mathcal{H}}\widehat{L}_{in}(\omega, h)$ in the above expression, *i.e.*,

$$\widehat{\nabla\mathcal{F}}(\omega) = \partial_\omega\widehat{L}_{out}(\omega,\hat{h}_\omega) - \partial^2_{\omega,h}\widehat{L}_{in}(\omega,\hat{h}_\omega)\left(\partial^2_h\widehat{L}_{in}(\omega,\hat{h}_\omega)\right)^{-1}\partial_h\widehat{L}_{out}(\omega,\hat{h}_\omega).$$

To simplify notations, for any $h \in \mathcal{H}$, we introduce the following operators $R(h), \hat{R}(h) : \mathcal{H} \to \Omega$:

$$R(h) = \partial^2_{\omega,h}L_{in}(\omega, h)\left(\partial^2_h L_{in}(\omega, h)\right)^{-1} \quad \text{and} \quad \hat{R}(h) = \partial^2_{\omega,h}\widehat{L}_{in}(\omega, h)\left(\partial^2_h\widehat{L}_{in}(\omega, h)\right)^{-1}.$$

The difference $\nabla\mathcal{F}(\omega) - \widehat{\nabla\mathcal{F}}(\omega)$ can be decomposed as:

$$\begin{aligned}\nabla\mathcal{F}(\omega) - \widehat{\nabla\mathcal{F}}(\omega) = &\left(\partial_\omega L_{out}(\omega, h^\star_\omega) - \partial_\omega\widehat{L}_{out}(\omega, h^\star_\omega)\right) + \left(\partial_\omega\widehat{L}_{out}(\omega, h^\star_\omega) - \partial_\omega\widehat{L}_{out}(\omega,\hat{h}_\omega)\right)\\ &- \hat{R}(\hat{h}_\omega)\Big(\left(\partial_h L_{out}(\omega, h^\star_\omega) - \partial_h\widehat{L}_{out}(\omega, h^\star_\omega)\right) + \left(\partial_h\widehat{L}_{out}(\omega, h^\star_\omega) - \partial_h\widehat{L}_{out}(\omega,\hat{h}_\omega)\right)\Big)\\ &- \left(R(h^\star_\omega) - \hat{R}(\hat{h}_\omega)\right)\partial_h L_{out}(\omega, h^\star_\omega).\end{aligned}$$

By taking the norm of the above equality and using triangle inequality, we obtain the following upper-bound:

$$\begin{aligned}\left\|\nabla\mathcal{F}(\omega)-\widehat{\nabla\mathcal{F}}(\omega)\right\| \leq& \underbrace{\left\|\partial_\omega L_{out}(\omega,h^\star_\omega)-\partial_\omega \widehat{L}_{out}(\omega,h^\star_\omega)\right\|}_{\partial_\omega \delta^{out}_\omega}+\underbrace{\left\|\partial_\omega \widehat{L}_{out}(\omega,h^\star_\omega)-\partial_\omega \widehat{L}_{out}(\omega,\hat{h}_\omega)\right\|}_{\partial_\omega E^{out}_\omega}\\ &+\left\|\hat{R}(\hat{h}_\omega)\right\|_{\mathrm{op}}\left(\underbrace{\left\|\partial_h L_{out}(\omega,h^\star_\omega)-\partial_h \widehat{L}_{out}(\omega,h^\star_\omega)\right\|_{\mathcal{H}}}_{\partial_h \delta^{out}_\omega}+\underbrace{\left\|\partial_h \widehat{L}_{out}(\omega,h^\star_\omega)-\partial_h \widehat{L}_{out}(\omega,\hat{h}_\omega)\right\|_{\mathcal{H}}}_{\partial_h E^{out}_\omega}\right)\\ &+\left\|R(h^\star_\omega)-\hat{R}(\hat{h}_\omega)\right\|_{\mathrm{op}}\|\partial_h L_{out}(\omega,h^\star_\omega)\|_{\mathcal{H}}.\end{aligned} \tag{17}$$

Next, we provide upper-bounds on $\left\|R(h^\star_\omega)-\hat{R}(\hat{h}_\omega)\right\|_{\mathrm{op}}$ and $\left\|\hat{R}(\hat{h}_\omega)\right\|_{\mathrm{op}}$ in terms of derivatives of L_{in} and $\widehat{L}_{in}$.

Upper-bounds on $\left\|R(h^\star_\omega)-\hat{R}(\hat{h}_\omega)\right\|_{\mathrm{op}}$ and $\left\|\hat{R}(\hat{h}_\omega)\right\|_{\mathrm{op}}$. By application of Propositions A.2 and A.3, we deduce that $\partial^2_{\omega,h}L_{in}(\omega,h^\star_\omega)$, $\partial^2_h L_{in}(\omega,h^\star_\omega)$, $\partial^2_{\omega,h}\widehat{L}_{in}(\omega,\hat{h}_\omega)$ and $\partial^2_h\widehat{L}_{in}(\omega,\hat{h}_\omega)$ are all bounded operators. Moreover, since L_{in} and $\widehat{L}_{in}$ are λ-strongly convex in their second argument by Proposition A.3, it follows that $\partial^2_h L_{in}(\omega,h^\star_\omega)\geq\lambda\,\mathrm{Id}_{\mathcal{H}}$ and $\partial^2_h\widehat{L}_{in}(\omega,\hat{h}_\omega)\geq\lambda\,\mathrm{Id}_{\mathcal{H}}$. We can therefore apply Lemma G.1 which yields the following inequalities:

$$\begin{aligned}\left\|R(h^\star_\omega)-\hat{R}(\hat{h}_\omega)\right\|_{\mathrm{op}} \leq& \frac{1}{\lambda^2}\left\|\partial^2_{\omega,h}L_{in}(\omega,h^\star_\omega)\right\|_{\mathrm{op}}\left\|\partial^2_h L_{in}(\omega,h^\star_\omega)-\partial^2_h\widehat{L}_{in}(\omega,\hat{h}_\omega)\right\|_{\mathrm{op}}\\ &+\frac{1}{\lambda}\left\|\partial^2_{\omega,h}L_{in}(\omega,h^\star_\omega)-\partial^2_{\omega,h}\widehat{L}_{in}(\omega,\hat{h}_\omega)\right\|_{\mathrm{op}},\\ \left\|\hat{R}(\hat{h}_\omega)\right\|_{\mathrm{op}} \leq& \frac{1}{\lambda}\left\|\partial^2_{\omega,h}\widehat{L}_{in}(\omega,\hat{h}_\omega)\right\|_{\mathrm{op}}.\end{aligned}$$

By applying the triangle inequality to both terms of the first inequality above, we obtain:

$$\begin{aligned}\left\|R(h^\star_\omega)-\hat{R}(\hat{h}_\omega)\right\|_{\mathrm{op}} \leq& \frac{1}{\lambda^2}\left\|\partial^2_{\omega,h}L_{in}(\omega,h^\star_\omega)\right\|_{\mathrm{op}}\left(\underbrace{\left\|\partial^2_h L_{in}(\omega,h^\star_\omega)-\partial^2_h\widehat{L}_{in}(\omega,h^\star_\omega)\right\|_{\mathrm{op}}}_{\partial^2_h\delta^{in}_\omega}+\underbrace{\left\|\partial^2_h\widehat{L}_{in}(\omega,h^\star_\omega)-\partial^2_h\widehat{L}_{in}(\omega,\hat{h}_\omega)\right\|_{\mathrm{op}}}_{\partial^2_h E^{in}_\omega}\right)\\ &+\frac{1}{\lambda}\left(\underbrace{\left\|\partial^2_{\omega,h}L_{in}(\omega,h^\star_\omega)-\partial^2_{\omega,h}\widehat{L}_{in}(\omega,h^\star_\omega)\right\|_{\mathrm{op}}}_{\partial^2_{\omega,h}\delta^{in}_\omega}+\underbrace{\left\|\partial^2_{\omega,h}\widehat{L}_{in}(\omega,h^\star_\omega)-\partial^2_{\omega,h}\widehat{L}_{in}(\omega,\hat{h}_\omega)\right\|_{\mathrm{op}}}_{\partial^2_{\omega,h}E^{in}_\omega}\right).\end{aligned}$$

Final bound. We can now substitute the above bounds on $\left\|R(h^\star_\omega)-\hat{R}(\hat{h}_\omega)\right\|_{\mathrm{op}}$ and $\left\|\hat{R}(\hat{h}_\omega)\right\|_{\mathrm{op}}$ into Equation (17) to obtain the following upper-bound on the gradient error:

$$\begin{aligned}\left\|\nabla\mathcal{F}(\omega)-\widehat{\nabla\mathcal{F}}(\omega)\right\| \leq& \partial_\omega\delta^{out}_\omega+\partial_\omega E^{out}_\omega+\frac{1}{\lambda}\left\|\partial^2_{\omega,h}\widehat{L}_{in}(\omega,\hat{h}_\omega)\right\|_{\mathrm{op}}\left(\partial_h\delta^{out}_\omega+\partial_h E^{out}_\omega\right)\\ &+\|\partial_h L_{out}(\omega,h^\star_\omega)\|_{\mathcal{H}}\left(\frac{1}{\lambda^2}\left\|\partial^2_{\omega,h}L_{in}(\omega,h^\star_\omega)\right\|_{\mathrm{op}}\left(\partial^2_h\delta^{in}_\omega+\partial^2_h E^{in}_\omega\right)+\frac{1}{\lambda}\left(\partial^2_{\omega,h}\delta^{in}_\omega+\partial^2_{\omega,h}E^{in}_\omega\right)\right).\end{aligned} \tag{18}$$

Furthermore, by Proposition D.3, we have the following upper-bounds on the derivatives of L_{in} and L_{out}:

$$\|\partial_h L_{out}(\omega,h^\star_\omega)\|_{\mathcal{H}}\leq C_{out},\quad \left\|\partial^2_{\omega,h}L_{in}(\omega,h^\star_\omega)\right\|_{\mathrm{op}}\leq C_{in},\quad \left\|\partial^2_{\omega,h}\widehat{L}_{in}(\omega,\hat{h}_\omega)\right\|_{\mathrm{op}}\leq C_{in}.$$

Incorporating the above bounds into Equation (18), we further get:

$$\begin{aligned}\left\|\nabla\mathcal{F}(\omega)-\widehat{\nabla\mathcal{F}}(\omega)\right\| \leq& \partial_\omega\delta^{out}_\omega+\partial_\omega E^{out}_\omega+\frac{C_{in}}{\lambda}\left(\partial_h\delta^{out}_\omega+\partial_h E^{out}_\omega\right)\\ &+C_{out}\left(\frac{C_{in}}{\lambda^2}\left(\partial^2_h\delta^{in}_\omega+\partial^2_h E^{in}_\omega\right)+\frac{1}{\lambda}\left(\partial^2_{\omega,h}\delta^{in}_\omega+\partial^2_{\omega,h}E^{in}_\omega\right)\right).\end{aligned}$$

By Proposition D.2, we can upper-bound the error terms $\partial_\omega E_\omega^{out}, \partial_h E_\omega^{out}$ by $C_{out}\left\|h_\omega^\star - \hat{h}_\omega\right\|_{\mathcal{H}}$ and $\partial_h^2 E_\omega^{in}, \partial_{\omega,h}^2 E_\omega^{in}$ by the difference $C_{in}\left\|h_\omega^\star - \hat{h}_\omega\right\|_{\mathcal{H}}$. Furthermore, since $\left\|h_\omega^\star - \hat{h}_\omega\right\|_{\mathcal{H}} \leq \frac{1}{\lambda}\partial_h \delta_\omega^{in}$ by Proposition D.1, we can further show that gradient error satisfies the desired bound:

$$\left\|\nabla\mathcal{F}(\omega) - \widehat{\nabla\mathcal{F}}(\omega)\right\| \leq \partial_\omega \delta_\omega^{out} + \frac{C_{in}}{\lambda}\partial_h \delta_\omega^{out} + C_{out}\left(\frac{C_{in}}{\lambda^2}\partial_h^2 \delta_\omega^{in} + \frac{1}{\lambda}\partial_{\omega,h}^2 \delta_\omega^{in}\right) + \frac{C_{out}}{\lambda}\left(1 + 2\frac{C_{in}}{\lambda} + \frac{C_{in}^2}{\lambda^2}\right)\partial_h \delta_\omega^{in}.$$

□

D.2. Maximal inequalities

Proposition D.5 (Maximal inequalities for empirical processes)**.** *Let Λ be a positive constant. Under Assumptions (A) to (D), the following maximal inequalities hold for any $0 < \lambda \leq \Lambda$:*

$$\mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega} \delta_\omega^{out}\right] \leq \sqrt{\frac{1}{\lambda^2 m}}c(\Omega)\max(M_{out}\mathrm{Lip}_{out}\,\mathrm{diam}(\Omega), \Lambda M_{out}^2)$$
$$\mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega} \partial_\omega \delta_\omega^{out}\right] \leq \sqrt{\frac{d}{\lambda^2 m}}c(\Omega)\max(M_{out}\,\mathrm{Lip}_{out}\,\mathrm{diam}(\Omega), \Lambda M_{out}^2),$$

where $c(\Omega)$ is a positive constant greater than 1 that depends only on Ω and d, while Lip_{out} and M_{out} are positive constants defined in Propositions C.3 and C.4.

Proof. We will apply the result of Proposition E.3 which provides maximal inequalities for real-valued empirical processes that are uniformly bounded and Lipschitz in their parameter. To this end, consider the parametric families:

$$\mathcal{T}_l^{out} := \{\mathcal{X}\times\mathcal{Y} \ni (x,y) \mapsto \partial_{w_l}\ell_{out}(\omega, h_\omega^\star(x), y) \mid \omega\in\Omega\}, \qquad 1 \leq l \leq d$$
$$\mathcal{T}_0^{out} := \{\mathcal{X}\times\mathcal{Y} \ni (x,y) \mapsto \ell_{out}(\omega, h_\omega^\star(x), y) \mid \omega\in\Omega\}.$$

For any $0 \leq l \leq d$, these real-valued functions are uniformly bounded by a positive constant M_{out}, thanks to Proposition C.3. Moreover, by Proposition C.4, the functions $\omega \mapsto \partial_{\omega_l}\ell_{out}(\omega, h_\omega^\star(x), y)$, and $\omega \mapsto \ell_{out}(\omega, h_\omega^\star(x), y)$ are all $\lambda^{-1}\,\mathrm{Lip}_{out}$-Lipschitz for any $(x,y) \in \mathcal{X}\times\mathcal{Y}$. Hence, Proposition E.3 is applicable to each of these families, with $\mathbb{D}$ set to $\mathbb{Q}$ and $\mathcal{Z}$ set to $\mathcal{X}\times\mathcal{Y}$. We treat both δ_ω^{out} and $\partial_\omega\delta_\omega^{out}$ separately.

A maximal inequality for δ_ω^{out}. For $l = 0$, we readily apply Proposition E.3 with $p = 1$ to get the following maximal inequality for δ_ω^{out}:

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega} \delta_\omega^{out}\right] :=& \mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega}\left|\mathbb{E}_{(x,y)\sim\mathbb{Q}}[\ell_{out}(\omega, h_\omega^\star(x), y)] - \frac{1}{m}\sum_{j=1}^m \ell_{out}(\omega, h_\omega^\star(\tilde{x}_j), \tilde{y}_j)\right|\right] \\ \leq& \sqrt{\frac{1}{\lambda^2 m}}c(\Omega)\max(\mathrm{Lip}_{out}\,\mathrm{diam}(\Omega), \Lambda M_{out}^2).\end{aligned}$$

A maximal inequality for $\partial_\omega\delta_\omega^{out}$. We now turn to $\partial_\omega\delta_\omega^{out}$, which involves vector-valued processes (as an error between the gradient and its estimate). While the maximal inequalities in Proposition E.3 hold for real-valued processes, we will first obtain maximal inequalities for each component appearing in $\partial_\omega\delta_\omega^{out}$ and then sum these to control $\partial_\omega\delta_\omega^{out}$. To this end, we first use the Cauchy-Schwarz inequality which implies that $\mathbb{E}_{\mathbb{Q}}[\sup_{\omega\in\Omega}(\partial_\omega\delta_\omega^{out})] \leq \mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega}(\partial_\omega\delta_\omega^{out})^2\right]^{\frac{1}{2}}$. Thus we

only need to control $\mathbb{E}_{\mathbb{Q}}\big[\sup_{\omega\in\Omega}(\partial_\omega \delta_\omega^{out})^2\big]$. Simple calculations show that:

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega}\partial_\omega\delta_\omega^{out}\right]^2 &\le \mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega}(\partial_\omega\delta_\omega^{out})^2\right]\\ &\le \sum_{l=1}^{d}\mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega}\left|\mathbb{E}_{(x,y)\sim\mathbb{Q}}[\partial_{w_l}\ell_{out}(\omega,h_\omega^\star(x),y)] - \frac{1}{m}\sum_{j=1}^{m}\partial_{\omega_l}\ell_{out}(\omega,h_\omega^\star(\tilde{x}_j),\tilde{y}_j)\right|^2\right]\\ &\le \left(\sqrt{\frac{d}{\lambda^2 m}}c(\Omega)\max(\mathrm{Lip}_{out}\,\mathrm{diam}(\Omega),\Lambda M_{out}^2)\right)^2,\end{aligned}$$

where the last inequality follows by application of Proposition E.3 with $p=2$ to each term in the right-hand side of the first inequality for $1\le l\le d$. We get the desired bound on $\mathbb{E}_{\mathbb{Q}}[\sup_{\omega\in\Omega}\partial_\omega\delta_\omega^{out}]$ by taking the square root of the above inequality. □

Proposition D.6 (Maximal inequalities for RKHS-valued empirical processes)**.** *Let Λ be a positive constant. Under Assumptions (A) to (D), the following maximal inequalities hold for any $0<\lambda\le\Lambda$:*

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega}\partial_h\delta_\omega^{out}\right] &\le \lambda^{-\frac{1}{4}}m^{-\frac{1}{2}}\left(c(\Omega)\max\left(\widetilde{M}_{out,1}\widetilde{L}_{out,1}\,\mathrm{diam}(\Omega),\Lambda\widetilde{M}_{out,1}^2\right)\right)^{\frac{1}{4}},\\ \mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}\partial_h\delta_\omega^{in}\right] &\le \lambda^{-\frac{1}{4}}n^{-\frac{1}{2}}\left(c(\Omega)\max\left(\widetilde{M}_{in,1}\widetilde{L}_{in,1}\,\mathrm{diam}(\Omega),\Lambda\widetilde{M}_{in,1}^2\right)\right)^{\frac{1}{4}},\\ \mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}\partial_{\omega,h}^2\delta_\omega^{in}\right] &\le \lambda^{-\frac{1}{4}}n^{-\frac{1}{2}}d^{\frac{1}{2}}\left(c(\Omega)\max\left(\widetilde{M}_{in,1}\widetilde{L}_{in,1}\,\mathrm{diam}(\Omega),\Lambda\widetilde{M}_{in,1}^2\right)\right)^{\frac{1}{4}},\\ \mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}\partial_h^2\delta_\omega^{in}\right] &\le \lambda^{-\frac{1}{4}}n^{-\frac{1}{2}}\left(c(\Omega)\max\left(\widetilde{M}_{in,2}^2\widetilde{L}_{in,2}\,\mathrm{diam}(\Omega),\Lambda\widetilde{M}_{in,2}^2\right)\right)^{\frac{1}{4}},\end{aligned}$$

where $c(\Omega)$ is a positive constant greater than 1 that depends only on Ω and d, $\widetilde{L}_{out,1},\widetilde{L}_{in,1},\widetilde{L}_{in,2},\widetilde{M}_{out,1},\widetilde{M}_{in,1},\widetilde{M}_{in,2}$ are positive constants defined as:

$$\widetilde{L}_{out,1}\coloneqq 2\,\mathrm{Lip}_{out}\,M_{out}\kappa,\quad \widetilde{L}_{in,1}\coloneqq 2\,\mathrm{Lip}_{in}\,M_{in}\kappa,\quad \widetilde{L}_{in,2}\coloneqq 2\,\mathrm{Lip}_{in}\,M_{in}\kappa^2,$$
$$\widetilde{M}_{out,1}\coloneqq M_{out}^2\kappa,\quad \widetilde{M}_{in,1}\coloneqq M_{in}^2\kappa,\quad \widetilde{M}_{in,2}\coloneqq M_{in}^2\kappa^2,$$

and $\mathrm{Lip}_{out},\mathrm{Lip}_{in},M_{out},M_{in}$ are positive constants given in Propositions C.3 and C.4.

Proof. Consider parametric families of real-valued functions indexed by Ω of the form:

$$\mathcal{T}_{s,a}\coloneqq\{t_\omega:((x,y),(x',y'))\mapsto f_s(\omega,x,y)f_s(\omega,x',y')K^a(x,x')\mid\omega\in\Omega\},$$

where $a\in\{1,2\}$, s is an integer satisfying $0\le s\le d+2$, and $f_s(\omega,x,y)$ are real-valued functions given by:

$$f_0:(\omega,x,y)\mapsto\partial_v\ell_{out}(\omega,h_\omega^\star(x),y),\qquad f_1:(\omega,x,y)\mapsto\partial_v\ell_{in}(\omega,h_\omega^\star(x),y),\qquad f_2:(\omega,x,y)\mapsto\partial_v^2\ell_{in}(\omega,h_\omega^\star(x),y),$$
$$f_{2+l}:(\omega,x,y)\mapsto\partial_{\omega_l,v}^2\ell_{in}(\omega,h_\omega^\star(x),y),\quad 1\le l\le d.$$

For any $1\le s\le d+2$, the real-valued functions f_s are uniformly bounded by a positive constant M_{in} thanks to Proposition C.3. Moreover, since the kernel K is bounded by κ due to Assumption (A), it follows that all elements t_ω of $\mathcal{T}_{s,a}$ are uniformly bounded by $\widetilde{M}_{in,a}\coloneqq M_{in}^2\kappa^a$. Moreover, for $1\le s\le d+2$, the functions $\omega\mapsto f_s(\omega,x,y)$, are $\lambda^{-1}\,\mathrm{Lip}_{in}$-Lipschitz for any $(x,y)\in\mathcal{X}\times\mathcal{Y}$, by Proposition C.4. Hence, it follows that the maps $\omega\mapsto t_\omega((x,y),(x',y'))$ are $\lambda^{-1}\widetilde{L}_{in,a}$-Lipschitz with $\widetilde{L}_{in,a}\coloneqq 2\,\mathrm{Lip}_{in}\,M_{in}\kappa^a$ for any (x,y) and (x',y') in $\mathcal{X}\times\mathcal{Y}$. Similarly, for $s=0$, we get the same properties, albeit, with different constants, *i.e.*, the family $\mathcal{T}_{0,a}$ is uniformly bounded by a constant $\widetilde{M}_{out,a}\coloneqq M_{out}^2\kappa^a$ with M_{out} introduced in Proposition C.3, and is $\lambda^{-1}\widetilde{L}_{out,a}$-Lipschitz in its parameter with $\widetilde{L}_{out,a}\coloneqq 2\,\mathrm{Lip}_{out}\,M_{out}\kappa^a$ where Lip_{out} is given in Proposition C.4. Hence, the maximal inequality in Proposition E.4 is applicable to each of these

families with $\mathcal{Z}$ set to $\mathcal{X} \times \mathcal{Y}$, and $\mathbb{D}$ set either to $\mathbb{P}$ for $1 \leq s \leq d+2$, or to $\mathbb{Q}$ for $s = 0$. For conciseness, in all what follows, we will write $z = (x, y)$ and $z_i = (x_i, y_i)$ and $\tilde{z}_j = (\tilde{x}_j, \tilde{y}_j)$ for $1 \leq i \leq n$ and $1 \leq j \leq m$.

Maximal inequalities for $\partial_h \delta_\omega^{out}$ and $\partial_h \delta_\omega^{in}$. We control $\partial_h \delta_\omega^{out}$ first as $\partial_h \delta_\omega^{in}$ will be dealt with similarly. Using Cauchy-Schwarz inequality and standard calculus, we have that:

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega} \partial_h \delta_\omega^{out}\right]^2 &\leq \mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega} (\partial_h \delta_\omega^{out})^2\right] \\
&\coloneqq \mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega} \left\| \mathbb{E}_{(x,y)\sim\mathbb{Q}}[\partial_v \ell_{out}(\omega, h_\omega^\star(x), y) K(x,\cdot)] - \frac{1}{m}\sum_{j=1}^m \partial_v \ell_{out}(\omega, h_\omega^\star(\tilde{x}_j), \tilde{y}_j) K(\tilde{x}_j, \cdot) \right\|_{\mathcal{H}}^2\right] \\
&= \mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega} \mathbb{E}_{z,z'\sim\mathbb{Q}\otimes\mathbb{Q}}\left[t_\omega(z,z')\right] + \frac{1}{m^2}\sum_{i,j=1}^m t_\omega(z_i, z_j) - \frac{2}{m}\sum_{j=1}^m \mathbb{E}_{z\sim\mathbb{Q}}\left[t_\omega(z, \tilde{z}_j)\right]\right],
\end{aligned}$$

where $t_\omega(z,z') \coloneqq \partial_v \ell_{out}(\omega, h_\omega^\star(x), y) \partial_v \ell_{out}(\omega, h_\omega^\star(x'), y') K(x,x') \in \mathcal{T}_{0,1}$. The last term is precisely what Proposition E.4 controls when applying it to the family $\mathcal{T}_{0,1}$ and choosing $\mathbb{D}$ to be $\mathbb{Q}$. Therefore, the following maximal inequality holds by application of Proposition E.4:

$$\mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega} \partial_h \delta_\omega^{out}\right] \leq \lambda^{-\frac{1}{4}} m^{-\frac{1}{2}} \left(c(\Omega) \max\left(\widetilde{M}_{out,1} \widetilde{L}_{out,1} \operatorname{diam}(\Omega), \Lambda \widetilde{M}_{out,1}^2\right)\right)^{\frac{1}{4}},$$

where $c(\Omega)$ is a positive constant greater than 1 the depends only on Ω and d. We obtain a similar inequality for $\partial_h \delta_\omega^{in}$ by carrying out similar calculations, then applying Proposition E.4 to the family $\mathcal{T}_{1,1}$ and choosing $\mathbb{P}$ for the probability distribution $\mathbb{D}$. The resulting bound is then of the form:

$$\mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega} \partial_h \delta_\omega^{in}\right] \leq \lambda^{-\frac{1}{4}} n^{-\frac{1}{2}} \left(c(\Omega) \max\left(\widetilde{M}_{in,1} \widetilde{L}_{in,1} \operatorname{diam}(\Omega), \Lambda \widetilde{M}_{in,1}^2\right)\right)^{\frac{1}{4}}.$$

A maximal inequality for $\partial^2_{\omega,h} \delta_\omega^{in}$. We have:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega} \partial^2_{\omega,h} \delta_\omega^{in}\right]^2 &\overset{(a)}{\leq} \mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega} (\partial^2_{\omega,h} \delta_\omega^{in})^2\right] \\
&\overset{(b)}{\coloneqq} \mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega} \left\| \mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\partial^2_{\omega,v} \ell_{in}(\omega, h_\omega^\star(x), y) K(x,\cdot)\right] - \frac{1}{n}\sum_{i=1}^n \partial^2_{\omega,v} \ell_{in}(\omega, h_\omega^\star(x_i), y_i) K(x_i, \cdot) \right\|_{\mathrm{op}}^2\right] \\
&\overset{(c)}{\leq} \mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega} \left\| \mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\partial^2_{\omega,v} \ell_{in}(\omega, h_\omega^\star(x), y) K(x,\cdot)\right] - \frac{1}{n}\sum_{i=1}^n \partial^2_{\omega,v} \ell_{in}(\omega, h_\omega^\star(x_i), y_i) K(x_i, \cdot) \right\|_{\mathrm{HS}}^2\right] \\
&\overset{(d)}{=} \sum_{l=1}^d \mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega} \left\| \mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\partial^2_{\omega_l,v} \ell_{in}(\omega, h_\omega^\star(x), y) K(x,\cdot)\right] - \frac{1}{n}\sum_{i=1}^n \partial^2_{\omega_l,v} \ell_{in}(\omega, h_\omega^\star(x_i), y_i) K(x_i, \cdot) \right\|_{\mathcal{H}}^2\right] \\
&\overset{(e)}{=} \sum_{l=1}^d \mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega} \mathbb{E}_{z,z'\sim\mathbb{P}\otimes\mathbb{P}}\left[t_{\omega,l}(z,z')\right] + \frac{1}{n^2}\sum_{i,j=1}^n t_{\omega,l}(z_i, z_j) - \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{z\sim\mathbb{P}}\left[t_{\omega,l}(z, z_i)\right]\right],
\end{aligned}$$

where we introduced $t_{\omega,l}(z,z') \coloneqq \partial^2_{\omega_l,v} \ell_{in}(\omega, h_\omega^\star(x), y) \partial^2_{\omega_l,v} \ell_{in}(\omega, h_\omega^\star(x'), y') K(x,x') \in \mathcal{T}_{2+l,1}$. Here, (a) follows by Cauchy-Schwarz inequality, (b) is obtained by definition of $\partial^2_{\omega,h} \delta_\omega^{in}$, while (c) uses the general fact that the operator norm of an operator is upper-bounded by its Hilbert-Schmidt norm which is finite in our case by application of Proposition A.2. Moreover, (d) further uses the Hilbert-Schmidt norm of an operator in terms of the norm of its rows, while (e) simply expands the squared RKHS norm and uses the reproducing property in the RKHS $\mathcal{H}$. Each term in the last item (e) is precisely what Proposition E.4 controls when applying it to the families $\mathcal{T}_{2+l,1}$ for $1 \leq l \leq d$ and choosing $\mathbb{D}$ to be $\mathbb{P}$.

Therefore, the following maximal inequality holds by a direct application of Proposition E.4:

$$\mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}\partial^2_{\omega,h}\delta^{in}_{\omega}\right] \leq \lambda^{-\frac{1}{4}}n^{-\frac{1}{2}}d^{\frac{1}{2}}\left(c(\Omega)\max\left(\widetilde{M}_{in,1}\widetilde{L}_{in,1}\operatorname{diam}(\Omega),\Lambda\widetilde{M}^2_{in,1}\right)\right)^{\frac{1}{4}},$$

where $c(\Omega)$ is a positive constant greater than 1 that depends only on Ω and d.

A maximal inequality for $\partial^2_h\delta^{in}_{\omega}$. We will use a similar approach as for $\partial^2_{\omega,h}\delta^{in}_{\omega}$. We have:

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}\partial^2_h\delta^{in}_{\omega}\right]^2\\
&\overset{(a)}{\leq}\mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}(\partial^2_h\delta^{in}_{\omega})^2\right]\\
&\overset{(b)}{:=}\mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}\left\|\mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\partial^2_v\ell_{in}(\omega,h^\star_\omega(x),y)K(x,\cdot)\otimes K(x,\cdot)\right]-\frac{1}{n}\sum_{i=1}^n\partial^2_v\ell_{in}(\omega,h^\star_\omega(x_i),y_i)K(x_i,\cdot)\otimes K(x_i,\cdot)\right\|^2_{\mathrm{op}}\right]\\
&\overset{(c)}{\leq}\mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}\left\|\mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\partial^2_v\ell_{in}(\omega,h^\star_\omega(x),y)K(x,\cdot)\otimes K(x,\cdot)\right]-\frac{1}{n}\sum_{i=1}^n\partial^2_v\ell_{in}(\omega,h^\star_\omega(x_i),y_i)K(x_i,\cdot)\otimes K(x_i,\cdot)\right\|^2_{\mathrm{HS}}\right]\\
&\overset{(d)}{=}\mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}\mathbb{E}_{z,z'\sim\mathbb{P}\otimes\mathbb{P}}\left[t_\omega(z,z')\right]+\frac{1}{n^2}\sum_{i,j=1}^n t_\omega(z_i,z_j)-\frac{2}{n}\sum_{i=1}^n\mathbb{E}_{z\sim\mathbb{P}}\left[t_\omega(z,z_i)\right]\right],
\end{aligned}$$

where we introduced $t_\omega(z,z') \coloneqq \partial^2_v\ell_{in}(\omega,x,y)\partial^2_v\ell_{in}(\omega,x',y')K^2(x,x') \in \mathcal{T}_{2,2}$. Here, (a) follows by Cauchy-Schwarz inequality, (b) is obtained by definition of $\partial^2_h\delta^{in}_{\omega}$, while (c) uses the general fact that the operator norm of an operator is upper-bounded by its Hilbert-Schmidt norm which is finite in our case by application of Proposition A.2. Moreover, (d) further uses the identity in Lemma G.3 for computing the Hilbert-Schmidt norm of sum/expectation of tensor-product operators. The last item (d) is precisely what Proposition E.4 controls when applying it to the family $\mathcal{T}_{2,2}$ and choosing $\mathbb{D}$ to be $\mathbb{P}$. Therefore, the following maximal inequality holds by direct application of Proposition E.4:

$$\mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}\partial^2_h\delta^{in}_{\omega}\right] \leq \lambda^{-\frac{1}{4}}n^{-\frac{1}{2}}\left(c(\Omega)\max\left(\widetilde{M}_{in,2}\widetilde{L}_{in,2}\operatorname{diam}(\Omega),\Lambda\widetilde{M}^2_{in,2}\right)\right)^{\frac{1}{4}},$$

where $c(\Omega)$ is a positive constant greater than 1 that depends only on Ω and d. □

D.3. Proof of Theorem 4.1

Theorem D.7 (Generalization bounds)**.** *The following holds under Assumptions (A) to (D):*

$$\begin{aligned}
\mathbb{E}\left[\sup_{\omega\in\Omega}\left|\mathcal{F}(\omega)-\widehat{\mathcal{F}}(\omega)\right|\right] &\lesssim \frac{1}{\lambda m^{\frac{1}{2}}}+\frac{C_{out}}{\lambda^{\frac{5}{4}}n^{\frac{1}{2}}}\\
\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\nabla\mathcal{F}(\omega)-\widehat{\nabla\mathcal{F}}(\omega)\right\|\right] &\lesssim \frac{1}{\lambda}\left(d^{\frac{1}{2}}+\frac{C_{in}}{\lambda^{\frac{1}{4}}}\right)\frac{1}{m^{\frac{1}{2}}}+\frac{C_{out}}{\lambda^{\frac{5}{4}}}\left(2+3\frac{C_{in}}{\lambda}+\frac{C^2_{in}}{\lambda^2}\right)\frac{1}{n^{\frac{1}{2}}},
\end{aligned}$$

where the constants C_{in} and C_{out} are given in Proposition D.2.

Proof. Using the point-wise estimates in Proposition D.4 and taking their supremum over Ω followed by the expectations

over data, the following error bounds hold:

$$\begin{aligned}
\mathbb{E}\left[\sup_{\omega\in\Omega}\left|\mathcal{F}(\omega)-\widehat{\mathcal{F}}(\omega)\right|\right] \leq& \mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega}\delta_{\omega}^{out}\right]+\frac{C_{out}}{\lambda}\mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}\partial_h\delta_{\omega}^{in}\right],\\
\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\nabla\mathcal{F}(\omega)-\widehat{\nabla\mathcal{F}}(\omega)\right\|\right] \leq& \mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega}\partial_{\omega}\delta_{\omega}^{out}\right]+\frac{C_{in}}{\lambda}\mathbb{E}_{\mathbb{Q}}\left[\sup_{\omega\in\Omega}\partial_h\delta_{\omega}^{out}\right]\\
&+\frac{C_{out}}{\lambda}\left(1+2\frac{C_{in}}{\lambda}+\frac{C_{in}^2}{\lambda^2}\right)\mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}\partial_h\delta_{\omega}^{in}\right]\\
&+\frac{C_{out}C_{in}}{\lambda^2}\mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}\partial_h^2\delta_{\omega}^{in}\right]+\frac{C_{out}}{\lambda}\mathbb{E}_{\mathbb{P}}\left[\sup_{\omega\in\Omega}\partial_{\omega,h}^2\delta_{\omega}^{in}\right].
\end{aligned}$$

Furthermore, we can use the maximal inequalities in Propositions D.5 and D.6 to control each term appearing in the right-hand side of the above inequalities:

$$\begin{aligned}
\mathbb{E}\left[\sup_{\omega\in\Omega}\left|\mathcal{F}(\omega)-\widehat{\mathcal{F}}(\omega)\right|\right] \leq& R\Big(m^{-\frac{1}{2}}\lambda^{-1}+C_{out}n^{-\frac{1}{2}}\lambda^{-(1+\frac{1}{4})}\Big)\\
\mathbb{E}\left[\sup_{\omega\in\Omega}\left\|\nabla\mathcal{F}(\omega)-\widehat{\nabla\mathcal{F}}(\omega)\right\|\right] \leq& R\Big(m^{-\frac{1}{2}}\lambda^{-1}d^{\frac{1}{2}}+C_{in}m^{-\frac{1}{2}}\lambda^{-(1+\frac{1}{4})}+C_{out}n^{-\frac{1}{2}}\lambda^{-(1+\frac{1}{4})}\left(1+2\frac{C_{in}}{\lambda}+\frac{C_{in}^2}{\lambda^2}\right)\\
&+C_{out}C_{in}n^{-\frac{1}{2}}\lambda^{-(2+\frac{1}{4})}+C_{out}n^{-\frac{1}{2}}\lambda^{-(1+\frac{1}{4})}\Big),
\end{aligned}$$

where the constant R depends only on the Lipschitz constants Lip_{in}, Lip_{out}, the upper-bounds M_{in}, M_{out}, the bound κ on the kernel, the set Ω and the dimension d. Rearranging the obtained upper-bounds concludes the proof. □

E. Maximal Inequalities for Bounded and Lipschitz Family of Functions

Let $\mathcal{Z}$ be a subset of a Euclidean space and Ω be a compact subset of $\mathbb{R}^d$. Denote by $\otimes^k\mathcal{Z}$ the k-th tensor power of $\mathcal{Z}$, for any $k\geq 1$. Consider a parametric family $\mathcal{T}$ of real-valued functions defined over $\mathcal{Z}$ and indexed by a parameter $\omega\in\Omega$, *i.e.*,

$$\mathcal{T}\coloneqq\{\mathcal{Z}\ni z\mapsto t_\omega(z)\in\mathbb{R}\mid\omega\in\Omega\}. \tag{19}$$

For a given probability measure μ, denote by $L_2(\mu)$ the space of square μ-integrable real-valued functions. We denote by $\|f\|_{\mathbb{D},2}\coloneqq\mathbb{E}_{\mathbb{D}}\left[f(z)^2\right]^{\frac{1}{2}}$ the $L_2(\mu)$-norm of any function $f\in L_2(\mu)$. For any $\epsilon>0$, we denote by $D\left(\epsilon,\mathcal{T},L_2(\mu)\right)$ the ϵ-packing number of $\mathcal{T}$ w.r.t. $L_2(\mu)$. The next proposition provides a control on such a number under regularity conditions on the family $\mathcal{T}$.

Proposition E.1 (Control on the packing number)**.** *Assume that Ω is a compact subset of $\mathbb{R}^d$, that the parametric family $\mathcal{T}$ defined in Equation (19) is uniformly bounded by a positive constant M, and that there exists a positive constant L so that, for any measure μ, $\omega\mapsto t_\omega(z)$ is L-Lipschitz for any $z\in\mathcal{Z}$. Then, there exists a positive constant $c(\Omega)$ greater than 1 that depends only on Ω and d so that, for any probability measure μ on $\mathcal{Z}$, the following bound holds for any $0<\epsilon\leq M$:*

$$D\left(\epsilon,\mathcal{T},L_2(\mu)\right)\leq c(\Omega)\left(\frac{\max\left(L\,\mathrm{diam}(\Omega),M\right)}{\epsilon}\right)^d.$$

Proof. First using (Kosorok, 2008, Lemma 9.18) and (Kosorok, 2008, Paragraph 8.1.2), we know that the ϵ-packing number $D\left(\epsilon,\mathcal{T},L_2(\mu)\right)$ is smaller than the $\frac{\epsilon}{2}$-bracketing number $N_{[]}\left(\frac{\epsilon}{2},\mathcal{T},L_2(\mu)\right)$. Hence, we only need to control the bracketing number. To this end, we recall that the function $\omega\mapsto t_\omega(z)$ is L-Lipschitz for any $z\in\mathcal{Z}$, so that (Van der Vaart, 2000, Example 19.7) ensures the existence of a positive constant $c(\Omega)$ that depends only on Ω for which the following inequality holds for any $0<\epsilon<L\mathrm{diam}(\Omega)$:

$$1\leq N_{[]}\left(\epsilon,\mathcal{T},L_2(\mu)\right)\leq c(\Omega)\left(\frac{L\,\mathrm{diam}(\Omega)}{\epsilon}\right)^d.$$

Moreover, since the ϵ-bracketing number is decreasing in ϵ, it holds that:

$$N_{[]}\left(\epsilon,\mathcal{T},L_2(\mu)\right)\leq N_{[]}\left(\epsilon_-,\mathcal{T},L_2(\mu)\right)\leq c(\Omega)\left(\frac{L\,\mathrm{diam}(\Omega)}{\epsilon_-}\right)^d,$$

for any $\epsilon \geq L\text{diam}(\Omega)$ and $\epsilon_- \leq L\,\text{diam}(\Omega)$. Taking the limit when ϵ_- approaches $L\text{diam}(\Omega)$ yields $N_{[]}\left(\epsilon, \mathcal{T}, L_2(\mu)\right) \leq c(\Omega)$ for any $\epsilon \geq L\,\text{diam}(\Omega)$. Hence, we have shown so far that for any $\epsilon > 0$:

$$N_{[]}\left(\epsilon, \mathcal{T}, L_2(\mu)\right) \leq c(\Omega) \max\left(1, \left(\frac{L\,\text{diam}(\Omega)}{\epsilon}\right)^d\right).$$

Moreover, by noticing that $\max(1, \frac{L\,\text{diam}(\Omega)}{\epsilon}) \leq \frac{\max(M, L\,\text{diam}(\Omega))}{\epsilon}$ for any $\epsilon \leq M$, we further have that:

$$N_{[]}\left(\epsilon, \mathcal{T}, L_2(\mu)\right) \leq c(\Omega)\left(\frac{\max\left(M, L\,\text{diam}(\Omega)\right)}{\epsilon}\right)^d.$$

Finally, recalling that $D\left(\epsilon, \mathcal{T}, L_2(\mu)\right) \leq N_{[]}\left(\frac{\epsilon}{2}, \mathcal{T}, L_2(\mu)\right)$, we get that $D\left(\epsilon, \mathcal{T}, L_2(\mu)\right) \leq 2^d c(\Omega)\left(\frac{\max\left(M, L\,\text{diam}(\Omega)\right)}{\epsilon}\right)^d$. The desired bound follows after redefining $c(\Omega)$ to include the factor 2^d (*i.e.*, $c(\Omega) \to 2^d c(\Omega)$). □

Theorem E.2 (Maximal inequality for degenerate bounded and Lipschitz U-processes)**.** *Let k be either 1 or 2. Consider a parametric family $\mathcal{T} \coloneqq \left\{\otimes^k \mathcal{Z} \ni (z_1, \ldots, z_k) \mapsto t_\omega(z_1, \ldots, z_k) \in \mathbb{R} \mid \omega \in \Omega\right\}$ of real-valued functions over $\otimes^k \mathcal{Z}$ and indexed by a parameter $\omega \in \Omega$, where Ω is a compact subset of $\mathbb{R}^d$. For a given probability distribution $\mathbb{D}$ over $\mathcal{Z}$, assume that all elements t_ω are degenerate w.r.t. $\mathbb{D}$, meaning that:*

$$\begin{cases} \mathbb{E}_{\bar{z}\sim\mathbb{D}}[t_\omega(\bar{z})] = 0, & \text{if} \quad k = 1 \\ \mathbb{E}_{\bar{z}\sim\mathbb{D}}[t_\omega(z, \bar{z})] = \mathbb{E}_{\bar{z}\sim\mathbb{D}}[t_\omega(\bar{z}, z)] = 0, \quad \forall z \in \mathcal{Z}, & \text{if} \quad k = 2. \end{cases}$$

Furthermore, assume that all functions in $\mathcal{T}$ are uniformly bounded by a positive constant M and that there exists a positive constant L so that $\omega \mapsto t_\omega(z_1, \ldots, z_k)$ is L-Lipschitz for any $(z_1, \ldots, z_k) \in \otimes^k \mathcal{Z}$. Given i.i.d. samples $(z_i)_{1\leq i\leq n}$ from $\mathbb{D}$, consider the following U-statistic U_n^k:

$$U_n^k t_\omega \coloneqq \begin{cases} \frac{1}{n}\sum_{i=1}^n t_\omega(z_i), & \text{if} \quad k = 1 \\ \frac{1}{n(n-1)}\sum_{\substack{i,j=1 \\ i\neq j}}^n t_\omega(z_i, z_j), & \text{if} \quad k = 2. \end{cases}$$

Then, there exists a universal positive constant $c(\Omega)$ greater than 1 that depends only on Ω and d such that for any $p \in \{1, 2\}$:

$$\mathbb{E}_{\mathbb{D}}\left[\sup_{\omega\in\Omega}\left|U_n^k t_\omega\right|^p\right]^{\frac{1}{p}} \leq n^{-\frac{k}{2}} c(\Omega) \max\left(ML\,\text{diam}(\Omega), M^2\right)^{\frac{1}{2}}.$$

Proof. **Maximal inequality for degenerate U-processes.** We will first apply the general result in (Sherman, 1994, Maximal inequality) which controls $\mathbb{E}_{\mathbb{D}}\left[\sup_{\omega\in\Omega}\left|U_n^k t_\omega\right|\right]$ in terms of the packing number of $\mathcal{T}$. First note, by assumption, that the functions $t_\omega(z_1, \ldots, z_k)$ are uniformly bounded by a positive constant M. Therefore, the constant function $T(z_1, \ldots, z_k) \coloneqq M$ is an envelope for $\mathcal{T}$, *i.e.*, T satisfies $T(z_1, \ldots, z_k) \geq \sup_{\omega\in\Omega}|t_\omega(z_1, \ldots, z_k)|$ for any $(z_1, \ldots, z_k) \in \otimes^k \mathcal{Z}$. The envelope T is, a fortiori, square μ-integrable for any probability measure μ on $\otimes^k \mathcal{Z}$. Hence, we can apply (Sherman, 1994, Maximal inequality) with the choice T for the envelope function and set the integer m appearing in the result to $m = d$ to get the following bound:

$$\mathbb{E}_{\mathbb{D}}\left[\sup_{\omega\in\Omega}\left|U_n^k t_\omega\right|^p\right]^{\frac{1}{p}} \leq n^{-\frac{k}{2}}\Gamma\mathbb{E}\left[\|T\|_{\mu_n,2}\int_0^{\delta_n}\left(D\left(\epsilon\|T\|_{\mu_n,2}, \mathcal{T}, L_2(\mu_n)\right)\right)^{\frac{1}{2d}}\mathrm{d}\epsilon\right], \tag{20}$$

where Γ is a positive universal constant[2] that depends only on d and that we choose to be greater than 1, while μ_n are suitably chosen probability measures on $\otimes^k \mathcal{Z}$ that possibly depend on the samples $z_1, \ldots, z_n$ and other random variables, and $\delta_n\|T\|_{\mu_n,2} \coloneqq \sup_{\omega\in\Omega}\|t_\omega\|_{\mu_n,2}$. Here, the expectation symbol in the right-hand side is over all randomness on which

[2]The constant Γ appearing (Sherman, 1994, Maximal inequality) depends only on k, p and m, *i.e.*, $\Gamma \coloneqq g(k, p, m)$. Since, we are only interested in $k \leq 2$ and $p \leq 2$ and m is fixed to d, we choose Γ to be $\max_{1\leq k,p\leq 2} g(k, p, d)^{\frac{1}{p}}$, so that it is the same in all our cases.

μ_n might depend. Note that the original result in (Sherman, 1994, Maximal inequality) is stated using a slightly different definition of the packing number but which is still equivalent to the statement above in our setting[3].

In our setting, the envelope function is constant and equal to M, and by definition $\delta_n \leq 1$. Hence, the inequality in Equation (20) further becomes:

$$\mathbb{E}_{\mathbb{D}}\left[\sup_{\omega\in\Omega}\left|U_n^k t_\omega\right|^p\right]^{\frac{1}{p}} \leq n^{-\frac{k}{2}} M\Gamma\mathbb{E}_{\mathbb{D}}\left[\int_0^1 \left(D\Big(\epsilon\|T\|_{\mu_n,2}, \mathcal{T}, L_2(\mu_n)\Big)\right)^{\frac{1}{2dp}} \mathrm{d}\epsilon\right]. \tag{21}$$

We simply need to control the packing number $D\Big(\epsilon\|T\|_{\mu,2}, \mathcal{T}, L_2(\mu)\Big)$ independently of the probability measure μ.

Control on the packing number. We have shown that the constant function $T(z_1, \ldots, z_k) \coloneqq M$ is an envelope for $\mathcal{T}$ which is, a fortiori, square μ-integrable for any probability measure μ with $\|T\|_{\mu,2} = M < +\infty$. Moreover, the functions $\omega \mapsto t_\omega(z_1, \ldots, z_k)$ are L-Lipschitz for any $(z_1, \ldots, z_k) \in \otimes^k \mathcal{Z}$. We can therefore apply Proposition E.1 which ensures the existence of a positive constant $c(\Omega)$ greater than 1 and that depends only on Ω and d so that the following estimate on the ϵ-packing number of the class $\mathcal{T}$ w.r.t. $L_2(\mu)$ holds:

$$D\left(\epsilon\|T\|_{\mu,2}, \mathcal{T}, L_2(\mu)\right) \leq \underbrace{c(\Omega)\left(\max\left(\frac{L\,\mathrm{diam}(\Omega)}{M}, 1\right)\right)^d}_{A}\left(\frac{1}{\epsilon}\right)^d, \qquad \forall \epsilon \in (0, 1]. \tag{22}$$

Combining Equation (22) with Equation (21) yields:

$$\begin{aligned}\mathbb{E}_{\mathbb{D}}\left[\sup_{\omega\in\Omega}\left|U_n^k t_\omega\right|^p\right]^{\frac{1}{p}} &\leq n^{-\frac{k}{2}} M\Gamma\mathbb{E}_{\mathbb{D}}\left[\int_0^1 \left(A\epsilon^{-d}\right)^{\frac{1}{2dp}} \mathrm{d}\epsilon\right] = n^{-\frac{k}{2}} M\Gamma A^{\frac{1}{2dp}} \underbrace{\int_0^1 \epsilon^{-\frac{1}{2p}} \mathrm{d}\epsilon}_{\leq 2} \\ &\leq 2n^{-\frac{k}{2}}\Gamma c(\Omega)^{\frac{1}{2d}} \max\left(L\,\mathrm{diam}(\Omega), M^2\right)^{\frac{1}{2}},\end{aligned}$$

where, for the last inequality, we used that $A^{\frac{1}{2dp}} \leq A^{\frac{1}{2d}} = c(\Omega)^{\frac{1}{2d}} \max\left(\frac{L\,\mathrm{diam}(\Omega)}{M}, 1\right)^{\frac{1}{2}}$ since A is greater than 1. The desired result follows after redefining $c(\Omega)$ as $2\Gamma c(\Omega)^{\frac{1}{2d}}$ which is a positive constant that depends only on Ω and d. □

The following next propositions are particular instances of Theorem E.2 and will be used to obtain the main bounds.

Proposition E.3 (Maximal inequality for empirical processes)**.** *Consider a parametric family $\mathcal{T} \coloneqq \{\mathcal{Z} \ni z \mapsto t_\omega(z) \in \mathbb{R} \mid \omega \in \Omega\}$ of real-valued functions defined over a subset $\mathcal{Z}$ of a Euclidean space and indexed by a parameter $\omega \in \Omega$, where Ω is a compact subset of $\mathbb{R}^d$. Assume that all functions in $\mathcal{T}$ are uniformly bounded by a positive constant M and that there exists a positive constant L so that $\omega \mapsto t_\omega(z)$ is L-Lipschitz for any $z \in \mathcal{Z}$. Consider a probability distribution $\mathbb{D}$ over $\mathcal{Z}$ and let $(z_i)_{1\leq i\leq n}$ be i.i.d. samples drawn from $\mathbb{D}$, then there exists a positive constant $c(\Omega)$ greater than 1 that depends only on Ω and d, such that for any integer $p \in \{1, 2\}$:*

$$\mathbb{E}_{\mathbb{D}}\left[\sup_{\omega\in\Omega}\left|\mathbb{E}_{z\sim\mathbb{D}}[t_\omega(z)] - \frac{1}{n}\sum_{i=1}^n t_\omega(z_i)\right|^p\right]^{\frac{1}{p}} \leq \sqrt{\frac{1}{n}} c(\Omega)\max(ML\,\mathrm{diam}(\Omega), M^2)^{\frac{1}{2}}.$$

Proof. The upper-bound is a direct consequence of Theorem E.2. Indeed consider the family $\mathcal{S}$ of functions of the form $s_\omega(z) = t_\omega(z) - \mathbb{E}_{\bar{z}\sim\mathbb{D}}[t_\omega(\bar{z})]$, for any $z \in \mathcal{Z}$. Then clearly, the process $U_n^1 s_\omega \coloneqq \frac{1}{n}\sum_{i=1}^n s_\omega(z_i)$ is degenerate of order $k = 1$, and the family $\mathcal{S}$ is uniformly bounded by $2M$ and is $2L$-Lipschitz. Hence, by Theorem E.2, the following maximal inequality holds:

$$\mathbb{E}_{\mathbb{D}}\left[\sup_{\omega\in\Omega}\left|U_n^1 s_\omega\right|^p\right]^{\frac{1}{p}} \leq 2n^{-\frac{1}{2}} c(\Omega)\max\left(ML\,\mathrm{diam}(\Omega), M^2\right)^{\frac{1}{2}}.$$

[3] In (Sherman, 1994, Maximal inequality), the author considers a modified version of the ϵ-packing number (call it $\tilde{D}(\epsilon, \mathcal{T}, L_2(\mu))$) associated to $L_2(\mu)$ but endowed with a normalized version of the standard norm on $L_2(\mu)$: $\|f\|_\mu \coloneqq \frac{\|f\|_{\mu,2}}{\|T\|_{\mu,2}}$. Both numbers are related by the following identity: $\tilde{D}(\epsilon, \mathcal{T}, L_2(\mu)) = D(\epsilon\|T\|_{\mu,2}, \mathcal{T}, L_2(\mu))$, thus making the statement (20) equivalent to the original statement in (Sherman, 1994, Maximal inequality).

We get the desired upper-bound by redefining $c(\Omega)$ to contain the factor 2. □

Proposition E.4 (Maximal inequality for U-processes of order 2)**.** *Consider a parametric family $\mathcal{T} := \{\mathcal{Z} \times \mathcal{Z} \ni (z, z') \mapsto t_\omega(z, z') \in \mathbb{R} \mid \omega \in \Omega\}$ of real-valued functions indexed by a parameter $\omega \in \Omega$, where Ω is a compact subset of $\mathbb{R}^d$ and $\mathcal{Z}$ is a subset of a Euclidean space. Assume that the functions in $\mathcal{T}$ are symmetric in their arguments, i.e., $t_\omega(z, z') = t_\omega(z', z)$. Additionally, assume that all functions in $\mathcal{T}$ are uniformly bounded by a positive constant M and that there exists a positive constant L so that $\omega \mapsto t_\omega(z, z')$ is L-Lipschitz for any $(z, z') \in \mathcal{Z} \times \mathcal{Z}$. Consider a probability distribution $\mathbb{D}$ over $\mathcal{Z}$ and let $(z_i)_{1 \leq i \leq n}$ be i.i.d. samples drawn from $\mathbb{D}$, and define the following statistic:*

$$\tau_\omega := \mathbb{E}_{z,z' \sim \mathbb{D} \otimes \mathbb{D}} \left[t_\omega(z, z') \right] + \frac{1}{n^2} \sum_{i,j=1}^{n} t_\omega(z_i, z_j) - \frac{2}{n} \sum_{i=1}^{n} \mathbb{E}_{z \sim \mathbb{D}} \left[t_\omega(z, z_i) \right].$$

Then there exists a universal positive constant $c(\Omega)$ greater than 1 that depends only on Ω and d such that:

$$\mathbb{E}_{\mathbb{D}} \left[\sup_{\omega \in \Omega} |\tau_\omega| \right] \leq \frac{1}{n} c(\Omega) \max \left(ML \operatorname{diam}(\Omega), M^2 \right)^{\frac{1}{2}}.$$

Proof. The proof will proceed by first decomposing τ_ω into a sum of a degenerate U-process and a term of order $O(\frac{1}{n})$. The maximal inequality for degenerate U-processes from (Sherman, 1994) will be employed to obtain the desired bound.

Decomposition of τ_ω. Consider the following function defined over $\mathcal{Z} \times \mathcal{Z}$ and indexed by elements $\omega \in \Omega$:

$$s_\omega(z, z') = t_\omega(z, z') - \mathbb{E}_{\bar{z} \sim \mathbb{D}} \left[t_\omega(z, \bar{z}) \right] - \mathbb{E}_{\bar{z} \sim \mathbb{D}} \left[t_\omega(\bar{z}, z') \right] + \mathbb{E}_{\{\bar{z}, \underline{z}\} \sim \mathbb{D} \otimes \mathbb{D}} \left[t_\omega(\bar{z}, \underline{z}) \right]. \tag{23}$$

By direct calculation, we decompose τ_ω into two higher order terms and a third term, $U_n^2 s_\omega$, involving s_ω, which happens to be a U-statistic:

$$\tau_\omega = \overbrace{\frac{1}{(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} s_\omega(z_i, z_j)}^{U_n^2 s_\omega} - \frac{1}{n^2(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} t_\omega(z_i, z_j) + \frac{1}{n^2} \sum_{i=1}^{n} t_\omega(z_i, z_i).$$

Using the triangle inequality in the above equality and recalling that, by assumption, $t_\omega(z, z')$ is uniformly bounded by a positive constant M, it follows that:

$$|\tau_\omega| \leq \left| U_n^2 s_\omega \right| + \frac{1}{n^2(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} |t_\omega(z_i, z_j)| + \frac{1}{n^2} \sum_{i=1}^{n} |t_\omega(z_i, z_i)| \leq \left| U_n^2 s_\omega \right| + \frac{2M}{n}.$$

Furthermore, taking the supremum over ω followed by the expectation over samples yields:

$$\mathbb{E}_{\mathbb{D}} \left[\sup_{\omega \in \Omega} |\tau_\omega| \right] \leq \mathbb{E}_{\mathbb{D}} \left[\sup_{\omega \in \Omega} \left| U_n^2 s_\omega \right| \right] + \frac{2M}{n}. \tag{24}$$

Hence, it only remains to control the first term in the above inequality. To this end, we will use a maximal inequality for degenerate U-processes due to (Sherman, 1994).

Maximal inequality for degenerate U-processes. We will first check that $U_n^2 s_\omega$ is a degenerate statistic for a given $\omega \in \Omega$. To this end, simple calculations show that for any z in $\mathcal{Z}$:

$$\mathbb{E}_{\bar{z} \sim \mathbb{D}}[s_\omega(z, \bar{z})] = \mathbb{E}_{\bar{z} \sim \mathbb{D}}[s_\omega(\bar{z}, z)] = 0.$$

The above equalities precisely ensure that $U_n^2 s_\omega$ is a degenerate U-statistic for $\mathbb{D}$. Consider now the family $\mathcal{S} := \{\mathcal{Z} \times \mathcal{Z} \ni (z, z') \mapsto s_\omega(z, z') \in \mathbb{R} \mid \omega \in \Omega\}$. We show that $\mathcal{S}$ is uniformly bounded and Lipschitz which allows to directly apply the result stated in Theorem E.2, which is a special case of the more general result in (Sherman, 1994, Maximal inequality). First note, by assumption, that the functions $t_\omega(z, z')$ are uniformly bounded by a positive constant M. Hence, using Equation (23), it follows that $s_\omega(z, z')$ is uniformly bounded by $4M$. Moreover, the functions $\omega \mapsto t_\omega(z, z')$ are

L-Lipschitz for any z, z' in $\mathcal{Z}$. Hence, from Equation (23), we directly have that $\omega \mapsto s_\omega(z, z')$ is $4L$-Lipschitz for any $z, z' \in \mathcal{Z}$. We can directly apply Theorem E.2 with $k = 2$ and $p = 1$ to $\mathcal{S}$ and get the following maximal inequality:

$$\mathbb{E}_{\mathbb{D}}\left[\sup_{\omega\in\Omega}\left|U_n^2 s_\omega\right|\right] \leq 4n^{-1}c(\Omega)\max\left(ML\operatorname{diam}(\Omega), M^2\right)^{\frac{1}{2}}.$$

We obtain an upper-bound on $\mathbb{E}_{\mathbb{D}}[\sup_{\omega\in\Omega}|\tau_\omega|]$ by combining the above inequality with Equation (24), then noticing that $2M \leq 2c(\Omega)\max\left(L\operatorname{diam}(\Omega), M\right)$ so that:

$$\mathbb{E}_{\mathbb{D}}\left[\sup_{\omega\in\Omega}|\tau_\omega|\right] \leq 6n^{-1}c(\Omega)\max\left(ML\operatorname{diam}(\Omega), M^2\right)^{\frac{1}{2}}.$$

Finally, the desired result follows by redefining $c(\Omega)$ to include the factor 6 in the above inequality. □

F. Differentiability Results

The proofs of Propositions A.1 and A.2 are direct applications of the following more general result.

Proposition F.1. *Let $\mathcal{U}$ be an open non-trivial subset of $\mathbb{R}^d$. Consider a real-valued function $\ell : (\omega, v, y) \mapsto \ell(\omega, v, y)$ defined on $\mathcal{U} \times \mathbb{R} \times \mathcal{Y}$ that is of class C^3 jointly in (ω, v) and whose derivatives are jointly continuous in (ω, v, y). For a given probability distribution $\mathbb{D}$ over $\mathcal{X} \times \mathcal{Y}$, consider the following functional defined over $\mathcal{U} \times \mathcal{H}$:*

$$L(\omega, h) \coloneqq \mathbb{E}_{\mathbb{D}}[\ell(\omega, h(x), y)].$$

*Under Assumptions **(A)** and **(B)**, the following properties hold for L:*

- *L admits finite values for any $(\omega, h) \in \mathcal{U} \times \mathcal{H}$.*
- *$(\omega, h) \mapsto L(\omega, h)$ is Fréchet differentiable with partial derivatives $\partial_\omega L(\omega, h)$ and $\partial_h L(\omega, h)$ at any point $(\omega, h) \in \mathcal{U} \times \mathcal{H}$ given by:*

$$\begin{aligned}\partial_\omega L(\omega, h) &= \mathbb{E}_{\mathbb{D}}[\partial_\omega \ell(\omega, h(x), y)] \in \mathbb{R}^d,\\ \partial_h L(\omega, h) &= \mathbb{E}_{\mathbb{D}}[\partial_v \ell(\omega, h(x), y)K(x, \cdot)] \in \mathcal{H}.\end{aligned}$$

- *The map $(\omega, h) \mapsto \partial_h L(\omega, h)$ is differentiable. Moreover, for any $(\omega, h) \in \mathcal{U} \times \mathcal{H}$, its partial derivatives $\partial^2_{\omega,h} L(\omega, h)$ and $\partial^2_{\omega,h} L(\omega, h)$ at (ω, h) are Hilbert-Schmidt operators given by:*

$$\begin{aligned}\partial^2_{\omega,h} L(\omega, h) &= \mathbb{E}_{\mathbb{D}}\left[\partial^2_{\omega,v} \ell(\omega, h(x), y)K(x, \cdot)\right] \in \mathcal{L}(\mathcal{H}, \mathbb{R}^d),\\ \partial^2_h L(\omega, h) &= \mathbb{E}_{\mathbb{D}}\left[\partial^2_v \ell(\omega, h(x), y)K(x, \cdot) \otimes K(x, \cdot)\right] \in \mathcal{L}(\mathcal{H}, \mathcal{H}).\end{aligned}$$

Proof. **Finite values.** Fix $\omega \in \mathcal{U}$ and $h \in \mathcal{H}$. We will first show that h is bounded on $\mathcal{X}$. By the reproducing property, we know that $|h(x)| \leq \|h\|_{\mathcal{H}}\sqrt{K(x, x)}$ for any $x \in \mathcal{X}$. Moreover, the kernel K is bounded by a constant κ thanks to Assumption **(A)**. Consequently, $|h(x)|$ is upper-bounded by $\|h\|_{\mathcal{H}}\sqrt{\kappa}$ for any $x \in \mathcal{X}$.

Denote by $\mathcal{I}$ the compact interval defined as $\mathcal{I} = [-\|h\|_{\mathcal{H}}\sqrt{\kappa}, \|h\|_{\mathcal{H}}\sqrt{\kappa}]$. By Assumption **(B)**, the set $\mathcal{Y}$ is compact so that $\mathcal{I} \times \mathcal{Y}$ is also compact. Moreover, we know, by assumption on ℓ, that $(\omega, v, y) \mapsto \ell(\omega, v, y)$ is continuous on $\mathcal{U} \times \mathbb{R} \times \mathcal{Y}$. Therefore, $(v, y) \mapsto \ell(\omega, v, y)$ must be bounded by some finite constant C on the compact set $\mathcal{I} \times \mathcal{Y}$. This allows to deduce that $(x, y) \mapsto \ell(\omega, h(x), y)$ is bounded by C for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and a fortiori $\mathbb{D}$-integrable, which shows that $L(\omega, h)$ is finite.

Fréchet differentiability of L. Let $(\omega, h) \in \mathcal{U} \times \mathcal{H}$. Consider $(\omega_j, h_j)_{j\geq 1}$ a sequence of elements in $\mathcal{U} \times \mathcal{H}$ converging to it, *i.e.*, $(\omega_j, h_j) \to (\omega, h)$ with $(\omega_j, h_j) \neq (\omega, h)$ for any $j \geq 0$. Define the sequence of functions $r_j : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ as follows:

$$r_j(x, y) = \frac{\ell\left(\omega_j, h_j(x), y\right) - \ell\left(\omega, h(x), y\right) - \left\langle \partial_v \ell\left(\omega, h(x), y\right) K(x, \cdot), h_j - h\right\rangle_{\mathcal{H}} - \left\langle \partial_\omega \ell(\omega, h(x), y), \omega_j - \omega\right\rangle}{\|(\omega_j, h_j) - (\omega, h)\|}. \quad (25)$$

We will first show that $\mathbb{E}_{\mathbb{D}}[|r_j(x,y)|]$ convergence to 0 by the dominated convergence theorem (Rudin, 1987, Theorem 1.34). By the reproducing property, note that $\ell(\omega, h(x), y) = \ell(\omega, \langle h, K(x,\cdot)\rangle_{\mathcal{H}}, y)$. Hence, since ℓ is jointly differentiable in (ω, v) for any y, it follows that $(\omega, h) \mapsto \ell(\omega, h(x), y)$ is also differentiable for any (x, y) by composition with the evaluation map $(\omega, h) \mapsto (\omega, \langle h, K(x,\cdot)\rangle_{\mathcal{H}}$ which is differentiable. Hence, the sequence $r_j(x,y)$ converges to 0 for any $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Moreover, by the mean-value theorem, there exists $0 \leq c_j \leq 1$ such that:

$$r_j(x,y) = \frac{\left\langle \left(\partial_v \ell\left(\bar{\omega}_j, \bar{h}_j(x), y\right) - \partial_v \ell\left(\omega, h(x), y\right)\right) K(x,\cdot), h_j - h\right\rangle_{\mathcal{H}} - \left\langle \partial_\omega \ell\left(\bar{\omega}_j, \bar{h}_j(x), y\right) - \partial_\omega \ell\left(\omega, h(x), y\right), \omega_j - \omega\right\rangle}{\|(\omega_j, h_j) - (\omega, h)\|},$$

where $(\bar{\omega}_j, \bar{h}_j) \coloneqq (1-c_j)(\omega, h) + c_j(\omega_j, h_j)$. We will show that $r_j(x,y)$ are bounded for j large enough. We first construct a compact set that will contain all elements of the form $(\omega_j, h_j(x), y)$ and $(\bar{\omega}_j, \bar{h}_j(x), y)$ for all j large enough. Since ω is an element in the open set $\mathcal{U}$, there exists a closed ball $\mathcal{B}(\omega, R)$ centered in ω and with some radius R small enough so that $\mathcal{B}(\omega, R)$ is included in $\mathcal{U}$. For all j large enough, ω_j and $\bar{\omega}_j$ belong to $\mathcal{B}(\omega, R)$ as these sequences converge to ω. Moreover, h_j and $\bar{h}_j$ are convergent sequences. Consequently, they must be bounded by some constant B. By the reproducing property, and recalling that the kernel K is bounded by κ by Assumption **(A)**, it follows that $\max(|h_j(x)|, \left|\bar{h}_j(x)\right|) \leq B\kappa$. Consider now the set $\mathcal{W} \coloneqq \mathcal{B}(\omega, R) \times \mathcal{B}_1(0, B\kappa) \times \mathcal{Y}$ which is a product of compact sets (recalling that $\mathcal{Y}$ is compact by Assumption **(B)**), where $\mathcal{B}_1(0, B\kappa)$ is the closed ball in $\mathbb{R}$ centered at 0 and of radius $B\kappa$. For j large enough, we have established that $(\omega_j, h_j(x), y)$ and $(\bar{\omega}_j, \bar{h}_j(x), y)$ belong to $\mathcal{W}$ for any $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Since, by assumption on ℓ, $\partial_v \ell(\omega, v, y)$ and $\partial_\omega \ell(\omega, v, y)$ are continuous, they must be bounded on the compact set $\mathcal{W}$ by some constant C. This allows to deduce from the expression of $r_j(x,y)$ above that $r_j(x,y)$ is bounded, and a fortiori dominated by an integrable function (a constant function). We then deduce that $\mathbb{E}_{\mathbb{D}}[|r_j(x,y)|]$ converges to 0 by application of the dominated convergence theorem (Rudin, 1987, Theorem 1.34).

Recalling Equation (25), $\mathbb{E}_{\mathbb{D}}[r_j(x,y)]$ admits the following expression:

$$\mathbb{E}_{\mathbb{D}}[r_j(x,y)] = \frac{L(\omega_j, h_j) - L(\omega, h) - \mathbb{E}_{\mathbb{D}}\left[\langle \partial_v \ell\left(\omega, h(x), y\right) K(x,\cdot), h_j - h\rangle_{\mathcal{H}}\right] - \langle \mathbb{E}_{\mathbb{D}}[\partial_\omega \ell(\omega, h(x), y)], \omega_j - \omega\rangle}{\|(\omega_j, h_j) - (\omega, h)\|}. \quad (26)$$

The convergence to 0 of the above expression precisely means that L is differentiable at (ω, h) provided that the linear form $g \mapsto \mathbb{E}_{\mathbb{D}}[\langle \partial_v \ell\left(\omega, h(x), y\right) K(x,\cdot), g\rangle_{\mathcal{H}}]$ is bounded. To establish this fact, consider the RKHS-valued function $(x,y) \mapsto \partial_v \ell\left(\omega, h(x), y\right) K(x,\cdot)$. This function is Bochner-integrable in the sense that $\mathbb{E}_{\mathbb{D}}[\|\partial_v \ell\left(\omega, h(x), y\right) K(x,\cdot)\|_{\mathcal{H}}]$ is finite (Diestel & Uhl, 1977, Definition 1, Chapter 2). Indeed, we have the following:

$$\mathbb{E}_{\mathbb{D}}[\|\partial_v \ell\left(\omega, h(x), y\right) K(x,\cdot)\|_{\mathcal{H}}] \coloneqq \mathbb{E}_{\mathbb{D}}\left[|\partial_v \ell\left(\omega, h(x), y\right)| \sqrt{K(x,x)}\right] \leq \sqrt{\kappa} \mathbb{E}_{\mathbb{D}}[|\partial_v \ell\left(\omega, h(x), y\right)|] < +\infty,$$

where, for the inequality, we used that $(x,y) \mapsto \partial_v \ell(\omega, h(x), y)$ is bounded as shown previously. Consequently, $\mathbb{E}[\partial_v \ell\left(\omega, h(x), y\right) K(x,\cdot)]$ is an element in $\mathcal{H}$ satisfying:

$$\langle \mathbb{E}[\partial_v \ell\left(\omega, h(x), y\right) K(x,\cdot)], g\rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{D}}[\langle \partial_v \ell\left(\omega, h(x), y\right) K(x,\cdot), g\rangle_{\mathcal{H}}], \forall g \in \mathcal{H}.$$

The above property follows from (Diestel & Uhl, 1977, Theorem 6, Chapter 2) for Bochner-integrable functions that allows exchanging the integral and the application of a continuous linear map (here the scalar product with an element g). The above identity establishes that $g \mapsto \mathbb{E}_{\mathbb{D}}[\langle \partial_v \ell\left(\omega, h(x), y\right) K(x,\cdot), g\rangle_{\mathcal{H}}]$ is bounded and provides the desired expression for $\partial_h L(\omega, h)$. The expression for $\partial_\omega L(\omega, h)$ directly follows from the last term in Equation (26).

Fréchet differentiability of $\partial_h L$. We use the same proof strategy as for the differentiability of L.

Let $(\omega, h) \in \mathcal{U} \times \mathcal{H}$. Consider $(\omega_j, h_j)_{j \geq 1}$ a sequence of elements in $\mathcal{U} \times \mathcal{H}$ converging to it, *i.e.*, $(\omega_j, h_j) \to (\omega, h)$ with $(\omega_j, h_j) \neq (\omega, h)$ for any $j \geq 0$. Define the sequence of functions $s_j : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$ as follows:

$$\begin{aligned} \|(\omega_j, h_j) - (\omega, h)\| \, s_j(x,y) =& (\partial_v \ell\left(\omega_j, h_j(x), y\right) - \partial_v \ell\left(\omega, h(x), y\right)) K(x,\cdot) \\ & - \partial_v^2 \ell\left(\omega, h(x), y\right) K(x,\cdot) \otimes K(x,\cdot)(h_j - h) - (\omega_j - \omega)^\top \partial^2_{\omega,v} \ell(\omega, h(x), y) K(x,\cdot). \end{aligned} \quad (27)$$

We will first show that $\mathbb{E}_{\mathbb{D}}\left[\|s_j(x,y)\|_{\mathcal{H}}\right]$ convergence to 0 by the dominated convergence theorem for Bochner-integrable functions (Diestel & Uhl, 1977, Theorem 3, Chapter 2). By the reproducing property, note that $\partial_v \ell(\omega, h(x), y) K(x,\cdot) =$

$\partial_v \ell(\omega, \langle h, K(x,\cdot)\rangle_{\mathcal{H}}, y)K(x,\cdot)$. Hence, since $(\omega, v) \mapsto \partial_v \ell(\omega, v, y)$ is jointly differentiable in (ω, v) for any y, it follows that $(\omega, h) \mapsto \partial_v \ell(\omega, h(x), y)K(x,\cdot)$ is also differentiable for any (x, y) by composition with the evaluation map $(\omega, h) \mapsto (\omega, \langle h, K(x,\cdot)\rangle_{\mathcal{H}}$ which is differentiable. Hence, the sequence $s_j(x, y)$ converges to 0 for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Moreover, by the mean-value theorem, there exists $0 \leq c_j \leq 1$ such that:

$$\begin{aligned}\|(\omega_j, h_j) - (\omega, h)\|\, s_j(x,y) =& \partial_v^2 \ell\left(\bar{\omega}_j, \bar{h}_j(x), y\right) K(x,\cdot) \otimes K(x,\cdot)(h_j - h) + (\omega_j - \omega)^\top \partial^2_{\omega,v} \ell(\bar{\omega}_j, \bar{h}_j(x), y)K(x,\cdot) \\ &- \partial_v^2 \ell\left(\omega, h(x), y\right) K(x,\cdot) \otimes K(x,\cdot)(h_j - h) - (\omega_j - \omega)^\top \partial^2_{\omega,v} \ell(\omega, h(x), y)K(x,\cdot),\end{aligned}$$

where $(\bar{\omega}_j, \bar{h}_j) \coloneqq (1 - c_j)(\omega, h) + c_j(\omega_j, h_j)$. Using the same construction as for the Fréchet differentiability, we find a compact set $\mathcal{W}$ containing all elements $(\omega_j, h_j(x), y)$ and $(\bar{\omega}_j, \bar{h}_j(x), y)$ for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and all j large enough. On such set, $\partial_v^2 \ell(\omega, v, y)$ and $\partial^2_{\omega,v} \ell(\omega, v, y)$ are bounded by some constant C. Consequently, we can write:

$$\begin{aligned}\|(\omega_j, h_j) - (\omega, h)\|\, \|s_j(x,y)\|_{\mathcal{H}} \leq& 2C\|K(x,\cdot) \otimes K(x,\cdot)(h_j - h)\|_{\mathcal{H}} + 2C\|\omega_j - \omega\| \|K(x,\cdot)\|_{\mathcal{H}} \\ \leq& 2C\kappa\|h_j - h\|_{\mathcal{H}} + 2C\sqrt{\kappa}\|\omega_j - \omega\|.\end{aligned}$$

This already establishes that $s_j(x, y)$ is bounded so that $\mathbb{E}_{\mathbb{D}}\left[\|s_j(x,y)\|_{\mathcal{H}}\right]$ converges to 0 by application of the dominated convergence theorem. Recalling Equation (27), $\mathbb{E}_{\mathbb{D}}[s_j(x, y)]$ admits the following expression:

$$\begin{aligned}\|(\omega_j, h_j) - (\omega, h)\|\, \mathbb{E}_{\mathbb{D}}[s_j(x,y)] =& \partial_h L(\omega_j, h_j) - \partial_h L(\omega, h) - \mathbb{E}_{\mathbb{D}}\left[\partial_v^2 \ell\left(\omega, h(x), y\right) K(x,\cdot) \otimes K(x,\cdot)(h_j - h)\right] \\ &- (\omega_j - \omega)^\top \mathbb{E}_{\mathbb{D}}\left[\partial^2_{\omega,v} \ell(\omega, h(x), y)K(x,\cdot)\right].\end{aligned}$$

The convergence to 0 of the above expression precisely means that L is differentiable at (ω, h) provided that: (1) $\mathbb{E}_{\mathbb{D}}\left[\partial^2_{\omega,v} \ell(\omega, h(x), y)K(x,\cdot)\right]$ is an element in $\mathcal{H}^d$, and (2) the linear map $g \mapsto \mathbb{E}_{\mathbb{D}}\left[\partial_v^2 \ell\left(\omega, h(x), y\right)(K(x,\cdot) \otimes K(x,\cdot))g\right]$ is bounded. Using the same strategy to establish Bochner's inetgrability of $(x, y) \mapsto \partial_v \ell\left(\omega, h(x), y\right) K(x,\cdot)$, we can show that $(x, y) \mapsto \partial^2_{\omega,v} \ell\left(\omega, h(x), y\right) K(x,\cdot)$ is also Bochner-integrable so that $\mathbb{E}_{\mathbb{D}}\left[\partial^2_{\omega,v} \ell(\omega, h(x), y)K(x,\cdot)\right]$ is indeed an element in $\mathcal{H}^d$. This also establishes the expression of $\partial_{\omega,h} L(\omega, h)$. Similarly, we consider the operator-valued function $\xi : (x, y) \mapsto \partial_v^2 \ell\left(\omega, h(x), y\right) K(x,\cdot) \otimes K(x,\cdot)$ with values in the space of Hilbert-Schmidt operators on $\mathcal{H}$. The Hilbert-Schmidt (HS) norm of such function satisfies the following inequality:

$$\mathbb{E}_{\mathbb{D}}\left[\left\|\partial_v^2 \ell(\omega, h(x), y)K(x,\cdot) \otimes K(x,\cdot)\right\|_{\mathrm{HS}}\right] \coloneqq \mathbb{E}_{\mathbb{D}}\left[\left|\partial_v^2 \ell(\omega, h(x), y)\right| K(x,x)\right] \leq \kappa C < +\infty.$$

Therefore, the function ξ is Bochner-integrable, so that $\mathbb{E}_{\mathbb{D}}\left[\partial_v^2 \ell(\omega, h(x), y)K(x,\cdot) \otimes K(x,\cdot)\right]$ is a Hilbert-Schmidt operator satisfying:

$$\mathbb{E}_{\mathbb{D}}\left[\partial_v^2 \ell(\omega, h(x), y)K(x,\cdot) \otimes K(x,\cdot)\right] g = \mathbb{E}_{\mathbb{D}}\left[\partial_v^2 \ell\left(\omega, h(x), y\right)(K(x,\cdot) \otimes K(x,\cdot))g\right], \forall g \in \mathcal{H}.$$

The above property follows from (Diestel & Uhl, 1977, Theorem 6, Chapter 2) for Bochner-integrable functions that allows exchanging the integral and the application of a continuous linear map (here the scalar product with an element g). Hence, from the above identity we deduce the desired expression for $\partial_h^2 L(\omega, h)$. □

G. Auxiliary Technical Lemmas

Lemma G.1. *Let A and A' be two bounded operators from $\mathcal{H}$ to $\mathbb{R}^d$, and B and B' be two bounded and invertible operators from $\mathcal{H}$ to itself. Assume that $B \geq \lambda \operatorname{Id}_{\mathcal{H}}$ and $B' \geq \lambda \operatorname{Id}_{\mathcal{H}}$. Then, the following inequalities hold:*

$$\begin{aligned}\left\|AB^{-1} - A'(B')^{-1}\right\|_{\mathrm{op}} &\leq \frac{\|A\|_{\mathrm{op}}}{\lambda^2}\|B - B'\|_{\mathrm{op}} + \frac{1}{\lambda}\|A - A'\|_{\mathrm{op}}, \\ \left\|AB^{-1}\right\|_{\mathrm{op}} \leq \lambda^{-1}\|A\|_{\mathrm{op}}, &\qquad \left\|A'(B')^{-1}\right\|_{\mathrm{op}} \leq \lambda^{-1}\|A'\|_{\mathrm{op}}.\end{aligned}$$

Proof. By the triangle inequality and the sub-multiplicative property of the operator norm $\|\cdot\|_{\mathrm{op}}$, we have:

$$\begin{aligned}\left\|AB^{-1} - A'(B')^{-1}\right\|_{\mathrm{op}} &\leq \left\|AB^{-1} - A(B')^{-1}\right\|_{\mathrm{op}} + \left\|A(B')^{-1} - A'(B')^{-1}\right\|_{\mathrm{op}} \\ &\leq \left\|A\left(B^{-1} - (B')^{-1}\right)\right\|_{\mathrm{op}} + \left\|(A - A')(B')^{-1}\right\|_{\mathrm{op}} \\ &\leq \|A\|_{\mathrm{op}} \left\|B^{-1} - (B')^{-1}\right\|_{\mathrm{op}} + \|A - A'\|_{\mathrm{op}} \left\|(B')^{-1}\right\|_{\mathrm{op}} \\ &\leq \|A\|_{\mathrm{op}} \left\|B^{-1}(B' - B)(B')^{-1}\right\|_{\mathrm{op}} + \|A - A'\|_{\mathrm{op}} \left\|(B')^{-1}\right\|_{\mathrm{op}} \\ &\leq \|A\|_{\mathrm{op}} \left\|B^{-1}\right\|_{\mathrm{op}} \|B' - B\|_{\mathrm{op}} \left\|(B')^{-1}\right\|_{\mathrm{op}} + \|A - A'\|_{\mathrm{op}} \left\|(B')^{-1}\right\|_{\mathrm{op}}. \qquad (28)\end{aligned}$$

Since $B \geq \lambda \operatorname{Id}_{\mathcal{H}}$ and $B' \geq \lambda \operatorname{Id}_{\mathcal{H}}$, we obtain:

$$\left\|B^{-1}\right\|_{\mathrm{op}} \leq \frac{1}{\lambda} \quad \text{and} \quad \left\|(B')^{-1}\right\|_{\mathrm{op}} \leq \frac{1}{\lambda}.$$

Substituting these into Equation (28), we get:

$$\left\|AB^{-1} - A'(B')^{-1}\right\|_{\mathrm{op}} \leq \frac{\|A\|_{\mathrm{op}}}{\lambda^2}\|B - B'\|_{\mathrm{op}} + \frac{1}{\lambda}\|A - A'\|_{\mathrm{op}}.$$

This proves the first inequality. The remaining two inequalities follow directly from the sub-multiplicative property of the operator norm $\|\cdot\|_{\mathrm{op}}$ and the assumptions $B \geq \lambda \operatorname{Id}_{\mathcal{H}}$ and $B' \geq \lambda \operatorname{Id}_{\mathcal{H}}$. □

Lemma G.2. *Let $f : \mathcal{H} \to \mathbb{R}$ be a λ-strongly convex and Fréchet differentiable function. Denote by $h^\star \in \mathcal{H}$ its minimizer. Then, for any $h \in \mathcal{H}$, the following holds:*

$$\|h - h^\star\|_{\mathcal{H}} \leq \frac{1}{\lambda} \|\partial_h f(h)\|_{\mathcal{H}} .$$

Proof. Let $h \in \mathcal{H}$.

Case 1: $h = h^\star$**.** The proof is straightforward.

Case 2: $h \neq h^\star$**.** Given that f is λ-strongly convex, we have:

$$f(h) - f(h^\star) \geq \langle \partial_h f(h^\star), h - h^\star \rangle_{\mathcal{H}} + \frac{\lambda}{2} \|h - h^\star\|_{\mathcal{H}}^2 ,$$
$$\text{and } f(h^\star) - f(h) \geq \langle \partial_h f(h), h^\star - h \rangle_{\mathcal{H}} + \frac{\lambda}{2} \|h - h^\star\|_{\mathcal{H}}^2 .$$

After summing these two inequalities, noticing that $\partial_h f(h^\star) = 0$, and rearranging the terms, we obtain:

$$\langle \partial_h f(h), h - h^\star \rangle_{\mathcal{H}} \geq \lambda \|h - h^\star\|_{\mathcal{H}}^2 .$$

After using the Cauchy-Schwarz inequality, we get:

$$\|\partial_h f(h)\|_{\mathcal{H}} \|h - h^\star\|_{\mathcal{H}} \geq \lambda \|h - h^\star\|_{\mathcal{H}}^2 .$$

Dividing by $\lambda \|h - h^\star\|_{\mathcal{H}} \neq 0$ concludes the proof. □

Lemma G.3. *Let $\mathcal{X}$ be a subset of $\mathbb{R}^p$, $\mathcal{Y}$ be a subset of $\mathbb{R}^q$, and $\mathbb{D}$ be a probability distribution over $\mathcal{X} \times \mathcal{Y}$. Given i.i.d. samples $(x_i, y_i)_{1 \leq i \leq n}$ drawn from $\mathbb{D}$, consider a function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ of class C^1 such that the operator $A : \mathcal{H} \to \mathcal{H}$ defined as:*

$$A \coloneqq \mathbb{E}_{(x,y)\sim\mathbb{D}}[g(x,y)K(x,\cdot) \otimes K(x,\cdot)] - \frac{1}{n}\sum_{i=1}^{n} g(x_i,y_i)K(x_i,\cdot) \otimes K(x_i,\cdot)$$

is Hilbert-Schmidt. Then, the following holds:

$$\|A\|_{\mathrm{HS}}^2 = \mathbb{E}_{(x,y),(x',y')\sim\mathbb{D}\otimes\mathbb{D}}\Big[g(x,y)g(x',y')K^2(x,x')\Big] + \frac{1}{n^2}\sum_{i,j=1}^{n} g(x_i,y_i)g(x_j,y_j)K^2(x_i,x_j)$$
$$- \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{(x,y)\sim\mathbb{D}}\Big[g(x_i,y_i)g(x,y)K^2(x,x_i)\Big].$$

Proof. Define $s \coloneqq \mathbb{E}_{(x,y)\sim\mathbb{D}}\left[g(x,y)K(x,\cdot) \otimes K(x,\cdot)\right]$ and $\hat{s} \coloneqq \frac{1}{n}\sum_{i=1}^{n} g(x_i,y_i)K(x_i,\cdot) \otimes K(x_i,\cdot)$. We have:

$$\|A\|_{\mathrm{HS}}^2 = \|s - \hat{s}\|_{\mathrm{HS}}^2 = \|s\|_{\mathrm{HS}}^2 + \|\hat{s}\|_{\mathrm{HS}}^2 - 2\langle s, \hat{s}\rangle_{\mathrm{HS}} . \tag{29}$$

Next, we compute each of the following quantities: $\|s\|_{\mathrm{HS}}^2$, $\|\hat{s}\|_{\mathrm{HS}}^2$, and $\langle s, \hat{s}\rangle_{\mathrm{HS}}$, separately. Simple calculations yield:

$$
\begin{aligned}
\|s\|_{\mathrm{HS}}^2 &= \left\|\mathbb{E}_{(x,y)\sim\mathbb{D}}\left[g(x,y)K(x,\cdot)\otimes K(x,\cdot)\right]\right\|_{\mathrm{HS}}^2 \\
&= \left\langle \mathbb{E}_{(x,y)\sim\mathbb{D}}\left[g(x,y)K(x,\cdot)\otimes K(x,\cdot)\right], \mathbb{E}_{(x',y')\sim\mathbb{D}}\left[g(x',y')K(x',\cdot)\otimes K(x',\cdot)\right]\right\rangle_{\mathrm{HS}} \\
&= \mathbb{E}_{(x,y),(x',y')\sim\mathbb{D}\otimes\mathbb{D}}\Big[g(x,y)g(x',y')\left\langle K(x,\cdot)\otimes K(x,\cdot), K(x',\cdot)\otimes K(x',\cdot)\right\rangle_{\mathrm{HS}}\Big] \\
&= \mathbb{E}_{(x,y),(x',y')\sim\mathbb{D}\otimes\mathbb{D}}\Big[g(x,y)g(x',y')K^2(x,x')\Big], \\
\|\hat{s}\|_{\mathrm{HS}}^2 &= \frac{1}{n^2}\left\|\sum_{i=1}^n g(x_i,y_i)K(x_i,\cdot)\otimes K(x_i,\cdot)\right\|_{\mathrm{HS}}^2 \\
&= \frac{1}{n^2}\left\langle \sum_{i=1}^n g(x_i,y_i)K(x_i,\cdot)\otimes K(x_i,\cdot), \sum_{j=1}^n g(x_j,y_j)K(x_j,\cdot)\otimes K(x_j,\cdot)\right\rangle_{\mathrm{HS}} \\
&= \frac{1}{n^2}\sum_{i,j=1}^n \partial_v^2 g(x_i,y_i)g(x_j,y_j)\left\langle K(x_i,\cdot)\otimes K(x_i,\cdot), K(x_j,\cdot)\otimes K(x_j,\cdot)\right\rangle_{\mathrm{HS}} \\
&= \frac{1}{n^2}\sum_{i,j=1}^n g(x_i,y_i)g(x_j,y_j)K^2(x_i,x_j), \\
\langle s,\hat{s}\rangle_{\mathrm{HS}} &= \left\langle \mathbb{E}_{(x,y)\sim\mathbb{D}}\left[g(x,y)K(x,\cdot)\otimes K(x,\cdot)\right], \frac{1}{n}\sum_{i=1}^n g(x_i,y_i)K(x_i,\cdot)\otimes K(x_i,\cdot)\right\rangle_{\mathrm{HS}} \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{(x,y)\sim\mathbb{D}}\left[g(x_i,y_i)g(x,y)\left\langle K(x,\cdot)\otimes K(x,\cdot), K(x_i,\cdot)\otimes K(x_i,\cdot)\right\rangle_{\mathrm{HS}}\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{(x,y)\sim\mathbb{D}}\Big[g(x_i,y_i)g(x,y)K^2(x,x_i)\Big].
\end{aligned}
$$

After substituting the obtained results into Equation (29) and rearranging, we obtain:

$$
\begin{aligned}
\|A\|_{\mathrm{HS}}^2 =& \mathbb{E}_{(x,y),(x',y')\sim\mathbb{D}\otimes\mathbb{D}}\Big[g(x,y)g(x',y')K^2(x,x')\Big] + \frac{1}{n^2}\sum_{i,j=1}^n g(x_i,y_i)g(x_j,y_j)K^2(x_i,x_j) \\
&- \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{(x,y)\sim\mathbb{D}}\Big[g(x_i,y_i)g(x,y)K^2(x,x_i)\Big].
\end{aligned}
$$

□