



Sparsity regret bounds for individual sequences in online linear regression

Sébastien Gerchinovitz

► To cite this version:

Sébastien Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. [Research Report] RR-7504, 2011. inria-00552267v1

HAL Id: inria-00552267

<https://inria.hal.science/inria-00552267v1>

Submitted on 5 Jan 2011 (v1), last revised 12 Apr 2013 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sparsity regret bounds for individual sequences in online linear regression

Sébastien Gerchinovitz

École Normale Supérieure*

Paris, France

sebastien.gerchinovitz@ens.fr

Abstract

We consider the problem of online linear regression on arbitrary deterministic sequences when the ambient dimension d can be much larger than the number of time rounds T . In this framework we prove deterministic online counterparts of the so-called sparsity oracle inequalities introduced in the stochastic setting in the past decade. They indicate that the task consisting in predicting almost as well as an unknown high-dimensional target vector is still statistically feasible if this target vector has only few non-zero coordinates. Our online-learning algorithm SeqSEW is based on exponential weighting and data-driven truncation. In a second part we apply a parameter-independent version of this algorithm to the regression model with random or fixed design. In this setting the sparsity regret bounds proved on arbitrary deterministic sequences yield sparsity oracle inequalities with leading constant 1 which are of the same flavor as in Dalalyan and Tsybakov (2008; 2010) but are adaptive (up to a logarithmic factor) to the unknown variance of the noise if the latter is Gaussian (weaker bounds are also proved under weaker assumptions). The framework of prediction of individual sequences thus offers a unifying setting to address tuning issues in both the random and the fixed design cases.

1 Introduction

Sparsity has been extensively studied in the stochastic setting over the past decade. Among the tools introduced for this purpose the notion of *sparsity oracle inequality* plays a fundamental role. In high-dimensional linear regression, such inequalities imply that the task consisting in predicting almost as well as an unknown target vector is still statistically feasible if the target vector has only few non-zero coordinates.

In this paper, we bring the notion of sparsity oracle inequality into the deterministic setting known as the *framework of prediction of individual sequences*. The corresponding deterministic inequalities are called *sparsity regret bounds*. We prove such bounds for an online-learning algorithm called SeqSEW which is inspired from the Sparse Exponential Weighting algorithm introduced in the stochastic setting by [DT07]. Thanks to individual sequences techniques (e.g., online truncation and online tuning), the most sophisticated version of our algorithm is fully automatic in the sense that no a priori knowledge is needed for the choice of the tuning parameters.

*This research was carried out within the INRIA project CLASSIC hosted by École Normale Supérieure and CNRS.

The second contribution of this paper deals with fruitful connections between the framework of individual sequences and the stochastic setting. More precisely, we show that the online truncation and the online parameter tuning performed by the algorithm SeqSEW for deterministic purposes yield, in the regression model with random or fixed design, sparsity oracle inequalities with leading constant 1 which are of the same flavour as in [DT08, DT10a] but are adaptive (up to a logarithmic factor) to the unknown variance σ^2 of the noise at least whenever the latter is Gaussian (weaker bounds are also proved under weaker assumptions). The framework of prediction of individual sequences thus offers a unifying framework to address tuning issues in both the random and the fixed design cases.

This paper is organized as follows. In Section 2 we describe our deterministic and stochastic settings. In Section 3 we detail the contributions of this paper in view of existing results on sparsity in the stochastic and the deterministic settings. In Section 4 we prove the aforementioned sparsity regret bounds for our algorithm SeqSEW, first when the forecaster has access to some a priori knowledge on the observations (Sections 4.1 and 4.2), and then when no a priori information is available (Section 4.3), which yields a fully automatic algorithm. In Section 5 we apply the algorithm SeqSEW to the regression model with random design (Section 5.1) and to the regression model with fixed design (Section 5.2). Some technical tools are finally given in appendix.

2 Setting and notations

As mentioned in the introduction, we consider three closely related prediction settings: a repeated game with deterministic data and two batch games with random data known in the statistical literature respectively as the *regression model with random design* and the *regression model with fixed design*.

Our prediction model in the deterministic setting is an equivalent variant of an extension of the game of prediction with expert advice called *prediction with side information*; see [CBL06, chapter 11] for references on this setting. A forecaster has to predict in a sequential fashion the values of an unknown sequence of real-valued observations $(y_t)_{t \in \mathbb{N}^*}$ given some input data $(x_t)_{t \in \mathbb{N}^*}$ from a certain set \mathcal{X} . At each time round $t \in \mathbb{N}^*$, the forecaster has access to $x_t \in \mathcal{X}$ and some base forecasts $\varphi_j(x_t) \in \mathbb{R}$, $1 \leq j \leq d$, on the basis of which he outputs a prediction $\hat{y}_t \in \mathbb{R}$ whose quality is assessed by the square loss $(y_t - \hat{y}_t)^2$. The goal of the forecaster is to minimize his cumulative loss

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2.$$

This cumulative loss will be compared to the cumulative losses $\sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2$ of the linear regressors $\mathbf{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^d u_j \varphi_j$, $\mathbf{u} \in \mathbb{R}^d$. Since the comparison of the latter cumulative losses will be carried out on *any* arbitrary deterministic sequence $(x_t, y_t)_{t \geq 1}$, this setting is a particular case of the so-called framework of prediction of *individual sequences*. In this setting no assumption is made on the way the data is generated (i.e., the sequence $(x_t, y_t)_{t \geq 1}$ is not assumed to be a realization of a given stochastic process with statistically interesting properties such as ergodicity and stationarity); we refer to [CBL06] for a detailed overview of this framework. We give in Figure 1 a detailed description of the repeated game of prediction with side information under the square loss.

As for the batch setting with random data, we first consider the so-called *regression model with random design*, a game-theoretic description of which is given in Figure 2. This setting is a

Online prediction with side information under the square loss

Parameters: input data set \mathcal{X} , base forecasters $\varphi = (\varphi_1, \dots, \varphi_d)$ with $\varphi_j : \mathcal{X} \rightarrow \mathbb{R}$, $1 \leq j \leq d$.

Initial step: the environment chooses a sequence of observations $(y_t)_{t \geq 1}$ in \mathbb{R} and a sequence of input data $(x_t)_{t \geq 1}$ in \mathcal{X} but the forecaster has not access to them.

At each time round $t \in \mathbb{N}^*$,

1. The environment reveals the input data $x_t \in \mathcal{X}$.
2. The forecaster chooses a prediction $\hat{y}_t \in \mathbb{R}$ (possibly as a linear combination of the $\varphi_j(x_t)$, but this is not necessary).
3. The environment reveals the observation $y_t \in \mathbb{R}$.
4. Each linear forecaster $\mathbf{u} \cdot \varphi \triangleq \sum_{j=1}^d u_j \varphi_j$, $\mathbf{u} \in \mathbb{R}^d$, incurs the loss $(y_t - \mathbf{u} \cdot \varphi(x_t))^2$ and the forecaster incurs the loss $(y_t - \hat{y}_t)^2$.

Figure 1: Description of the repeated game of online prediction with side information under the square loss.

batch counterpart of the previous online setting. The pairs $(x_t, y_t) \in \mathcal{X} \times \mathbb{R}$, $t \geq 1$, are drawn independently at random according to a common probability distribution; we thus use the capital letters X_t and Y_t instead of x_t and y_t . Moreover, contrary to the online deterministic setting, all the data from time 1 to time T is available to the forecaster at the beginning of the prediction game. The goal of the forecaster is not to minimize the cumulative loss $\sum_{t=1}^T (Y_t - \hat{y}_t)^2$ as in the deterministic setting (for it would be trivial since the Y_t are known to the forecaster), but to output a data-based regressor $\hat{f}_T : \mathcal{X} \rightarrow \mathbb{R}$ whose expected L^2 -risk

$$\mathbb{E} \left[\left\| f - \hat{f}_T \right\|_{L^2}^2 \right] \triangleq \mathbb{E} \left[(f(X) - \hat{f}_T(X))^2 \right]$$

is as small as possible. In the above equation, f is the unknown regression function, the expectation on the left-hand side is taken with respect to the sample $(X_t, Y_t)_{1 \leq t \leq T}$, and the expectation on the right-hand side is taken with respect to $(X_t, Y_t)_{1 \leq t \leq T}$ and X , where X is a random variable independent of $(X_t, Y_t)_{1 \leq t \leq T}$ which has the same distribution as X_1 . The expected L^2 -risk above will be compared to the L^2 -risks $\|f - \mathbf{u} \cdot \varphi\|_{L^2}^2$ of the linear regressors $\mathbf{u} \cdot \varphi \triangleq \sum_{j=1}^d u_j \varphi_j$, $\mathbf{u} \in \mathbb{R}^d$.

Finally, in the regression model with fixed design, the prediction game is still of the batch style (i.e., all the data is available to the forecaster at the beginning of the game). A game-theoretic description of this setting is given in Figure 3. Note that, similarly to the regression model with random design, the observations Y_t are random, but, in contrast with the latter setting, the input data x_t are deterministic. The goal of the forecaster is to make his expected prediction loss

$$\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T (f(x_t) - \hat{f}_T(x_t))^2 \right]$$

as small as possible, where the expectation is taken with respect to $(Y_t)_{1 \leq t \leq T}$. It will be compared to the prediction losses of the base linear regressors $\mathbf{u} \cdot \varphi \triangleq \sum_{j=1}^d u_j \varphi_j$, where $\mathbf{u} \in \mathbb{R}^d$, i.e., to

Aggregation in the regression model with random design under the square loss

Parameters: time horizon T , input data set \mathcal{X} (endowed with some σ -field \mathcal{B}), and base regressors $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_d)$ with $\varphi_j : \mathcal{X} \rightarrow \mathbb{R}$, $1 \leq j \leq d$.

One-shot game:

1. The environment chooses a probability distribution for $(X, Y) \in \mathcal{X} \times \mathbb{R}$ such that $\mathbb{E}[Y^2] < +\infty$ and $\sup_{1 \leq j \leq d} \mathbb{E}[\varphi_j^2(X)] < +\infty$. The corresponding regression function f and noise ε are given by

$$f(x) \triangleq \mathbb{E}[Y|X = x] \quad \text{and} \quad \varepsilon \triangleq Y - f(X).$$

(all these quantities are unknown to the forecaster)

2. The environment draws independent random copies $(X_1, Y_1), \dots, (X_T, Y_T)$ of (X, Y) . They are revealed to the forecaster.
3. The forecaster chooses a data-based regressor $\hat{f}_T : \mathcal{X} \rightarrow \mathbb{R}$ (possibly as a linear combination of the φ_j , but this is not necessary).
4. Each base linear regressor $\mathbf{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^d u_j \varphi_j$, where $\mathbf{u} \in \mathbb{R}^d$, incurs the L^2 -risk

$$\|f - \mathbf{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 \triangleq \mathbb{E}[(f(X) - \mathbf{u} \cdot \boldsymbol{\varphi}(X))^2]$$

and the forecaster incurs the L^2 -risk

$$\left\| f - \hat{f}_T \right\|_{L^2}^2 \triangleq \mathbb{E}[(f(X) - \hat{f}_T(X))^2 \mid (X_t, Y_t)_{1 \leq t \leq T}].$$

Figure 2: Aggregation in the regression model with random design under the square loss. This classical model is described here in a game-theoretic fashion.

the quantities

$$\frac{1}{T} \sum_{t=1}^T (f(x_t) - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2.$$

We end this section with some notations. The ambient dimension will be denoted by d contrary to several works on sparsity in the stochastic setting where it is denoted by p or by M (in this paper, the letter p will be used to denote some probability distributions on \mathbb{R}^d). Vectors in \mathbb{R}^d will be denoted by bold letters. For all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, the standard inner product in \mathbb{R}^d between $\mathbf{u} = (u_1, \dots, u_d)$ and $\mathbf{v} = (v_1, \dots, v_d)$ will be denoted by $\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^d u_i v_i$.

For all $\mathbf{u} \in \mathbb{R}^d$, the ℓ^0 -norm, ℓ^1 -norm, and ℓ^2 -norm of $\mathbf{u} = (u_1, \dots, u_d)$ are respectively defined by

$$\|\mathbf{u}\|_0 \triangleq \sum_{j=1}^d \mathbb{I}_{\{u_j \neq 0\}} = |\{j : u_j \neq 0\}|,$$

Aggregation in the regression model with fixed design under the square loss

Parameters: time horizon T , input data set \mathcal{X} , and base regressors $\varphi = (\varphi_1, \dots, \varphi_d)$ with $\varphi_j : \mathcal{X} \rightarrow \mathbb{R}$, $1 \leq j \leq d$.

One-shot game:

1. The environment chooses a function $f : \mathcal{X} \rightarrow \mathbb{R}$ and a probability distribution for $\varepsilon \in \mathbb{R}$ such that $\mathbb{E}[\varepsilon^2] < +\infty$ and $\mathbb{E}[\varepsilon] = 0$. The function f and the distribution of ε are not revealed to the forecaster.
2. The environment picks deterministic elements x_1, \dots, x_T in \mathcal{X} , draws independent random copies $\varepsilon_1, \dots, \varepsilon_T$ of ε , and forms the random variables

$$Y_t = f(x_t) + \varepsilon_t .$$

The sample $(x_1, Y_1), \dots, (x_T, Y_T)$ is revealed to the forecaster.

3. The forecaster chooses a data-based regressor $\hat{f}_T : \mathcal{X} \rightarrow \mathbb{R}$ (possibly as a linear combination of the φ_j , but this is not necessary).
4. Each base linear regressor $\mathbf{u} \cdot \varphi \triangleq \sum_{j=1}^d u_j \varphi_j$, where $\mathbf{u} \in \mathbb{R}^d$, incurs the prediction loss

$$\frac{1}{T} \sum_{t=1}^T (f(x_t) - \mathbf{u} \cdot \varphi(x_t))^2$$

and the forecaster incurs the prediction loss

$$\frac{1}{T} \sum_{t=1}^T (f(x_t) - \hat{f}_T(x_t))^2 .$$

Figure 3: Aggregation in the regression model with fixed design under the square loss. This classical model is described here in a game-theoretic fashion.

by

$$\|\mathbf{u}\|_1 \triangleq \sum_{j=1}^d |u_j| ,$$

and by

$$\|\mathbf{u}\|_2 \triangleq \left(\sum_{j=1}^d u_j^2 \right)^{1/2} .$$

The set of all probability distributions on a Borel subset Θ of \mathbb{R}^d (endowed with its Borel σ -algebra) will be denoted by $\mathcal{M}_1^+(\Theta)$. For all $\rho, \pi \in \mathcal{M}_1^+(\Theta)$, the Kullback-Leibler divergence between ρ and π is defined by

$$\mathcal{K}(\rho, \pi) \triangleq \begin{cases} \int_{\mathbb{R}^d} \ln \left(\frac{d\rho}{d\pi} \right) d\rho & \text{if } \rho \text{ is absolutely continuous with respect to } \pi; \\ +\infty & \text{otherwise,} \end{cases}$$

where $\frac{d\rho}{d\pi}$ denotes the Radon-Nikodym derivative of ρ with respect to π .

For all $x \in \mathbb{R}$ and $B > 0$, we denote by $\lceil x \rceil$ the smallest integer larger than or equal to x , and by $[x]_B$ its thresholded value:

$$[x]_B \triangleq \begin{cases} -B & \text{if } x < -B; \\ x & \text{if } -B \leq x \leq B; \\ B & \text{if } x > B. \end{cases}$$

Finally, we will use the (natural) convention $0 \ln(1 + U/0) = 0$ for all $U > 0$.

3 Overview of the literature and contributions of this paper

In this section we detail the contributions of this paper in view of existing results on sparsity both in the stochastic setting and in the framework of prediction of (deterministic) individual sequences. Our first contribution consists in the introduction of a deterministic counterpart of the so-called *sparsity oracle inequalities* in the framework of prediction of individual sequences. In Section 3.1 we motivate this new type of regret bounds – called *sparsity regret bounds* thereafter – and describe the main features of the algorithm SeqSEW, which satisfies such sparsity regret bounds. We then detail in Section 3.2 the second contribution of this paper, namely, how these sparsity regret bounds (of deterministic nature) can yield adaptive sparsity oracle inequalities in the stochastic setting.

3.1 Introduction of a deterministic counterpart of sparsity oracle inequalities

3.1.1 Sparsity oracle inequalities in the stochastic setting

Consider first the (generalized) linear regression model with fixed or random design, i.e., to be consistent with the notations introduced in the previous section,

$$(1) \quad Y_t = \mathbf{u}^* \cdot \boldsymbol{\varphi}(X_t) + \varepsilon_t, \quad 1 \leq t \leq T,$$

where the $X_t \in \mathcal{X}$ are either i.i.d random variables (random design) or fixed elements (fixed design), denoted in both cases by capital letters in this section, where $\mathbf{u}^* \in \mathbb{R}^d$ is the unknown linear combination (recall that $\mathbf{u}^* \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^d u_j^* \varphi_j$), and where the ε_t are i.i.d. square-integrable real random variables with zero mean (conditionally on the X_t if the design is random).

Three main statistical problems arise in this linear regression framework:

- prediction: estimating $(\mathbf{u}^* \cdot \boldsymbol{\varphi}(X_t))_{1 \leq t \leq T}$ (fixed design) or $\mathbf{u}^* \cdot \boldsymbol{\varphi}$ (random design);
- estimation: estimating \mathbf{u}^* ;
- support estimation: estimating the set of the non-zero coordinates of \mathbf{u}^* .

These three tasks have been extensively studied over the past decade in the high-dimensional setting under a sparsity scenario, namely, when

$$(2) \quad \|\mathbf{u}^*\|_0 \ll T \ll d.$$

In this paper, we will only address the prediction problem (under a sparsity scenario). As we will see later, this problem can be addressed in rather general regression models both in the stochastic

setting (see the next paragraphs) and in the deterministic setting known as the framework of prediction of individual sequences (see Section 3.1.2).

In the stochastic setting, most risk bounds for the prediction problem take the form of sparsity oracle inequalities. We explain below the basic idea that underlies this type of risks bounds. In the (generalized) linear regression model with fixed or random design (1), the ordinary least squares estimator

$$\hat{\mathbf{u}}_T \in \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^d} \frac{1}{T} \sum_{t=1}^T (Y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(X_t))^2$$

has an expected risk $\mathbb{E}[R(\hat{\mathbf{u}}_T)]$ which is at most of order d/T (see [GKKW02] and the references therein for the fixed design, and the more recent advances in [AC09] for the random design under weak assumptions on the output distribution). Here, we defined the risk $R(\mathbf{u})$ of any linear combination $\mathbf{u} \in \mathbb{R}^d$ by

$$R(\mathbf{u}) \triangleq \begin{cases} \|\mathbf{u}^* \cdot \boldsymbol{\varphi} - \mathbf{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 & \text{(random design)} \\ \frac{1}{T} \sum_{t=1}^T (\mathbf{u}^* \cdot \boldsymbol{\varphi}(X_t) - \mathbf{u} \cdot \boldsymbol{\varphi}(X_t))^2 & \text{(fixed design).} \end{cases}$$

When the ambient dimension d is much larger than the sample size T , a direct minimization of the least-squares criterion on \mathbb{R}^d can lead to overfitting, which is reflected in the non-vanishing upper bound d/T . However, one can still hope to achieve a small risk under the additional assumption $\|\mathbf{u}^*\|_0 = s \ll T$. Indeed, if the support $J(\mathbf{u}^*)$ of \mathbf{u}^* was known in advance, the oracle applying the ordinary least squares estimator to the linear subspace $\{\mathbf{u} \in \mathbb{R}^d : \forall j \notin J(\mathbf{u}^*), u_j = 0\}$ would have a risk of order at most $s/T \ll 1$.

The rate s/T of the latter oracle can actually be achieved up to a $\ln d$ factor without the prior knowledge of the set $J(\mathbf{u}^*)$ nor even of its cardinality $s = |J(\mathbf{u}^*)|$. This was first done by adding a ℓ^0 -complexity penalty to the least-squares criterion (penalty proportional to the number of non-zero coefficients; see [Aka71] for the AIC criterion, [Mal73] for Mallows' C_p , and [Sch78, FG94] for the BIC criterion). As proved in [BTW04, BTW07a] through model selection arguments, without any assumption on the dictionary, such ℓ^0 -regularized methods “mimic” the ℓ^0 -oracle in the sense that their risks are at most of order

$$\frac{s \ln(d/s)}{T}$$

without any prior knowledge on $J(\mathbf{u}^*)$. Since this rate is actually minimax optimal on ℓ^0 -balls for fixed or Gaussian random designs [RWY09, Ver10], such ℓ^0 -regularized methods are termed *adaptive* to the unknown sparsity of \mathbf{u}^* .

The risk bounds proved in [BTW04, BTW07a] are actually stronger. They indeed hold in the more general regression model (sometimes called the *nonparametric regression setup*)

$$(3) \quad Y_t = f(X_t) + \varepsilon_t, \quad 1 \leq t \leq T,$$

with random design (where the X_t are i.i.d. random variables, as in Figure 2) or fixed design (where the X_t are deterministic elements, as in Figure 3). Since f is not necessarily assumed to be of the form $f = \mathbf{u}^* \cdot \boldsymbol{\varphi}$, the risk bounds mentioned earlier

$$\mathbb{E}[R(\hat{\mathbf{u}}_T)] = \mathcal{O}\left(\frac{s \ln(d/s)}{T}\right)$$

are replaced with bounds on the differences $\mathbb{E}[R(\hat{\mathbf{u}}_T)] - R(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{R}^d$, where the risk $R(\mathbf{u})$ is now defined by

$$R(\mathbf{u}) \triangleq \begin{cases} \|f - \mathbf{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 & \text{(random design)} \\ \frac{1}{T} \sum_{t=1}^T (f(X_t) - \mathbf{u} \cdot \boldsymbol{\varphi}(X_t))^2 & \text{(fixed design).} \end{cases}$$

The risk bounds proved in [BTW04, BTW07a] are indeed of the form

$$(4) \quad \mathbb{E}[R(\hat{\mathbf{u}}_T)] \leq (1+a) \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ R(\mathbf{u}) + C(a) \frac{\|\mathbf{u}\|_0}{T} \ln \left(\frac{e d}{\max\{\|\mathbf{u}\|_0, 1\}} \right) \right\},$$

where $C(a) \sim a^{-1} \rightarrow +\infty$ as $a \rightarrow 0$. The above upper bound is a typical example of what is called a *sparsity oracle inequality*, i.e., in the prediction problem, a risk bound in terms of the risk $R(\mathbf{u})$ and the number of non-zero coordinates $\|\mathbf{u}\|_0$ of all $\mathbf{u} \in \mathbb{R}^d$.

Numerous other works have proved sparsity oracle inequalities over the last decade. For the case of the prediction problem we are interested in, we can refer, among others, to [BTW07b, vdG08, BRT09, Kol09b] for sparsity oracle inequalities with leading constant strictly larger than 1 (as in (4)) which hold for Lasso-type procedures (which we comment on later). Recently, *sharp* sparsity oracle inequalities, i.e., sparsity oracle inequalities with leading constant equal to 1, have been proved for procedures based on exponential weighting; see, e.g., [DT08, DT10a, RT10, AL10] and the references therein.

As detailed in [BTW06, BTW07a, DT08], sparsity oracle inequalities have interesting consequences. They indeed imply that:

1. in the high-dimensional linear regression model (1), prediction is still statistically feasible under a sparsity scenario (this is the main motivation we chose to introduce the notion of sparsity oracle inequality);
2. statistical procedures satisfying such sparsity oracle inequalities can be used to perform adaptive nonparametric regression (i.e., for an appropriately well chosen basis, these procedures are adaptive to the unknown smoothness of the regression function f);
3. statistical procedures satisfying sharp sparsity oracle inequalities achieve (quasi-)optimal rates of model-selection, convex, and linear aggregation in the sense of [Nem00, Tsy03]. Namely, up to some small remainder terms, these procedures predict at least as well as the best among the base predictors φ_j (model-selection aggregation), the best convex combination of the φ_j (convex aggregation), and the best linear combination of the φ_j (linear aggregation); the corresponding remainder terms are the smallest possible ones. We refer the reader to, e.g., [RT10, Section 6] and the references therein (note that there are also other types of aggregation than the three ones mentioned above).

We end this section with a brief computational efficiency-oriented overview of alternatives to ℓ^0 -regularization. A major drawback of ℓ^0 -regularization is that the corresponding non-convex minimization problems are not computationally tractable. This complexity issue has been handled by replacing the ℓ^0 -penalty with a ℓ^1 -penalty (proportional to the (weighted) sum of the absolute values of the coefficients). The ℓ^1 -regularization can indeed be seen as a ‘convex relaxation’ of ℓ^0 -regularization, i.e., from a geometrical viewpoint, it behaves similarly to the ℓ^0 -penalty but leads to

convex and thus computationally tractable minimization problems. The ℓ^1 -regularization was first proposed by [Tib96] for the so-called Lasso estimator and by [DJ94] for a soft thresholding-based estimator in the context of wavelet regression. The ℓ^0 -oracle properties of the Lasso, i.e., risk bounds typically of the form $(s \ln d)/T$, together with those of several variants, have then been extensively studied over the past decade (see, e.g., [BTW07b, BRT09, vdG08, Kol09a, Kol09b] and references therein). We also mention that [MM10] recently addressed the ℓ^1 -oracle properties of the Lasso estimator through a model selection viewpoint.

Despite their computational efficiency, these ℓ^1 -regularized methods suffer from another drawback: their ℓ^0 -oracle properties hold under rather restrictive assumptions on the (fixed or random) design; namely, that the covariates should be nearly orthogonal (see [vdGB09] for a detailed discussion on these assumptions). Recently an attempt has thus been made to reach a compromise between strong theoretical guarantees (that hold under very weak assumptions on the design) and computational efficiency. In this respect [DT08, DT09, DT10a] proposed an aggregation algorithm which is based on exponential weighting, which satisfies sparsity oracle inequalities on a fixed or random design under almost no assumption on the dictionary, and which can be approximated numerically at a reasonable computational cost for large values of the ambient dimension d . This is the algorithm from which our algorithm SeqSEW is inspired (see Section 3.1.3 for more details). More recently [RT10, AL10] designed algorithms based on aggregation which satisfy optimal rates of sparse aggregation in the regression model with fixed design (in the sense of [RT10]). [AL10] also addressed the regression model with random design but the corresponding risk bounds depend as in [DT08] on the logarithms of $\|\mathbf{u}^*\|_1$ and T (but their bound holds with large probability). In both papers [RT10, AL10] the corresponding algorithms were shown to be well approximated by MCMC methods with conclusive experimental results.

3.1.2 Sparsity in the framework of individual sequences

In the deterministic setting described in Figure 1, we prove regret bounds on individual sequences $(x_t, y_t)_{1 \leq t \leq T} \in (\mathcal{X} \times \mathbb{R})^T$ of the form

$$(5) \quad \sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + c_1 B_y^2 \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\boldsymbol{\varphi}\|_\infty \|\mathbf{u}\|_1}{c_2 B_y \|\mathbf{u}\|_0} \right) \right\} + c_3 B_y^2,$$

where $\boldsymbol{\varphi}(x) \triangleq (\varphi_1(x), \dots, \varphi_d(x))$, where $\|\boldsymbol{\varphi}\|_\infty \triangleq \max_{1 \leq j \leq d} \sup_{x \in \mathcal{X}} |\varphi_j(x)|$, where c_1, c_2, c_3 are positive absolute constants, where B_y is a bound on all the observations $|y_t|$, $1 \leq t \leq T$ (whose knowledge is not required by the most sophisticated version of our algorithm), and where $\|\mathbf{u}\|_0$ denotes the *sparsity* of the vector $\mathbf{u} \in \mathbb{R}^d$, i.e., the number of its non-zero coordinates:

$$\|\mathbf{u}\|_0 \triangleq \sum_{j=1}^d \mathbb{I}_{\{u_j \neq 0\}} = |\{j : u_j \neq 0\}|.$$

We thus provide an online and deterministic counterpart of the so-called *sharp sparsity oracle inequalities* introduced in the stochastic setting, which we call *sparsity regret bounds* (see Equation (7) below for a general formulation). Our work thus draws the following parallel. In 2001, Azoury and Warmuth [AW01] and Vovk [Vov01] proved the first (deterministic) regret bounds on a variant of the Ridge regression forecaster, which had already been theoretically studied in the stochastic setting since the 1970s (see, e.g., [HK70, CML07, AC09]). In the same way, we extend to the deterministic setting the work on Sparse Exponential Weighting by Dalalyan and Tsybakov

[DT07, DT08, DT10a] who addressed the (stochastic) regression model with fixed or random design under a sparsity scenario; see Section 3.1.3 below for more details on our algorithm.

Note that our work draws another parallel. In the same way as sparsity-oriented methods in the stochastic literature are adaptive to the unknown sparsity, our algorithm provides adaptivity to the unknown sparsity in the deterministic setting. The upper bound (5) can indeed be rewritten in the following way (since $(x, y) \mapsto x \ln(1 + y/x)$ is nondecreasing in $x \in \mathbb{R}^+$ and in $y \in \mathbb{R}^+$ as explained in Corollary 4).

For all $s \in \mathbb{N}$ and all $U > 0$,

(6)

$$\sup_{\substack{x_1, \dots, x_T \\ y_1, \dots, y_T}} \left\{ \sum_{t=1}^T (y_t - \hat{y}_t)^2 - \inf_{\substack{\|\mathbf{u}\|_0 \leq s \\ \|\mathbf{u}\|_1 \leq U}} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \right\} \leq c_1 B_y^2 s \ln \left(1 + \frac{\sqrt{dT} \|\boldsymbol{\varphi}\|_\infty U}{c_2 B_y s} \right) + c_3 B_y^2,$$

where the supremum is taken over all the input data $x_1, \dots, x_T \in \mathcal{X}$ and all the observations $y_1, \dots, y_T \in [-B_y, B_y]$.

The upper bound (6) is of order $B_y^2 s \ln(\sqrt{dT} \|\boldsymbol{\varphi}\|_\infty U / (s B_y))$, which can be compared to the bound $B_y^2 s \ln(T \|\boldsymbol{\varphi}\|_\infty^2 \|\mathbf{u}^*\|_2^2 / (s B_y^2))$ that a variant of the sequential Ridge regression forecaster would satisfy when applied to the unknown support $\{j : u_j^* \neq 0\}$ of a “good” linear combination \mathbf{u}^* of sparsity $\|\mathbf{u}^*\|_0 = s$. More precisely, this sequential forecaster, which was studied in [AW01, Vov01], predicts at time t as a linear combination of the base forecasts, i.e., as $\hat{y}_t = \hat{\mathbf{u}}_t \cdot \boldsymbol{\varphi}(x_t)$, where

$$\hat{\mathbf{u}}_t \in \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + \lambda \|\mathbf{u}\|_2^2 \right\},$$

for some tuning parameter $\lambda > 0$. Using the same notations as above, it is proved in [AW01, Vov01] that this forecaster satisfies the regret bound

$$\begin{aligned} \sum_{t=1}^T (y_t - \hat{y}_t)^2 &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + \lambda \|\mathbf{u}\|_2^2 \right\} + dB_y^2 \ln \left(1 + \frac{T \|\boldsymbol{\varphi}\|_\infty^2}{\lambda} \right) \\ &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + dB_y^2 \ln \left(1 + \frac{T \|\mathbf{u}\|_2^2 \|\boldsymbol{\varphi}\|_\infty^2}{dB_y^2} \right) \right\} + dB_y^2. \end{aligned}$$

The last inequality follows from the quasi-optimal¹ choice of $\lambda = (dB_y^2) / \|\mathbf{u}^*\|_2^2$, where $\mathbf{u}^* \in \mathbb{R}^d$ minimizes the right-hand side of the last inequality. We can thus deduce the aforementioned bound $B_y^2 s \ln(T \|\boldsymbol{\varphi}\|_\infty^2 \|\mathbf{u}^*\|_2^2 / (s B_y^2))$ with $s = \|\mathbf{u}^*\|_0$ when the latter forecaster is only applied to the support of \mathbf{u}^* . However, just like in the stochastic setting, this support is unknown (and so is $\|\mathbf{u}^*\|_2^2$), so that the latter forecaster can only be used as an ideal benchmark. The discussion above shows that our regret bound (6), which is derived for a legitimate algorithm, is comparable to this benchmark bound.

¹This choice of λ minimizes over \mathbb{R}_+ the function $\lambda \mapsto \lambda \|\mathbf{u}^*\|_2^2 + dB_y^2 \ln(T \|\boldsymbol{\varphi}\|_\infty^2 / \lambda)$.

As far as optimality is concerned, the upper bound (6) is not improvable up to logarithmic factors. Indeed, one can straightforwardly derive from [Vov01, Section 3.3] (see also [SKF02]) that if $s \ll T$, then the regret of any forecaster on the subset $\{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\|_0 \leq s \text{ and } \|\mathbf{u}\|_1 \leq 1\}$ is bounded from below by a quantity of order $B_y^2 s \ln T$. Our algorithm is thus minimax optimal on intersections of ℓ^0 - and ℓ^1 -balls (up to logarithmic factors).

The main interesting consequence of our sparsity regret bound (6) is that, just like in the stochastic setting, the prediction task in a high-dimensional deterministic setting is still statistically feasible under a sparsity scenario. Indeed, our new sparsity regret bound is roughly of order $s \ln(dT)$, which is sublinear in T if $s \ll T / \ln(dT)$. Therefore, it can yield vanishing per-round regret bounds even when $d \gg T$ provided that there is a sparse $\mathbf{u} \in \mathbb{R}^d$ (i.e., such that $\|\mathbf{u}\|_0 \ll T / \ln(dT)$) whose cumulative loss after T time rounds is small enough.

To the best of our knowledge, Proposition 2 and its refinements (Proposition 3 and Theorem 1) provide the first examples of sparsity regret bounds in the framework of prediction of individual sequences, i.e., bounds of the form

$$(7) \quad \sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + \Delta(\|\mathbf{u}\|_0, \|\mathbf{u}\|_1, T, d, B_y, \|\boldsymbol{\varphi}\|_\infty) \right\},$$

where the regret term $\Delta(\|\mathbf{u}\|_0, \|\mathbf{u}\|_1, T, d, B_y, \|\boldsymbol{\varphi}\|_\infty)$ grows at most linearly in $\|\mathbf{u}\|_0$ and B_y^2 and logarithmically in $\|\mathbf{u}\|_1$, T , and d . We stress that Δ should not depend more than logarithmically on d and $\|\mathbf{u}\|_1$ in view of the lower bound $B_y^2 s \ln T$ mentioned in the previous paragraph.

Recent works in the field of online convex optimization addressed the sparsity issue in the online deterministic setting, but from a quite different angle. See, e.g., [LLZ09, SST09, Xia10, DSSST10] and the references therein. All these articles focus on convex regularization. In the particular case of ℓ^1 -regularization under the square loss, the aforementioned works prove bounds on the ℓ^1 -regularized regret of the form

$$\begin{aligned} & \sum_{t=1}^T \left((y_t - \hat{\mathbf{u}}_t \cdot \mathbf{x}_t)^2 + \lambda \|\hat{\mathbf{u}}_t\|_1 \right) \\ & \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T \left((y_t - \mathbf{u} \cdot \mathbf{x}_t)^2 + \lambda \|\mathbf{u}\|_1 \right) + \tilde{\Delta}(\|\mathbf{u}\|_1, T, d, B_y, \|\boldsymbol{\varphi}\|_\infty) \right\}, \end{aligned}$$

where we used the same notations as above (and denoted by $\hat{\mathbf{u}}_t$ the linear combination output by the algorithm at time t), and where the function $\tilde{\Delta}$ grows at least linearly in $\|\mathbf{u}\|_1$. This is in contrast with the slow logarithmic dependence we are looking for and prove, e.g., in Proposition 2.

However, in the works mentioned above [LLZ09, SST09, Xia10, DSSST10] and in related papers cited therein, a particular attention is paid to sparsity-inducing algorithms, i.e., algorithms that predict as a sparse linear combination of the base forecasts, namely, that output

$$\hat{y}_t = \hat{\mathbf{u}}_t \cdot \boldsymbol{\varphi}(x_t) \quad \text{with} \quad \|\hat{\mathbf{u}}_t\|_0 \ll T / \ln(dT),$$

which we do not do in this paper. Our predictions \hat{y}_t are not even linear combinations of the base forecasts since they are given by averages of truncated linear combinations of the $\varphi_j(x_t)$, $1 \leq j \leq d$. An open question is thus whether it is possible to conciliate both goals, i.e., to design an online learning algorithm which both outputs sparse linear combinations (which is statistically useful and sometimes essential for computational issues) and satisfies a sparsity regret bound of the form (7).

3.1.3 Sequential Sparse Exponential Weighting

Our prediction algorithm, called *SeqSEW* thereafter (for *Sequential Sparse Exponential Weighting*), is an online variant of the Sparse Exponential Weighting algorithm introduced in the stochastic setting by [DT07]. The key contribution of these papers was to establish a link between aggregation with exponential weights and sparsity. In this way they provide a statistical procedure with ℓ^0 -oracle guarantees under almost no assumption on the dictionary $(\varphi_j)_{1 \leq j \leq d}$ which is at the same time computationally efficient for large values of the dimension d . The results developed in Section 4 show that this connection between exponential weighting and sparsity still holds in the framework of prediction of individual sequences.

Aggregation (or mixing) with exponential weights has now quite a long history in the theory of prediction of individual sequences, from the seminal works [Vov90, LW94] to more recent parameter-tuning oriented papers such as [CBMS07]; see [CBL06] for a detailed overview of the subject. The idea of mixing estimators was initiated earlier in the stochastic setting and has been concurrently developed since; see, e.g., the references given in [Yan01, LB06]. For the particular case of mixing with exponential weights in the stochastic setting, we can refer, among other works, to [Cat99, Cat04] for density estimation, to [Cat99, Cat04, Aud04, BN08, DT10a, AL10] for regression on a random design, and to [LB06, Gir08, DT08, AL10, RT10] for regression on a fixed design (or in the related Gaussian sequence model).

In addition to using exponential weights, our algorithm is based on a trick which proves to be crucial for deriving some theoretical guarantees: truncation of the base forecasts before the exponential averaging step. More precisely, the probability distribution $p_t \in \mathcal{M}_1^+(\mathbb{R}^d)$ output by our algorithm at the end of time round $t - 1$ is computed on the basis of the truncated base forecasts instead of the base forecasts themselves, i.e.,

$$p_t(\mathbf{d}\mathbf{u}) \triangleq \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(y_s - [\mathbf{u} \cdot \varphi(x_s)]_B\right)^2\right)}{W_t} \pi_\tau(\mathbf{d}\mathbf{u}),$$

where W_t is a normalizing constant, where π_τ is a prior distribution on \mathbb{R}^d to be specified later, and where $B > 0$ and $\eta > 0$ are tuning parameters. Accordingly, the prediction \hat{y}_t output by our algorithm at time t is a convex mixture (with respect to p_t) of the truncated base forecasts, i.e.,

$$\hat{y}_t \triangleq \int_{\mathbb{R}^d} [\mathbf{u} \cdot \varphi(x_t)]_B p_t(\mathbf{d}\mathbf{u}).$$

Truncation was already used many times in the literature; for the case of least squares regression, see, e.g., [GKKW02]. Truncation also turned out to be useful in the sequential prediction of unbounded time series under the square loss [GO07, BBGO10]. As can be seen from the proof of Lemma 1, truncation to a level larger than a bound B_y on the observations $|y_t|$ can only improve prediction, while it enables to take advantage of the exponential concavity of the square loss. When no such bound B_y is available to the forecaster, or, in the stochastic setting, when the finite-time stochastic process $(Y_t)_{1 \leq t \leq T}$ is unbounded, one can automatically and almost surely adapt to the unknown bound $\max_{1 \leq t \leq T} |y_t|$ (resp., the unknown random bound $\max_{1 \leq t \leq T} |Y_t|$); see Section 3.2 below.

This paper only focuses on the theoretical guarantees of the algorithm SeqSEW. However, it is possible to slightly adapt its statement to make it computationally tractable by means of Langevin Monte-Carlo approximation while not affecting its statistical properties. To do so, one can follow the same lines as in [DT09, DT10b]. The technical details are however omitted in this paper.

3.2 From individual sequences to stochastic sequences

A major criticism that has sometimes been addressed to the theory of prediction of individual sequences is that the losses are assumed to be uniformly bounded, while this is not the case in the stochastic setting (e.g., observations Y_t deviating from $f(X_t)$ by a Gaussian noise can take values in the whole real line).

This remark could lead to think that the framework of individual sequences on the one hand (which does not make any stochastic assumption on the way the observed data are generated but only deals with bounded losses) and the stochastic setting on the other hand (in which the observed sequence is assumed to be ergodic and stationary or even i.i.d., but for which the observations do not need to be bounded) are in some sense independent frameworks. This is not true as was already illustrated in the past: bounds in the stochastic setting can be derived from bounds on individual sequences, and we make further progress in this direction. In particular, our (deterministic) sparsity regret bounds yield adaptive sparsity oracle inequalities both in the random- and the fixed-design regression models (with unbounded noise).

This section is organized as follows. In Section 3.2.1 we explain how online-learning algorithms designed for individual sequences with bounded observations can be applied to the regression model with random design if the observations Y_1, \dots, Y_T are unbounded. In Section 3.2.2 we show that the algorithm SeqSEW answers two open questions raised in [DT10a] on a random design and also provides adaptive counterparts of results in [DT08] on a fixed design.

3.2.1 Exploiting the fact that a finite stochastic sequence is almost surely bounded

In this section we only deal with the regression model with random design. The regression model with fixed design can however be handled similarly; see Section 5.2.

The basic idea underlying the use of online-learning algorithms in this stochastic setting consists in treating each realization of the stochastic sequence $(X_1, Y_1), \dots, (X_T, Y_T)$ as an individual sequence. As a first consequence, even if the whole sample $(X_1, Y_1), \dots, (X_T, Y_T)$ is available at the beginning of the prediction game, we process it in a sequential fashion.

As a second sequence, the regret guarantees of the underlying online-learning algorithm hold *almost surely*. They can usually take the following form: for some comparison subset \mathcal{U} of \mathbb{R}^d , with probability one,

$$(8) \quad \sum_{t=1}^T (Y_t - \tilde{f}_t(X_t))^2 \leq \inf_{\mathbf{u} \in \mathcal{U}} \left\{ \sum_{t=1}^T (Y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + \psi_{T,d,B_y}(\mathbf{u}) \right\},$$

for some small remainder term $\psi_{T,d,B_y}(\mathbf{u})$ such as the one given in the sparsity regret bound (7). The comparison set $\mathcal{U} \subset \mathbb{R}^d$ can be taken as, e.g., the set of the vertices of the simplex in \mathbb{R}^d (which corresponds to the problem of *model-selection* aggregation), the whole simplex (which corresponds to *convex* aggregation), or the whole \mathbb{R}^d space (which corresponds to *linear* aggregation). In the above equation $\tilde{f}_t(x)$ is the prediction output by the algorithm at time t when given the input data $x \in \mathcal{X}$, and B_y is a bound on the observations $|Y_1|, \dots, |Y_T|$.

The derivation of a risk bound from the above almost-sure regret bound (8) is now quite standard and follows from a mirror-averaging step; see, e.g., [CBCG04, CBG08, JRT08] for a reference.

More precisely, as is detailed in Section 5.1, one can show from Jensen's inequality and elementary calculations that the data-based regressor $\hat{f}_T \triangleq \frac{1}{T} \sum_{t=1}^T \hat{f}_t$ satisfies the risk bound

$$\mathbb{E} \left[(f(X) - \hat{f}_T(X))^2 \right] \leq \inf_{\mathbf{u} \in \mathcal{U}} \left\{ \mathbb{E} \left[(f(X) - \mathbf{u} \cdot \boldsymbol{\varphi}(X))^2 \right] + \mathbb{E} [\psi_{T,d,B_y}(\mathbf{u})] \right\}.$$

If the observations $|Y_t|$ are almost surely bounded by a known deterministic bound B_y , then there are many online-learning algorithms which satisfy the regret bound (8). They usually depend on some tuning parameters which can be chosen as a function of B_y ; see, e.g., [CBL06] for a reference on the exponentially weighted average forecaster or other potential based forecasters.

However, a proper attention should be paid to the case when the observations Y_t are unbounded, e.g., when Y_1, \dots, Y_T are i.i.d. random variables such that

$$\forall B > 0, \quad \mathbb{P}(|Y_1| > B) > 0.$$

In this case, [BN08] propose truncating the observations Y_t to some threshold $\gamma = b$ or $\gamma = b \ln T$ (up to constant factors) for some known bound b on the infinity norms of the regression function f and the base regressors φ_j , $1 \leq j \leq d$. They then apply an exponentially weighted average forecaster to the truncated observations

$$\tilde{Y}_t = [Y_t]_\gamma, \quad 1 \leq t \leq T.$$

The authors then prove a risk bound for \hat{f}_T in terms of the truncated observation $[Y]_\gamma$, from which they derive by an approximation argument a risk bound in terms of the non-truncated (and possibly unbounded) observation Y .

A drawback of the previous approach is not only that b is assumed to be known in advance, but more importantly that f is assumed to be uniformly bounded. On the contrary one could hope to deal with the more general case where $\mathbb{E}[f^2(X)] < +\infty$. Besides, the base forecasters $\mathbf{u} \cdot \boldsymbol{\varphi}$, $\mathbf{u} \in \mathcal{U}$, are not uniformly bounded if $\boldsymbol{\varphi} \neq \mathbf{0}$ and $\mathcal{U} = \mathbb{R}^d$ (so that the approach followed in [BN08] for model-selection or convex aggregation does not readily apply to linear aggregation).

We thus suggest to truncate the base forecasts $\mathbf{u} \cdot \boldsymbol{\varphi}(X_t)$ instead of truncating the observations Y_t . Ideally we would like to truncate these base forecasts to the unknown random bound $B_y = \max_{1 \leq t \leq T} |Y_t|$, since this can only improve prediction and proves to be useful for the analysis. The actual truncation is thus performed with respect to a time-varying threshold which adapts to B_y (the latter random bound is almost surely finite since T is finite and the observations Y_t are integrable by assumption). To do so, we use parameter-tuning techniques provided, e.g., in [ACBG02, CBMS07]. Adaptation means here that we are able to prove bounds within constant factors of (quasi-optimal) bounds that could be proved if B_y was known in advance by the forecaster. In particular, this property enables us to derive sparsity oracle inequalities in the stochastic setting which are adaptive to the unknown variance of the noise at least whenever the latter is Gaussian; see the next section for more details.

3.2.2 Adaptivity of SeqSEW in the stochastic setting under a sparsity scenario

Our results in the stochastic setting answer two questions left open in [DT10a] for the regression model with random design and provide adaptive counterparts of results in [DT08] for the regression model with fixed design.

Following the mirror-averaging approach described in the previous section, the sparsity regret bounds derived in Section 4 on individual sequences lead to sparsity oracle inequalities in the regression model with random design (see Theorem 2 of Section 5.1). In the present section we only consider the particular case when the regression function f is bounded and when the i.i.d. errors $\varepsilon_t \triangleq Y_t - f(X_t)$ are subgaussian (conditionally on the X_t) in the sense that, for some constant $\sigma^2 > 0$,

$$\|f\|_\infty < +\infty \quad \text{and} \quad \mathbb{E}\left[e^{\lambda \varepsilon_1} \mid X_1\right] \leq e^{\lambda^2 \sigma^2 / 2} \quad \text{a.s.,} \quad \forall \lambda \in \mathbb{R}.$$

Under the above assumptions, we prove the following sparsity oracle inequality in Corollary 5 of Section 5.1:

(9)

$$\begin{aligned} \mathbb{E}\left[\|f - \hat{f}_T\|_{L^2}^2\right] &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \|f - \mathbf{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 + 128 \left(\frac{\mathbb{E}[Y]^2}{T} + \psi_T \right) \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \frac{1}{dT} \sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + 64 \left(\frac{\mathbb{E}[Y]^2}{T} + \psi_T \right), \end{aligned}$$

where $\psi_T \leq 8 \|f\|_\infty^2 / T + 2\sigma^2 / T + 32 \sigma^2 (\ln T) / T$ as shown in Remark 8 of Section 5.1. Besides, as explained in Remark 7, the two terms $\mathbb{E}[Y]^2 / T$ above can be avoided with a more sophisticated variant of the algorithm SeqSEW, so that we omit them in the following discussion.

As detailed in Remark 5 of Section 5.1, the above bound (59) is comparable to the sparsity oracle inequality stated in Proposition 1 of [DT10a], which we recall below with our notations (e.g., β therein corresponds to $1/\eta$ in this paper). Our bound might even be better if R is large, or if the $\|\varphi_j\|_{L^2}^2$ are much smaller than the $\|\varphi_j\|_\infty^2$, which is the case, e.g., when the φ_j are square integrable but unbounded.

Proposition 1 (Proposition 1 of [DT10a]). *Assume that for some constant $L_\varphi > 0$ we have $\max_{1 \leq j \leq d} \|\varphi_j\|_\infty \leq L_\varphi$. If, in addition, the errors ε_t have a bounded exponential moment:*

$$\exists b, \sigma^2 > 0 \quad \text{such that} \quad \mathbb{E}\left[e^{\lambda \varepsilon} \mid X\right] \leq e^{\sigma^2 \lambda^2 / 2}, \quad \forall |\lambda| \leq b, \quad \text{a.s.,}$$

then, for every $R > 0$ and $\eta \leq \min \left\{ \left(2\sigma^2 + 2 \sup_{\|\mathbf{u}\|_1 \leq R} \|\mathbf{u} \cdot \boldsymbol{\varphi} - f\|_\infty^2 \right)^{-1}, b/(4RL_\varphi) \right\}$, the mirror averaging aggregate $\hat{f}_T : \mathcal{X} \rightarrow \mathbb{R}$ defined in [DT10a, Equations (1) and (3)] satisfies

$$\mathbb{E}\left[\|f - \hat{f}_T\|_{L^2}^2\right] \leq \inf_{\mathbf{u}} \left\{ \|f - \mathbf{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 + \frac{4}{\eta(T+1)} \sum_{j=1}^d \ln(1 + |u_j|/\tau) \right\} + \Delta_T(\eta, \tau),$$

where the infimum is taken over all $\mathbf{u} \in \mathbb{R}^d$ such that $\|\mathbf{u}\|_1 \leq R - 2d\tau$ and where

$$\Delta_T(\eta, \tau) \triangleq 4\tau^2 \sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + \eta^{-1}(T+1)^{-1}.$$

We can now discuss the two questions left open in [DT10a]; further details are given in Remark 5 of Section 5.1. Despite the similarity of the two bounds, the sparsity oracle inequality

stated in Proposition 1 above only holds for vectors \mathbf{u} within ℓ^1 -balls of finite radii. The authors thus asked in [DT10a, Section 4.2] whether it was possible to extend the infimum to the whole \mathbb{R}^d space. Our results show that, thanks to truncation, the answer is positive.

The second open question, which was raised in [DT10a, Remark 7], deals with the prior knowledge of the variance factor σ^2 of the noise. The latter is indeed required by their algorithm for the choice of the inverse temperature parameter η . The authors thus asked whether adaptivity to σ^2 was possible. Our sparsity oracle inequality provides a positive answer (up to logarithmic factors) at least whenever the noise is Gaussian. Weaker bounds are also proved under weaker assumptions; see Corollary 5 of Section 5.1.

Since these adaptivity properties still hold within the regression model with fixed design (see Section 5.2), the framework of prediction of individual sequences seems to offer a unifying setting to address tuning issues both in the random and in the fixed design regression models.

4 Sparsity regret bounds for individual sequences

In this section, we prove sparsity regret bounds for different variants of an online-learning algorithm called SeqSEW and based on Sequential Sparse Exponential Weighting. We thus only consider the deterministic setting described in Figure 1. The stochastic setting will be addressed in Section 5.

Both when the forecaster has access in advance to a bound B_y on the observations $|y_t|$ (Section 4.1) and when this preliminary knowledge is not available (Sections 4.2 and 4.3), our analysis consists of two steps. We first prove in an online-learning fashion a deterministic PAC-Bayesian inequality with leading constant 1 (which actually holds for a generic prior), see Lemmas 1 and 2. The second step of our analysis follows directly from [DT08, Theorem 5] and a proper online tuning of the parameters at hand. Our main results are Propositions 2 and 3 when some a priori knowledge is available to the forecaster and Theorem 1 for the fully automatic case.

4.1 Known bounds B_y on the observations and B_Φ on the trace of the empirical Gram matrix

To simplify the analysis, we first assume that, at the beginning of the game, the number of rounds T is known to the forecaster and that he has access to a bound B_y on all the observations y_1, \dots, y_T , i.e.,

$$y_1, \dots, y_T \in [-B_y, B_y],$$

and to a bound B_Φ on the trace of the empirical Gram matrix, i.e.,

$$\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leq B_\Phi.$$

The first version of the algorithm studied in this paper is defined in Figure 4 (adaptive variants are to be introduced later). We name it *SeqSEW* for it is a variant of the Sparse Exponential Weighting algorithm introduced in the stochastic setting by [DT07, DT08] which is tailored for the prediction of individual sequences.

Algorithm: SeqSEW $_{\tau}^{B,\eta}$.

Parameters: threshold $B > 0$, inverse temperature $\eta > 0$, and prior scale $\tau > 0$ with which we associate the *sparsity prior* $\pi_{\tau} \in \mathcal{M}_1^+(\mathbb{R}^d)$ defined by

$$(10) \quad \pi_{\tau}(\mathrm{d}\mathbf{u}) \triangleq \prod_{j=1}^d \frac{(3/\tau) \mathrm{d}u_j}{2(1 + |u_j|/\tau)^4}.$$

Initialization: $p_1 \triangleq \pi_{\tau}$.

At each time round $t \geq 1$,

1. get the input data x_t and predict as

$$\hat{y}_t \triangleq \int_{\mathbb{R}^d} [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_B p_t(\mathrm{d}\mathbf{u}),$$

where

$$[z]_B \triangleq \begin{cases} -B & \text{if } z < -B \\ z & \text{if } -B \leq z \leq B \\ B & \text{if } z > B \end{cases} \quad \text{for any } z \in \mathbb{R};$$

2. get the observation y_t and compute the posterior distribution $p_{t+1} \in \mathcal{M}_1^+(\mathbb{R}^d)$ as

$$p_{t+1}(\mathrm{d}\mathbf{u}) \triangleq \frac{\exp\left(-\eta \sum_{s=1}^t \left(y_s - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_s)]_B\right)^2\right)}{W_{t+1}} \pi_{\tau}(\mathrm{d}\mathbf{u}),$$

where

$$W_{t+1} \triangleq \int_{\mathbb{R}^d} \exp\left(-\eta \sum_{s=1}^t \left(y_s - [\mathbf{v} \cdot \boldsymbol{\varphi}(x_s)]_B\right)^2\right) \pi_{\tau}(\mathrm{d}\mathbf{v}).$$

Figure 4: Definition of the SeqSEW $_{\tau}^{B,\eta}$ algorithm.

The following proposition is the main result of this section.

Proposition 2. Assume that, for a known constant $B_y > 0$, the $(x_1, y_1), \dots, (x_T, y_T)$ are such that

$$y_1, \dots, y_T \in [-B_y, B_y].$$

Then, for all $B \geq B_y$, all $\eta \leq 1/(8B^2)$, and all $\tau > 0$, the algorithm SeqSEW $_{\tau}^{B,\eta}$ satisfies

$$(11) \quad \sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + \frac{4}{\eta} \|\mathbf{u}\|_0 \ln \left(1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0 \tau} \right) \right\} + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t).$$

Proposition 2 yields the following corollary. Note that to set the parameters B , η , and τ the forecaster has to know in advance the upper bounds B_y and B_Φ . These two requirements will be removed one by one in the next two sections.

Corollary 1. *Assume that, for some known constants $B_y > 0$ and $B_\Phi > 0$, the $(x_1, y_1), \dots, (x_T, y_T)$ are such that*

$$y_1, \dots, y_T \in [-B_y, B_y] \quad \text{and} \quad \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leq B_\Phi.$$

Then, when used with

$$(12) \quad B = B_y, \quad \eta = 1/(8B_y^2), \quad \text{and} \quad \tau = \sqrt{16B_y^2/B_\Phi},$$

the algorithm $\text{SeqSEW}_\tau^{B,\eta}$ satisfies

$$(13) \quad \sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + 32 B_y^2 \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{B_\Phi} \|\mathbf{u}\|_1}{4 B_y \|\mathbf{u}\|_0} \right) \right\} + 16 B_y^2.$$

To prove Proposition 2, we first need the following deterministic PAC-Bayesian inequality which is at the core of our analysis.

Lemma 1. *Assume that for some known constant $B_y > 0$,*

$$y_1, \dots, y_T \in [-B_y, B_y].$$

For all $\tau > 0$, if the algorithm $\text{SeqSEW}_\tau^{B,\eta}$ is used with $B \geq B_y$ and $\eta \leq 1/(8B^2)$, then it satisfies

$$(14) \quad \sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_B)^2 \rho(d\mathbf{u}) + \frac{\mathcal{K}(\rho, \pi_\tau)}{\eta} \right\}$$

$$(15) \quad \leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \rho(d\mathbf{u}) + \frac{\mathcal{K}(\rho, \pi_\tau)}{\eta} \right\}.$$

Remark 1. *Lemma 1 still holds when π_τ is replaced with any prior $\pi \in \mathcal{M}_1^+(\mathbb{R}^d)$ (both in the statement of the lemma and in the definition of the algorithm SeqSEW). This fact is standard in the PAC-Bayesian approach; see, e.g., [Cat04] and [DT08]. As a consequence, any algorithm satisfying (15) will also satisfy Proposition 2 and Corollary 1.*

Remark 2. *Online PAC-Bayesian upper bounds of the same form were already proved in the framework of prediction with expert advice, e.g., in [FSSW97] and [KW99], or in the same setting as ours with a Gaussian prior π in [Vov01]. However, the Kullback-Leibler term was either obtained via an incremental update [FSSW97, KW99] or explicitly computed due to the particular form of the Gaussian prior without even the need to recognize a Kullback-Leibler divergence [Vov01]. Our analysis relies on a generic and simple convex duality argument for the Kullback-Leibler divergence used in the stochastic setting, e.g., in [Cat04] and [DT08, DT10a].*

Proof (of Lemma 1). As is usually done in the online learning setting for the study of the Exponentially Weighted Average Forecaster, our proof relies on the control of $\sum_t \eta^{-1} \ln(W_{t+1}/W_t)$ where we recall that $W_1 \triangleq 1$ and, for all $t \geq 2$,

$$W_t \triangleq \int_{\mathbb{R}^d} \exp \left(-\eta \sum_{s=1}^{t-1} \left(y_s - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_s)]_B \right)^2 \right) \pi_\tau(\mathbf{d}\mathbf{u}).$$

On the one hand, we have

$$\begin{aligned} \frac{1}{\eta} \ln \frac{W_{T+1}}{W_1} &= \frac{1}{\eta} \ln \int_{\mathbb{R}^d} \exp \left(-\eta \sum_{t=1}^T \left(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_B \right)^2 \right) \pi_\tau(\mathbf{d}\mathbf{u}) - \frac{1}{\eta} \ln 1 \\ (16) \quad &= \frac{1}{\eta} \sup_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \left(-\eta \sum_{t=1}^T \left(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_B \right)^2 \right) \rho(\mathbf{d}\mathbf{u}) - \mathcal{K}(\rho, \pi_\tau) \right\} - 0 \end{aligned}$$

$$(17) \quad = - \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \left(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_B \right)^2 \rho(\mathbf{d}\mathbf{u}) + \frac{\mathcal{K}(\rho, \pi_\tau)}{\eta} \right\},$$

where (16) follows from a convex duality argument for the Kullback-Leibler divergence (cf., e.g., [DZ98, p. 264] or [Cat04, p. 159]) which we recall in Proposition 5 in the Appendix.

On the other hand, we can rewrite W_{T+1}/W_1 as a telescopic product and get

$$\begin{aligned} \frac{1}{\eta} \ln \frac{W_{T+1}}{W_1} &= \sum_{t=1}^T \frac{1}{\eta} \ln \frac{W_{t+1}}{W_t} \\ &= \sum_{t=1}^T \frac{1}{\eta} \ln \int_{\mathbb{R}^d} \frac{\exp \left(-\eta \left(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_B \right)^2 \right) \exp \left(-\eta \sum_{s=1}^{t-1} \left(y_s - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_s)]_B \right)^2 \right)}{W_t} \pi_\tau(\mathbf{d}\mathbf{u}) \\ (18) \quad &= \sum_{t=1}^T \frac{1}{\eta} \ln \int_{\mathbb{R}^d} \exp \left(-\eta \left(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_B \right)^2 \right) p_t(\mathbf{d}\mathbf{u}), \end{aligned}$$

where (18) follows from the definition of p_t .

Let $t \in \{1, \dots, T\}$. First note that by assumption $y_t \in [-B_y, B_y] \subset [-B, B]$ so that both y_t and $[\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_B$ are $[-B, B]$ -valued for all $\mathbf{u} \in \mathbb{R}^d$. Moreover, from Proposition 6 in the Appendix, the square loss is $1/(8B^2)$ -exp-concave on $[-B, B]$ and thus η -exp-concave (since $\eta \leq 1/(8B^2)$ by assumption). Therefore, by Jensen's inequality,

$$\int_{\mathbb{R}^d} e^{-\eta \left(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_B \right)^2} p_t(\mathbf{d}\mathbf{u}) \leq \exp \left(-\eta \left(y_t - \int_{\mathbb{R}^d} [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_B p_t(\mathbf{d}\mathbf{u}) \right)^2 \right).$$

Taking the logarithms of both sides of the inequality yields

$$(19) \quad \begin{aligned} \ln \int_{\mathbb{R}^d} e^{-\eta(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_B)^2} p_t(d\mathbf{u}) &\leq -\eta \left(y_t - \int_{\mathbb{R}^d} [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_B p_t(d\mathbf{u}) \right)^2 \\ &= -\eta(y_t - \hat{y}_t)^2. \end{aligned}$$

Dividing the latter inequality by η , summing over $t \in \{1, \dots, T\}$ and combining with Equation (18), we get

$$(20) \quad \frac{1}{\eta} \ln \frac{W_{T+1}}{W_1} \leq - \sum_{t=1}^T (y_t - \hat{y}_t)^2.$$

Putting Inequalities (17) and (20) together, we get Inequality (14) of Lemma 1. As for Inequality (15), it follows from (14) by noting that

$$\forall y \in [-B, B], \quad \forall x \in \mathbb{R}, \quad |y - [x]_B| \leq |y - x|,$$

so that truncation to $[-B, B]$ can only improve prediction under the square loss if the observations are $[-B, B]$ -valued, which is the case here since by assumption $y_t \in [-B_y, B_y] \subset [-B, B]$ for all $t = 1, \dots, T$. \square

Proof (of Proposition 2). Our proof mimics the proof of Theorem 5 in [DT08]. We thus only write the outline of the proof and stress the minor changes that are needed to derive Inequality (11). The key technical tools provided in [DT08] are reproduced in appendix (Section 6.3) for the convenience of the reader.

Let $\mathbf{u}^* \in \mathbb{R}^d$. Since $B \geq B_y$ and $\eta \leq 1/(8B^2)$, we can apply Lemma 1 and get

$$(21) \quad \begin{aligned} \sum_{t=1}^T (y_t - \hat{y}_t)^2 &\leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \rho(d\mathbf{u}) + \frac{\mathcal{K}(\rho, \pi_\tau)}{\eta} \right\} \\ &\leq \underbrace{\int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \rho_{\mathbf{u}^*, \tau}(d\mathbf{u})}_{(1)} + \underbrace{\frac{\mathcal{K}(\rho_{\mathbf{u}^*, \tau}, \pi_\tau)}{\eta}}_{(2)}. \end{aligned}$$

In the last inequality, $\rho_{\mathbf{u}^*, \tau}$ is taken as the translated of π_τ at \mathbf{u}^* , namely,

$$\rho_{\mathbf{u}^*, \tau}(d\mathbf{u}) \triangleq \frac{d\pi_\tau(\mathbf{u} - \mathbf{u}^*)}{d\mathbf{u}} = \prod_{j=1}^d \frac{(3/\tau) du_j}{2(1 + |u_j - u_j^*|/\tau)^4}.$$

The two terms (1) and (2) can be upper bounded as in the proof of Theorem 5 in [DT08].

By a symmetry argument recalled in Lemma 3, the first term (1) can be rewritten as

$$(22) \quad \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \rho_{\mathbf{u}^*, \tau}(d\mathbf{u}) = \sum_{t=1}^T (y_t - \mathbf{u}^* \cdot \boldsymbol{\varphi}(x_t))^2 + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t).$$

As for the term (2), we have, as is recalled in Lemma 4,

$$(23) \quad \frac{\mathcal{K}(\rho_{\mathbf{u}^*, \tau}, \pi_\tau)}{\eta} \leq \frac{4}{\eta} \|\mathbf{u}^*\|_0 \ln \left(1 + \frac{\|\mathbf{u}^*\|_1}{\|\mathbf{u}^*\|_0 \tau} \right).$$

Combining (21), (22), and (23), which all hold for all $\mathbf{u}^* \in \mathbb{R}^d$, we get Inequality (11). \square

Proof (of Corollary 1). Applying Proposition 2, we have, since $B \geq B_y$ and $\eta \leq 1/(8B^2)$,

$$\begin{aligned} \sum_{t=1}^T (y_t - \hat{y}_t)^2 &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + \frac{4}{\eta} \|\mathbf{u}\|_0 \ln \left(1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0 \tau} \right) \right\} + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \\ (24) \quad &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + \frac{4}{\eta} \|\mathbf{u}\|_0 \ln \left(1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0 \tau} \right) \right\} + \tau^2 B_\Phi, \end{aligned}$$

since $\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leq B_\Phi$ by assumption. The particular choices for η and τ given in (12) then yield the desired inequality (13).

We end this proof with a remark on the choices of B , η , and τ suggested in (12). The best choice of (B, η) that satisfies the assumptions of Proposition 2 is $B = B_y$ and $\eta = 1/(8B_y^2)$. As for the choice of τ , it approximately minimizes the upper bound given in (24). Indeed, for all $C_1, C_2, C_3 > 0$, the function $f : (0, +\infty) \rightarrow \mathbb{R}$ defined by

$$f(\tau) \triangleq C_1 \ln \left(\frac{C_2}{\tau} \right) + C_3 \tau^2$$

has a derivative equal to $f'(\tau) = -C_1/\tau + 2C_3\tau = \tau^{-1}(2C_3\tau^2 - C_1)$, which is negative on $(0, \sqrt{C_1/(2C_3)})$ and positive on $(\sqrt{C_1/(2C_3)}, +\infty)$. The function f thus admits a global minimum in $\tau = \sqrt{C_1/(2C_3)}$. Since the sum of the last two terms of (24) is approximately of the form of $f(\tau)$ with $C_1 = 4/\eta = 32B_y^2$ and $C_3 = B_\Phi$, a reasonable choice for τ is given by

$$\tau = \sqrt{\frac{32B_y^2}{2B_\Phi}} = \sqrt{\frac{16B_y^2}{B_\Phi}}.$$

□

4.2 Unknown bound B_y on the observations but known bound B_Φ on the trace of the empirical Gram matrix

In the previous section, to prove the upper bounds stated in Lemma 1 and Proposition 2, we assumed that the forecaster had access to a bound B_y on the observations $|y_t|$. In this section, we remove this requirement and prove a sparsity regret bound for a variant of $\text{SeqSEW}_\tau^{B, \eta}$ automatically tuned in B and η and which is adaptive to the unknown bound B_y , see Proposition 3. We however still assume that the forecaster has access to the time horizon T or to a bound B_Φ on the trace of the empirical Gram matrix to derive Corollaries 2 and 3 respectively.

In this section, we consider the partially automatic algorithm defined in Figure 5 and called SeqSEW_τ^* . It differs from $\text{SeqSEW}_\tau^{B, \eta}$ defined in the previous section in that the threshold B and the inverse temperature η are now allowed to vary over time and are chosen at each time round as a function of the data available to the forecaster (we say that the threshold and the temperature parameters are *data-driven*).

The idea of truncating the base forecasts was already used in the past – see Section 3.1.3 and references therein. The key ingredient in the present paper is to perform truncation with respect to a data-driven threshold. The online tuning of this threshold is based on a pseudo-doubling-trick technique provided in [CBMS07] (we use the prefix *pseudo* since the algorithm does not need to

Algorithm: SeqSEW $^*_\tau$.

Parameter: prior scale $\tau > 0$ with which we associate the *sparsity prior* $\pi_\tau \in \mathcal{M}_1^+(\mathbb{R}^d)$ defined by

$$(10) \quad \pi_\tau(\mathbf{d}\mathbf{u}) \triangleq \prod_{j=1}^d \frac{(3/\tau) \mathbf{d}u_j}{2(1 + |u_j|/\tau)^4}.$$

Initialization: $B_1 \triangleq 0$, $\eta_1 \triangleq +\infty$, and $p_1 \triangleq \pi_\tau$.

At each time round $t \geq 1$,

1. get the input data x_t and predict as

$$\hat{y}_t \triangleq \int_{\mathbb{R}^d} [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t} p_t(\mathbf{d}\mathbf{u}),$$

where

$$[z]_{B_t} \triangleq \begin{cases} -B_t & \text{if } z < -B_t \\ z & \text{if } -B_t \leq z \leq B_t \\ B_t & \text{if } z > B_t \end{cases} \quad \text{for any } z \in \mathbb{R};$$

2. get the observation y_t and update the threshold

$$B_{t+1} \triangleq \left(2^{\lceil \log_2 \max_{1 \leq s \leq t} y_s^2 \rceil} \right)^{1/2},$$

the inverse temperature

$$\eta_{t+1} \triangleq \frac{1}{8B_{t+1}^2},$$

and the posterior distribution $p_{t+1} \in \mathcal{M}_1^+(\mathbb{R}^d)$ by

$$p_{t+1}(\mathbf{d}\mathbf{u}) \triangleq \frac{\exp \left(-\eta_{t+1} \sum_{s=1}^t \left(y_s - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_s)]_{B_s} \right)^2 \right)}{W_{t+1}} \pi_\tau(\mathbf{d}\mathbf{u}),$$

where

$$W_{t+1} \triangleq \int_{\mathbb{R}^d} \exp \left(-\eta_{t+1} \sum_{s=1}^t \left(y_s - [\mathbf{v} \cdot \boldsymbol{\varphi}(x_s)]_{B_s} \right)^2 \right) \pi_\tau(\mathbf{d}\mathbf{v}).$$

Figure 5: Definition of the SeqSEW $^*_\tau$ algorithm (with data-driven threshold and temperature parameters).

restart at the beginning of each new regime).

We first state our main results (Proposition 3 and the two corollaries of it). The two choices of τ suggested in Corollaries 2 and 3 have different purposes. The tuning given in (26) will be useful in the stochastic (batch) setting (Section 5) since in this setting T is known in advance (whereas this is not necessarily the case for B_Φ). As for the tuning given in (28), it approximately minimizes the bound of Proposition 3 (see the end of the proof of Proposition 2).

Proposition 3. *For any $\tau > 0$, the algorithm SeqSEW_τ^* satisfies*

$$(25) \quad \sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + 32B_{T+1}^2 \|\mathbf{u}\|_0 \ln \left(1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0 \tau} \right) \right\} \\ + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) + 16B_{T+1}^2 ,$$

where

$$B_{T+1}^2 \triangleq 2^{\lceil \log_2 \max_{1 \leq t \leq T} y_t^2 \rceil} \leq 2 \max_{1 \leq t \leq T} y_t^2 .$$

Remark 3. *In view of Proposition 2, the algorithm SeqSEW_τ^* satisfies a sparsity regret bound which is adaptive to the unknown bound $B_y = \max_{1 \leq t \leq T} |y_t|$. The price for the automatic tuning with respect to B_y consists only of a multiplicative factor smaller than 2 and the additive factor $16B_{T+1}^2$ which is smaller than $32B_y^2$.*

Corollary 2. *Assume that T is known to the forecaster at the beginning of the prediction game. Then, when used with*

$$(26) \quad \tau = \frac{1}{\sqrt{dT}} ,$$

the algorithm SeqSEW_τ^ satisfies*

$$(27) \quad \sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + 32B_{T+1}^2 \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ + \frac{1}{dT} \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) + 16B_{T+1}^2 ,$$

where

$$B_{T+1}^2 \triangleq 2^{\lceil \log_2 \max_{1 \leq t \leq T} y_t^2 \rceil} \leq 2 \max_{1 \leq t \leq T} y_t^2 .$$

Corollary 3. Assume that, for a known constant $B_\Phi > 0$, the $(x_1, y_1), \dots, (x_T, y_T)$ are such that

$$\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leq B_\Phi.$$

Then, when used with

$$(28) \quad \tau = \frac{1}{\sqrt{B_\Phi}},$$

the algorithm SeqSEW_τ^* satisfies

$$(29) \quad \sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + 32B_{T+1}^2 \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{B_\Phi} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} + 16B_{T+1}^2 + 1,$$

where

$$B_{T+1}^2 \triangleq 2^{\lceil \log_2 \max_{1 \leq t \leq T} y_t^2 \rceil} \leq 2 \max_{1 \leq t \leq T} y_t^2.$$

As in the previous section, to prove Proposition 3, we first need a key PAC-Bayesian inequality which is stated in the following lemma.

Lemma 2. For any $\tau > 0$, SeqSEW_τ^* satisfies

$$(30) \quad \sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t})^2 \rho(d\mathbf{u}) + 8B_{T+1}^2 \mathcal{K}(\rho, \pi_\tau) \right\} + 8B_{T+1}^2$$

$$(31) \quad \leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \rho(d\mathbf{u}) + 8B_{T+1}^2 \mathcal{K}(\rho, \pi_\tau) \right\} + 16B_{T+1}^2,$$

where

$$B_{T+1}^2 \triangleq 2^{\lceil \log_2 \max_{1 \leq t \leq T} y_t^2 \rceil}.$$

Remark 4. As noted in Remark 1, Lemma 2 still holds when π_τ is replaced with any prior $\pi \in \mathcal{M}_1^+(\mathbb{R}^d)$ (both in the statement of the lemma and in the definition of the algorithm SeqSEW). As a consequence, any algorithm satisfying (31) will also satisfy Proposition 3 and Corollaries 2 and 3.

Proof (of Lemma 2). The proof follows the same lines as for Lemma 1 except that we now need to deal with B and η changing over time. In the same spirit as in [ACBG02, CBMS07, GO07], the analysis thus now relies on the control of $(\ln W_{t+1})/\eta_{t+1} - (\ln W_t)/\eta_t$ where we recall that $W_1 \triangleq 1$ and, for all $t \geq 2$,

$$W_t \triangleq \int_{\mathbb{R}^d} \exp \left(-\eta_t \sum_{s=1}^{t-1} (y_s - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_s)]_{B_s})^2 \right) \pi_\tau(d\mathbf{u}).$$

On the one hand, we have

$$\begin{aligned}
 \frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} &= \frac{1}{\eta_{T+1}} \ln \int_{\mathbb{R}^d} \exp \left(-\eta_{T+1} \sum_{t=1}^T \left(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t} \right)^2 \right) \pi_{\tau}(\mathbf{d}\mathbf{u}) - \frac{1}{\eta_1} \ln 1 \\
 (32) \quad &= - \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \left(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t} \right)^2 \rho(\mathbf{d}\mathbf{u}) + \frac{\mathcal{K}(\rho, \pi_{\tau})}{\eta_{T+1}} \right\},
 \end{aligned}$$

where (32) follows like (17) from a convex duality argument for the Kullback-Leibler divergence.

On the other hand, we can rewrite $(\ln W_{T+1})/\eta_{T+1} - (\ln W_1)/\eta_1$ as a telescopic sum and get

$$\begin{aligned}
 \frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} &= \sum_{t=1}^T \left(\frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W_t}{\eta_t} \right) \\
 (33) \quad &= \sum_{t=1}^T \underbrace{\left(\frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W'_{t+1}}{\eta_t} \right)}_{(1)} + \underbrace{\frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t}}_{(2)},
 \end{aligned}$$

where W'_{t+1} is obtained from W_{t+1} by replacing η_{t+1} with η_t ; namely,

$$W'_{t+1} \triangleq \int_{\mathbb{R}^d} \exp \left(-\eta_t \sum_{s=1}^t \left(y_s - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_s)]_{B_s} \right)^2 \right) \pi_{\tau}(\mathbf{d}\mathbf{u}).$$

Let $t \in \{1, \dots, T\}$. The first term (1) is non-positive by Jensen's inequality. Indeed, we have,

$$\begin{aligned}
 \frac{\ln W_{t+1}}{\eta_{t+1}} &\triangleq \frac{1}{\eta_{t+1}} \ln \int_{\mathbb{R}^d} \exp \left(-\eta_{t+1} \sum_{s=1}^t \left(y_s - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_s)]_{B_s} \right)^2 \right) \pi_{\tau}(\mathbf{d}\mathbf{u}) \\
 &= \frac{1}{\eta_{t+1}} \ln \int_{\mathbb{R}^d} \exp^{\frac{\eta_{t+1}}{\eta_t}} \left(-\eta_t \sum_{s=1}^t \left(y_s - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_s)]_{B_s} \right)^2 \right) \pi_{\tau}(\mathbf{d}\mathbf{u}) \\
 (34) \quad &\leq \frac{1}{\eta_{t+1}} \ln \left(\int_{\mathbb{R}^d} \exp \left(-\eta_t \sum_{s=1}^t \left(y_s - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_s)]_{B_s} \right)^2 \right) \pi_{\tau}(\mathbf{d}\mathbf{u}) \right)^{\eta_{t+1}/\eta_t} \\
 &= \frac{\ln W'_{t+1}}{\eta_t},
 \end{aligned}$$

where (34) follows from Jensen's inequality since π is a probability distribution and the application $x \mapsto x^{\eta_{t+1}/\eta_t}$ is concave (because $\eta_{t+1} < \eta_t$). We thus get

$$(1) \triangleq \frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W'_{t+1}}{\eta_t} \leq 0.$$

As for the second term (2), we note that, as in the proof of Lemma 1, we have

$$\begin{aligned}
 &\frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t} \\
 &= \frac{1}{\eta_t} \ln \int_{\mathbb{R}^d} \frac{\exp \left(-\eta_t \left(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t} \right)^2 \right) \exp \left(-\eta_t \sum_{s=1}^{t-1} \left(y_s - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_s)]_{B_s} \right)^2 \right)}{W_t} \pi_{\tau}(\mathbf{d}\mathbf{u})
 \end{aligned}$$

$$(36) \quad = \frac{1}{\eta_t} \ln \int_{\mathbb{R}^d} \exp \left(-\eta_t \left(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t} \right)^2 \right) p_t(d\mathbf{u}),$$

where (36) follows from the definition of p_t .

We now distinguish between two cases, namely according to $t \notin \mathcal{T}$ or $t \in \mathcal{T}$, where

$$\mathcal{T} \triangleq \left\{ s \in \{1, \dots, T\}, B_{s+1} > B_s \right\}.$$

First case: $t \notin \mathcal{T}$ (i.e., $B_{t+1} = B_t$ since $(B_s)_{s \geq 1}$ is non-decreasing).

By definition of B_{t+1} , $|y_t| \leq B_{t+1} = B_t$ since $t \notin \mathcal{T}$. Thus, both $\mathbf{u} \mapsto [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t}$ and $|y_t|$ are $[-B_t, B_t]$ -valued. From the same arguments that led to Inequality (19) in the proof of Lemma 1 (essentially, the fact that the square loss is $\eta_t = 1/(8B_t^2)$ -exp-concave on $[-B_t, B_t]$), Equation (36) yields

$$(37) \quad \frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t} \leq -(y_t - \hat{y}_t)^2.$$

Second case: $t \in \mathcal{T}$ (i.e., $B_{t+1} > B_t$).

A rough upper bound on the right hand side of (36) is given by the fact that the exponential appearing in the integral is smaller than 1, so that

$$(38) \quad \begin{aligned} \frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t} &\leq \frac{1}{\eta_t} \ln 1 = 0 \\ &\leq -(y_t - \hat{y}_t)^2 + (2B_{t+1})^2, \end{aligned}$$

where (38) follows from $y_t \in [-B_{t+1}, B_{t+1}]$ (by definition of B_{t+1}) and $\hat{y}_t \in [-B_t, B_t] \subset [-B_{t+1}, B_{t+1}]$ so that $|y_t - \hat{y}_t| \leq 2B_{t+1}$.

Combining (33) with (35) and then with (37) and (38), we get

$$(39) \quad \frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} \leq - \sum_{t=1}^T (y_t - \hat{y}_t)^2 + 4 \sum_{\substack{t=1 \\ t \in \mathcal{T}}}^T B_{t+1}^2.$$

The last sum can be bounded from above as in [CBMS07] by noting that $B_{t+1}^2 \triangleq 2^{\lceil \log_2 \max_{1 \leq s \leq t} y_s^2 \rceil}$ is multiplied by at least a factor of 2 from one $t \in \mathcal{T}$ to the next one, so that

$$(40) \quad \sum_{\substack{t=1 \\ t \in \mathcal{T}}}^T B_{t+1}^2 \triangleq \sum_{\substack{t=1 \\ t \in \mathcal{T}}}^T 2^{\lceil \log_2 \max_{1 \leq s \leq t} y_s^2 \rceil} \leq 2^{1 + \lceil \log_2 \max_{1 \leq t \leq T} y_t^2 \rceil} \triangleq 2B_{T+1}^2.$$

Inequality (39) thus yields

$$(41) \quad \frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} \leq - \sum_{t=1}^T (y_t - \hat{y}_t)^2 + 8B_{T+1}^2.$$

Putting Equations (32) and (41) together, we get the PAC-Bayesian inequality

$$(42) \quad \sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \left(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t} \right)^2 \rho(d\mathbf{u}) + \frac{\mathcal{K}(\rho, \pi_\tau)}{\eta_{T+1}} \right\} + 8B_{T+1}^2,$$

which yields (30) by definition of $\eta_{T+1} \triangleq 1/(8B_{T+1}^2)$.

As for the PAC-Bayesian inequality (31), which is stated for non-truncated base forecasts, we can use the same arguments as in the proof of Lemma 1 except when $t \in \mathcal{T}$. More precisely, if $t \notin \mathcal{T}$, then $y_t \in [-B_{t+1}, B_{t+1}] = [-B_t, B_t]$ so that truncation to $[-B_t, B_t]$ can only improve prediction, i.e., for all $\mathbf{u} \in \mathbb{R}^d$,

$$\left(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t} \right)^2 \leq (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2.$$

But if $t \in \mathcal{T}$, truncation to $[-B_t, B_t]$ does not necessarily help and we only have the (rough) upper bound

$$\left(y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t} \right)^2 \leq (2B_{t+1})^2 \leq (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + 4B_{t+1}^2.$$

In view of (40), the first PAC-Bayesian inequality (30) combined with the last two upper bounds (for $t \notin \mathcal{T}$ and for $t \in \mathcal{T}$ respectively) yields (31). \square

Proof (of Proposition 3). The same arguments hold as in the proof of Proposition 2 where (11) was derived from Lemma 1 and Equations (21), (22), and (23). Here we apply Lemma 2 and restrict the infimum to the $\rho_{\mathbf{u}^*, \tau}$, $\mathbf{u}^* \in \mathbb{R}^d$, defined in (61) to get

$$\begin{aligned} \sum_{t=1}^T (y_t - \hat{y}_t)^2 &\leq \inf_{\mathbf{u}^* \in \mathbb{R}^d} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \rho_{\mathbf{u}^*, \tau}(d\mathbf{u}) + 8B_{T+1}^2 \mathcal{K}(\rho_{\mathbf{u}^*, \tau}, \pi_\tau) \right\} + 16B_{T+1}^2 \\ &\leq \inf_{\mathbf{u}^* \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u}^* \cdot \boldsymbol{\varphi}(x_t))^2 + 32B_{T+1}^2 \|\mathbf{u}^*\|_0 \ln \left(1 + \frac{\|\mathbf{u}^*\|_1}{\|\mathbf{u}^*\|_0 \tau} \right) \right\} \\ &\quad + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) + 16B_{T+1}^2, \end{aligned}$$

where the last inequality follows from Lemmas 3 and 4. \square

Proof (of Corollary 2). Equation (27) follows directly from Proposition 3 and the particular tuning $\tau = 1/\sqrt{dT}$ given in (26). \square

Proof (of Corollary 3). Equation (29) follows directly from Proposition 3 and the particular tuning $\tau = 1/\sqrt{B_\Phi}$ given in (28). \square

4.3 Unknown bounds B_y on the observations and B_Φ on the empirical Gram matrix

In the previous section, we assumed that the forecaster had access at the beginning of the prediction game to either the time horizon T (Corollary 2) or a bound B_Φ on the trace of the empirical Gram matrix (Corollary 3).

In this section, we remove these requirements and prove a sparsity regret bound for a fully automatic algorithm called SeqSEW $_*$ in the sequel. It is based on a doubling trick over the non-decreasing quantity

$$\gamma_t \triangleq \ln \left(1 + \sqrt{\sum_{s=1}^t \sum_{j=1}^d \varphi_j^2(x_s)} \right)$$

and repeatedly runs the algorithm SeqSEW $_*$ of the previous section for different values of τ as is detailed below.

More formally, the set of time rounds $t = 1, 2, \dots$ is partitioned into regimes $r = 0, 1, \dots$ whose final time instances t_r are data-driven. Let $t_{-1} \triangleq 0$ by convention. We call regime r , $r = 0, 1, \dots$, the sequence of time rounds $(t_{r-1} + 1, \dots, t_r)$ where t_r is the first date $t \geq t_{r-1} + 1$ such that $\gamma_t > 2^r$ (note that γ_t can be computed at the end of round t).

At the beginning of regime r , we restart the algorithm SeqSEW $_*$ defined in Figure 5 with the parameter $\tau = \tau_r$, where τ_r is the solution of the equation $2^r = \ln(1 + 1/\tau)$, i.e.,

$$(43) \quad \tau_r \triangleq \frac{1}{e^{2^r} - 1}.$$

Note that, because of the restarting, the threshold values B_t used during regime r are not computed on the basis of the whole past data but only on the basis of the past data observed during regime r . To avoid ambiguity, we thus rather use the notation $B_{r,t}$, so that

$$(44) \quad B_{r,t} \triangleq \left(2^{\lceil \log_2 \max_{t_{r-1}+1 \leq s \leq t-1} y_s^2 \rceil} \right)^{1/2}, \quad t \in \{t_{r-1} + 1, \dots, t_r\}.$$

Finally note that at each round t of every regime r , the values of $B_{r,t}$, $\eta_{r,t} \triangleq 1/(8B_{r,t}^2)$, and γ_t depend only on the data that is available at the time of computation (i.e., at the beginning of round t for B_t and η_t and at the end of round t for γ_t). The algorithm SeqSEW $_*$ is thus fully automatic.

Theorem 1. *Without requiring any preliminary knowledge at the beginning of the prediction game, SeqSEW $_*$ satisfies, for all $T \geq 1$ and all $(x_1, y_1), \dots, (x_T, y_T) \in \mathcal{X} \times \mathbb{R}$,*

$$(45) \quad \begin{aligned} \sum_{t=1}^T (y_t - \hat{y}_t)^2 &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + 256 \left(\max_{1 \leq t \leq T} y_t^2 \right) \|\mathbf{u}\|_0 \ln \left(e + \sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)} \right) \right. \\ &\quad \left. + 64 \left(\max_{1 \leq t \leq T} y_t^2 \right) A_T \|\mathbf{u}\|_0 \ln \left(1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \left(1 + 38 \max_{1 \leq t \leq T} y_t^2 \right) A_T, \end{aligned}$$

where $A_T \triangleq 2 + \log_2 \ln \left(e + \sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)} \right)$.

Corollary 4. Fix $s \in \mathbb{N}$ and $U > 0$. Then, for all $T \geq 1$ and all $(x_1, y_1), \dots, (x_T, y_T) \in \mathcal{X} \times \mathbb{R}$, the regret of the algorithm SeqSEW_*^* is uniformly bounded on $\{\mathbf{u} : \|\mathbf{u}\|_0 \leq s\} \cap \{\mathbf{u} : \|\mathbf{u}\|_1 \leq U\}$ by

$$(46) \quad \begin{aligned} & \sum_{t=1}^T (y_t - \hat{y}_t)^2 - \inf_{\substack{\|\mathbf{u}\|_0 \leq s \\ \|\mathbf{u}\|_1 \leq U}} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \\ & \leq 256 \left(\max_{1 \leq t \leq T} y_t^2 \right) s \ln \left(e + \sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)} \right) + 64 \left(\max_{1 \leq t \leq T} y_t^2 \right) A_T s \ln \left(1 + \frac{U}{s} \right) \\ & \quad + \left(1 + 38 \max_{1 \leq t \leq T} y_t^2 \right) A_T, \end{aligned}$$

where $A_T \triangleq 2 + \log_2 \ln \left(e + \sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)} \right)$.

Proof (of Theorem 1). We denote by R the index of the last regime and let $t_R \triangleq T$ (even if $\gamma_T \leq 2^R$).

We upper bound the regret of SeqSEW_*^* on each regime $r = 0, \dots, R$ and aggregate the resulting upper bounds by noting that

$$(47) \quad \begin{aligned} \sum_{t=1}^T (y_t - \hat{y}_t)^2 &= \sum_{r=0}^R \sum_{t=t_{r-1}+1}^{t_r} (y_t - \hat{y}_t)^2 \\ &= \sum_{r=0}^R \left((y_{t_r} - \hat{y}_{t_r})^2 + \sum_{t=t_{r-1}+1}^{t_r-1} (y_t - \hat{y}_t)^2 \right) \\ &\leq \sum_{r=0}^R \left(2(y_{t_r}^2 + B_{r,t_r}^2) + \sum_{t=t_{r-1}+1}^{t_r-1} (y_t - \hat{y}_t)^2 \right) \end{aligned}$$

$$(48) \quad \leq \sum_{r=0}^R \left(\sum_{t=t_{r-1}+1}^{t_r-1} (y_t - \hat{y}_t)^2 \right) + 6(R+1)y_T^{*2},$$

where we define y_T^* by

$$y_T^* \triangleq \max_{1 \leq t \leq T} |y_t|.$$

To get Inequality (47) we used the elementary upper bound $(y_{t_r} - \hat{y}_{t_r})^2 \leq 2(y_{t_r}^2 + \hat{y}_{t_r}^2) \leq 2(y_{t_r}^2 + B_{r,t_r}^2)$ (since $\hat{y}_{t_r} \in [-B_{r,t_r}, B_{r,t_r}]$ by definition of \hat{y}_{t_r}). As for Inequality (48), it follows from $y_{t_r}^2 \leq y_T^{*2}$ and

$$B_{r,t_r}^2 \triangleq 2^{\lceil \log_2 \max_{t_{r-1}+1 \leq t \leq t_r-1} y_t^2 \rceil} \leq 2 \max_{t_{r-1}+1 \leq t \leq t_r-1} y_t^2 \leq 2 y_T^{*2}.$$

But for all $r = 0, \dots, R$, by definition of t_r , we have $\gamma_{t_{r-1}} \leq 2^r$ which is equivalent to

$$\sum_{t=1}^{t_r-1} \sum_{j=1}^d \varphi_j^2(x_t) \leq (e^{2^r} - 1)^2,$$

so that, by the nonnegativity of the $\varphi_j^2(x_t)$ for $1 \leq t \leq t_{r-1}$,

$$\sum_{t=t_{r-1}+1}^{t_r-1} \sum_{j=1}^d \varphi_j^2(x_t) \leq (e^{2^r} - 1)^2.$$

Since in addition $\tau_r \triangleq 1/\sqrt{(e^{2^r} - 1)^2}$, we can apply Corollary 3 on each period $\{t_{r-1}+1, \dots, t_r-1\}$, $r = 0, \dots, R$, with $B_\Phi = (e^{2^r} - 1)^2$ and get from (48) the upper bound

$$(49) \quad \sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \sum_{r=0}^R \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=t_{r-1}+1}^{t_r-1} (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + \Delta_r(\mathbf{u}) \right\} + 6(R+1)y_T^{*2},$$

where

$$(50) \quad \Delta_r(\mathbf{u}) \triangleq 32B_{t_r}^2 \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{(e^{2^r} - 1)^2} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) + 16B_{r,t_r}^2 + 1.$$

Since the infimum is superadditive, i.e., $\sum_r \inf_u \leq \inf_u \sum_r$, (49) yields

$$(51) \quad \begin{aligned} \sum_{t=1}^T (y_t - \hat{y}_t)^2 &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \sum_{r=0}^R \left(\sum_{t=t_{r-1}+1}^{t_r-1} (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + \Delta_r(\mathbf{u}) \right) + 6(R+1)y_T^{*2} \\ &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{r=0}^R \sum_{t=t_{r-1}+1}^{t_r} (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + \sum_{r=0}^R \Delta_r(\mathbf{u}) \right\} + 6(R+1)y_T^{*2} \\ &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + \sum_{r=0}^R \Delta_r(\mathbf{u}) \right\} + 6(R+1)y_T^{*2}. \end{aligned}$$

Let $\mathbf{u} \in \mathbb{R}^d$. It now remains to upper bound $\sum_{r=0}^R \Delta_r(\mathbf{u})$ and $6(R+1)y_T^{*2}$. Together with the above-mentioned upper bound $B_{r,t_r}^2 \leq 2y_T^{*2}$ and the elementary inequality $\ln(1 + xy) \leq \ln((1+x)(1+y)) = \ln(1+x) + \ln(1+y)$ for all $x, y > 0$, (50) yields

$$\begin{aligned} \Delta_r(\mathbf{u}) &\leq 64 y_T^{*2} \|\mathbf{u}\|_0 \ln \left(1 + \frac{(e^{2^r} - 1) \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) + 32 y_T^{*2} + 1 \\ &\leq 64 y_T^{*2} \|\mathbf{u}\|_0 \ln(1 + (e^{2^r} - 1)) + 64 y_T^{*2} \|\mathbf{u}\|_0 \ln \left(1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) + 32 y_T^{*2} + 1 \\ &\leq 64 y_T^{*2} \|\mathbf{u}\|_0 2^r + 64 y_T^{*2} \|\mathbf{u}\|_0 \ln \left(1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) + 32 y_T^{*2} + 1. \end{aligned}$$

Summing over $r = 0, \dots, R$, we get

$$(52) \quad \sum_{r=0}^R \Delta_r(\mathbf{u}) \leq 64 y_T^{*2} \|\mathbf{u}\|_0 \sum_{r=0}^R 2^r + (R+1) \left(64 y_T^{*2} \|\mathbf{u}\|_0 \ln \left(1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) + 32 y_T^{*2} + 1 \right).$$

To finish the proof, we distinguish between the two cases $R = 0$ or $R \geq 1$.

First case: $R = 0$

Then $\sum_{r=0}^R 2^r = 1$ and $R + 1 = 1$ so that (52) yields

$$\sum_{r=0}^R \Delta_r(\mathbf{u}) \leq 64 y_T^{*2} \|\mathbf{u}\|_0 + 64 y_T^{*2} \|\mathbf{u}\|_0 \ln \left(1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) + 32 y_T^{*2} + 1.$$

Noting that $6(R + 1)y_T^{*2} = 6y_T^{*2} \triangleq 6 \max_{1 \leq t \leq T} y_t^2$, that $\ln \left(e + \sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)} \right) \geq 1$, and that $A_T \geq 2 + \log_2 1 \geq 1$, Inequality (51) combined with the last upper bound concludes the proof in the case when $R = 0$.

Second case: $R \geq 1$

Since $R \geq 1$ and by definition of t_{R-1} ,

$$\begin{aligned} 2^{R-1} &< \gamma_{t_{R-1}} \triangleq \ln \left(1 + \sqrt{\sum_{t=1}^{t_{R-1}} \sum_{j=1}^d \varphi_j^2(x_t)} \right) \\ (53) \quad &\leq \ln \left(e + \sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)} \right). \end{aligned}$$

From the above inequality and $\sum_{r=0}^R 2^r = 2^{R+1} - 1 \leq 4 \cdot 2^{R-1}$ we get

$$\sum_{r=0}^R 2^r \leq 4 \ln \left(e + \sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)} \right).$$

Note that from (53) we also have $R + 1 \leq 2 + \log_2 \ln \left(e + \sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)} \right) \triangleq A_T$, so that, on the one hand, via (52),

$$\begin{aligned} \sum_{r=0}^R \Delta_r(\mathbf{u}) &\leq 256 y_T^{*2} \|\mathbf{u}\|_0 \ln \left(e + \sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)} \right) + 64 y_T^{*2} A_T \|\mathbf{u}\|_0 \ln \left(1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \\ &\quad + A_T (32 y_T^{*2} + 1) \end{aligned}$$

and, on the other hand,

$$6(R + 1)y_T^{*2} \leq 6A_T y_T^{*2}.$$

Combining (51) with the two inequalities above and noting that $y_T^{*2} = \max_{1 \leq t \leq T} y_t^2$ conclude the proof in the case when $R \geq 1$. \square

Proof (of Corollary 4). The proof is straightforward. In view of Theorem 1, we just need to check that the quantity (continuously extended in $s = 0$)

$$256 \left(\max_{1 \leq t \leq T} y_t^2 \right) s \ln \left(e + \sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)} \right) + 64 \left(\max_{1 \leq t \leq T} y_t^2 \right) A_T s \ln \left(1 + \frac{U}{s} \right)$$

is non-decreasing in $s \in \mathbb{R}_+$ and in $U \in \mathbb{R}_+$.

This is clear for U . The fact that it is also non-decreasing in s comes from the following remark. For all $U \geq 0$, the function $s \in (0, +\infty) \mapsto s \ln(1 + U/s)$ has a derivative equal to

$$\ln\left(1 + \frac{U}{s}\right) - \frac{U/s}{1 + U/s} \quad \text{for all } s > 0.$$

From the elementary inequality

$$\ln(1 + u) = -\ln\left(\frac{1}{1 + u}\right) \geq -\left(\frac{1}{1 + u} - 1\right) = \frac{u}{1 + u},$$

which holds for all $u \in (-1, +\infty)$, the above derivative is nonnegative for all $s > 0$ so that the continuous extension $s \in \mathbb{R}_+ \mapsto s \ln(1 + \frac{U}{s})$ is non-decreasing. \square

5 Adaptivity to the unknown variance in the stochastic setting

In this section, we apply the online-learning algorithm SeqSEW_τ^* of Section 4.2 to the stochastic settings described in Figure 2 and 3, namely, the regression models with random and fixed design respectively. The sparsity regret bounds proved for this algorithm on individual sequences imply in both settings sparsity oracle inequalities with leading constant 1 which are of the same flavour as in [DT08, DT10a] but are adaptive (up to a logarithmic factor) to the unknown variance σ^2 of the noise if the latter is Gaussian. Weaker bounds are also proved under weaker assumptions like an exponential moment assumption on the noise distribution. Besides, for the case of random design, our inequalities hold with respect to linear combinations \mathbf{u} in the whole \mathbb{R}^d space instead of being restricted to ℓ^1 -balls with finite radii. These results thus answer two questions left open in [DT10a]; see Remark 5 below.

This section is organized as follows. In Section 5.1 we derive sharp sparsity oracle inequalities in the regression model with random design. We then explain in Section 5.2 how similar bounds can be obtained in the case of a fixed design.

5.1 Regression model with random design

In this section, we consider the regression model with random design under the square loss described in Figure 2. In particular, the time horizon T is known in advance to the forecaster and the observed pairs $(X_1, Y_1), \dots, (X_T, Y_T) \in \mathcal{X} \times \mathbb{R}$ are i.i.d. with an unknown distribution.

In this setting, even if the whole sample $(X_1, Y_1), \dots, (X_T, Y_T)$ is available at the beginning of the prediction game, we treat it in a sequential fashion. We run the algorithm SeqSEW_τ^* defined in Figure 5 from time 1 to time T with $\tau = 1/\sqrt{dT}$. We then define our data-based regressor \hat{f}_T as the uniform average

$$\hat{f}_T \triangleq \frac{1}{T} \sum_{t=1}^T \tilde{f}_t$$

of the regressors $\tilde{f}_t : \mathcal{X} \rightarrow \mathbb{R}$ sequentially built by the algorithm SeqSEW_τ^* as

$$\tilde{f}_t(x) \triangleq \int_{\mathbb{R}^d} [\mathbf{u} \cdot \boldsymbol{\varphi}(x)]_{B_t} p_t(d\mathbf{u}).$$

Therefore, \widehat{f}_T is a mirror-averaging type aggregation procedure in the same spirit as in [DT10a] or, e.g., the previous works [Cat04, CBCG04, JRT08]. However, contrary to [Cat04, JRT08, DT10a], the regressors $\widetilde{f}_t : \mathcal{X} \rightarrow \mathbb{R}$ are tuned online so that \widehat{f}_T does not depend on any prior knowledge on the unknown distribution of the (X_t, Y_t) , $1 \leq t \leq T$, such as the unknown variance σ^2 of the noise,

$$\sigma^2 \triangleq \mathbb{E}[\varepsilon^2] = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2],$$

the $\|\varphi_j\|_\infty$, or the $\|f - \varphi_j\|_\infty$ (actually, the φ_j and the $f - \varphi_j$ do not even need to be bounded).

Theorem 2. *Assume that $(X_1, Y_1), \dots, (X_T, Y_T) \in \mathcal{X} \times \mathbb{R}$ are independent random copies of $(X, Y) \in \mathcal{X} \times \mathbb{R}$, where $\mathbb{E}[Y^2] < +\infty$ and $\|\varphi_j\|_{L^2}^2 \triangleq \mathbb{E}[\varphi_j(X)^2] < +\infty$ for all $j = 1, \dots, d$. Then, the data-based regressor \widehat{f}_T defined above satisfies*

$$\begin{aligned} \mathbb{E} \left[\left\| f - \widehat{f}_T \right\|_{L^2}^2 \right] &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \|f - \mathbf{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 + 64 \frac{\mathbb{E}[\max_{1 \leq t \leq T} Y_t^2]}{T} \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \frac{1}{dT} \sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + 32 \frac{\mathbb{E}[\max_{1 \leq t \leq T} Y_t^2]}{T}. \end{aligned}$$

The previous theorem can be used in the following particular cases.

Corollary 5. *Assume that $(X_1, Y_1), \dots, (X_T, Y_T) \in \mathcal{X} \times \mathbb{R}$ are independent random copies of $(X, Y) \in \mathcal{X} \times \mathbb{R}$, that $\sup_{1 \leq j \leq d} \|\varphi_j\|_{L^2}^2 < +\infty$, that $\mathbb{E}|Y| < +\infty$, and that one of the following assumptions holds on the distribution of $\Delta Y \triangleq Y - \mathbb{E}[Y]$.*

(BD(B)) : $|\Delta Y| \leq B$ almost surely for a given constant $B > 0$;

(SG(σ^2)) : ΔY is subgaussian with variance factor $\sigma^2 > 0$, that is, $\mathbb{E}[e^{\lambda \Delta Y}] \leq e^{\lambda^2 \sigma^2 / 2}$ for all $\lambda \in \mathbb{R}$;

(BEM(α, M)) : ΔY has a bounded exponential moment, that is, $\mathbb{E}[e^{\alpha |\Delta Y|}] \leq M$ for some given constants $\alpha > 0$ and $M > 0$;

(BM(α, M)) : ΔY has a bounded moment, that is, $\mathbb{E}[|\Delta Y|^\alpha] \leq M$ for some given constants $\alpha > 2$ and $M > 0$.

Then, the data-based regressor \widehat{f}_T defined above satisfies

$$\begin{aligned} \mathbb{E} \left[\left\| f - \widehat{f}_T \right\|_{L^2}^2 \right] &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \|f - \mathbf{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 + 128 \left(\frac{\mathbb{E}[Y]^2}{T} + \psi_T \right) \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \frac{1}{dT} \sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + 64 \left(\frac{\mathbb{E}[Y]^2}{T} + \psi_T \right), \end{aligned}$$

where

$$\psi_T \triangleq \frac{1}{T} \mathbb{E} \left[\max_{1 \leq t \leq T} (Y_t - \mathbb{E}[Y_t])^2 \right] \leq \begin{cases} \frac{B^2}{T} & \text{under Assumption (BD(B))}, \\ \frac{\sigma^2}{T} + 2 C_T \sigma^2 \frac{\ln T}{T} & \text{under Assumption (SG}(\sigma^2)), \\ \frac{\ln^2((M+e)T)}{\alpha^2 T} & \text{under Assumption (BEM}(\alpha, M)), \\ \frac{M^{2/\alpha}}{T^{(\alpha-2)/\alpha}} & \text{under Assumption (BM}(\alpha, M)), \end{cases}$$

for some sequence $(C_T)_{T \geq 1}$ of positive real numbers depending only on T and such that

$$\sup_{T \geq 1} C_T \leq 8 \quad \text{and} \quad \lim_{T \rightarrow +\infty} C_T = 1.$$

The above corollary answers two questions left open in [DT10a] as detailed in Remark 5. First note that several classical assumptions on Y expressed in terms of $f(X)$ and ε are either particular cases of the above corollary or can be treated similarly. Indeed, each of the four assumptions above on $\Delta Y \triangleq Y - \mathbb{E}[Y] = f(X) - \mathbb{E}[f(X)] + \varepsilon$ is satisfied as soon as both the distribution of $f(X) - \mathbb{E}[f(X)]$ and the conditional distribution of ε (conditionally on X) satisfy the same type of assumption. For example, if $f(X) - \mathbb{E}[f(X)]$ is subgaussian with variance factor σ_X^2 and if ε is subgaussian conditionally on X with a variance factor uniformly bounded by a constant σ_ε^2 , then ΔY is subgaussian with variance factor $\sigma_X^2 + \sigma_\varepsilon^2$ (see also Remark 8 to avoid conditioning).

The assumptions on $f(X) - \mathbb{E}[f(X)]$ and ε can also be mixed together. For instance, as explained in Remark 8 after the proof of the present corollary, under the classical assumptions

$$(54) \quad \|f\|_\infty < +\infty \quad \text{and} \quad \mathbb{E}[e^{\alpha|\varepsilon|} \mid X] \leq M \quad \text{a.s.}$$

or

$$(55) \quad \|f\|_\infty < +\infty \quad \text{and} \quad \mathbb{E}[e^{\lambda\varepsilon} \mid X] \leq e^{\lambda^2 \sigma^2 / 2} \quad \text{a.s.,} \quad \forall \lambda \in \mathbb{R},$$

the key quantity ψ_T in the corollary can be bounded from above by

$$\psi_T \leq \begin{cases} \frac{8 \|f\|_\infty^2}{T} + \frac{2 \ln^2((M+e)T)}{\alpha^2 T} & \text{under the set of assumptions (54),} \\ \frac{8 \|f\|_\infty^2}{T} + \frac{2\sigma^2}{T} + 4 C_T \sigma^2 \frac{\ln T}{T} & \text{under the set of assumptions (55).} \end{cases}$$

The previous comments lead to the following remark, whose technical explanation is given at the end of the section.

Remark 5. *Our results answer two questions left open in [DT10a]. Indeed, under the assumptions*

$$\|f\|_\infty < +\infty \quad \text{and} \quad \mathbb{E}[e^{\lambda\varepsilon} \mid X] \leq e^{\lambda^2 \sigma^2 / 2} \quad \text{a.s.,} \quad \forall \lambda \in \mathbb{R},$$

the bound stated in Corollary 5 is of the same flavour as the one proved in Proposition 1 of [DT10a]. The authors however made two criticisms about their results. First, as noted in [DT10a, Section 4.2], their sparsity oracle inequality only holds on a ℓ^1 -ball with a finite radius R which

has to be chosen in advance by the forecaster (thus somehow assuming that the forecaster has access to an a priori bound R^* on the ℓ^1 -norm of a vector \mathbf{u}^* minimizing some regularized risk; see the proof of the present claim). Second, as noted in [DT10a, Section 5.1, Remark 7], their algorithm requires the a priori knowledge of the variance factor σ^2 of the noise. The authors thus asked the two following questions:

- Is it possible to extend the infimum to the whole \mathbb{R}^d space? (or equivalently, can we adapt to the unknown ℓ^1 -norm $\|\mathbf{u}^*\|_1$?)
- Is it possible to adapt to the unknown variance factor σ^2 ?

Our results show that both questions have a positive answer, as indicated in pages 38-39.

Proof (of Theorem 2). By Corollary 2 and the definitions of \tilde{f}_t above and $\hat{y}_t \triangleq \tilde{f}_t(X_t)$ in Figure 5, we have, almost surely,

$$\begin{aligned} \sum_{t=1}^T (Y_t - \tilde{f}_t(X_t))^2 &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (Y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(X_t))^2 + 64 \left(\max_{1 \leq t \leq T} Y_t^2 \right) \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \frac{1}{dT} \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(X_t) + 32 \max_{1 \leq t \leq T} Y_t^2. \end{aligned}$$

It remains to take the expectations of both sides with respect to $((X_1, Y_1), \dots, (X_T, Y_T))$. First note that for all $t = 1, \dots, T$, defining $\varepsilon_t \triangleq Y_t - f(X_t)$, we have

$$\begin{aligned} \mathbb{E} \left[(Y_t - \tilde{f}_t(X_t))^2 \right] &= \mathbb{E} \left[(\varepsilon_t + f(X_t) - \tilde{f}_t(X_t))^2 \right] \\ &= \sigma^2 + \mathbb{E} \left[(f(X_t) - \tilde{f}_t(X_t))^2 \right], \end{aligned}$$

since $\mathbb{E}[\varepsilon_t^2] = \mathbb{E}[\varepsilon^2] \triangleq \sigma^2$ on the one hand, and, on the other hand, \tilde{f}_t is a measurable function of $(X_s, Y_s)_{1 \leq s \leq t-1}$ and $\mathbb{E}[\varepsilon_t | (X_s, Y_s)_{1 \leq s \leq t-1}, X_t] = \mathbb{E}[\varepsilon_t | X_t] = 0$ (from the independence of $(X_s, Y_s)_{1 \leq s \leq t-1}$ and (X_t, Y_t) and by definition of f).

In the same way,

$$\mathbb{E} \left[(Y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(X_t))^2 \right] = \sigma^2 + \mathbb{E} \left[(f(X_t) - \mathbf{u} \cdot \boldsymbol{\varphi}(X_t))^2 \right].$$

Therefore, by Jensen's inequality and the concavity of the infimum, the last inequality becomes, after taking the expectations of both sides,

$$\begin{aligned} T\sigma^2 + \sum_{t=1}^T \mathbb{E} \left[(f(X_t) - \tilde{f}_t(X_t))^2 \right] &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ T\sigma^2 + \sum_{t=1}^T \mathbb{E} \left[(f(X_t) - \mathbf{u} \cdot \boldsymbol{\varphi}(X_t))^2 \right] \right. \\ &\quad \left. + 64 \mathbb{E} \left[\max_{1 \leq t \leq T} Y_t^2 \right] \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \frac{1}{dT} \sum_{j=1}^d \sum_{t=1}^T \mathbb{E} [\varphi_j^2(X_t)] + 32 \mathbb{E} \left[\max_{1 \leq t \leq T} Y_t^2 \right]. \end{aligned}$$

Noting that the $T\sigma^2$ cancel out, dividing the two sides by T , and using the fact that $X_t \sim X$ in the right-hand side, we get

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[(f(X_t) - \tilde{f}_t(X_t))^2 \right] &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \|f - \mathbf{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 \right. \\ &\quad \left. + 64 \frac{\mathbb{E} [\max_{1 \leq t \leq T} Y_t^2]}{T} \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \frac{1}{dT} \sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + 32 \frac{\mathbb{E} [\max_{1 \leq t \leq T} Y_t^2]}{T}. \end{aligned}$$

The right-hand side of the last inequality is exactly the upper bound stated in Theorem 2. To conclude the proof, we thus only need to check that $\|f - \hat{f}_T\|_{L^2}^2$ is bounded from above by the left-hand side. But by definition of \hat{f}_T and by convexity of the square loss we have

$$\begin{aligned} \mathbb{E} \left[\|f - \hat{f}_T\|_{L^2}^2 \right] &\triangleq \mathbb{E} \left[\left(f(X) - \frac{1}{T} \sum_{t=1}^T \tilde{f}_t(X) \right)^2 \right] \\ &\leq \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[(f(X) - \tilde{f}_t(X))^2 \right] = \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[(f(X_t) - \tilde{f}_t(X_t))^2 \right]. \end{aligned}$$

The last equality follows classically from the fact that, for all $t = 1, \dots, T$, $(X_s, Y_s)_{1 \leq s \leq t-1}$ (on which \tilde{f}_t is constructed) is independent from both X_t and X and the fact that $X_t \sim X$. \square

Remark 6. The fact that the inequality stated in Corollary 2 has a leading constant equal to 1 on individual sequences is crucial to derive in the stochastic setting an oracle inequality in terms of the (excess) risks $\mathbb{E} [\|f - \hat{f}_T\|_{L^2}^2]$ and $\|f - \mathbf{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2$. Indeed, if the constant appearing in front of the infimum was equal to $C > 1$, then the $T\sigma^2$ would not cancel out in the previous proof, so that the resulting expected inequality would contain a non-vanishing additive term $(C - 1)\sigma^2$.

Proof (of Corollary 5). We can apply Theorem 2. Then, to prove the upper bound on $\mathbb{E} [\|f - \hat{f}_T\|_{L^2}^2]$, it suffices to show that

$$(56) \quad \frac{\mathbb{E} [\max_{1 \leq t \leq T} Y_t^2]}{T} \leq 2 \left(\frac{\mathbb{E}[Y]^2}{T} + \psi_T \right).$$

Recall that

$$\psi_T \triangleq \frac{1}{T} \mathbb{E} \left[\max_{1 \leq t \leq T} (Y_t - \mathbb{E}[Y_t])^2 \right] = \frac{1}{T} \mathbb{E} \left[\max_{1 \leq t \leq T} (\Delta Y)_t^2 \right],$$

where we defined $(\Delta Y)_t \triangleq Y_t - \mathbb{E}[Y_t] = Y_t - \mathbb{E}[Y]$ for all $t = 1, \dots, T$.

From the elementary inequality $(x + y)^2 \leq 2x^2 + 2y^2$ for all $x, y \in \mathbb{R}$, we have

$$(57) \quad \mathbb{E} \left[\max_{1 \leq t \leq T} Y_t^2 \right] \triangleq \mathbb{E} \left[\max_{1 \leq t \leq T} (\mathbb{E}[Y] + (\Delta Y)_t)^2 \right] \leq 2\mathbb{E}[Y]^2 + 2\mathbb{E} \left[\max_{1 \leq t \leq T} (\Delta Y)_t^2 \right].$$

Dividing both sides by T , we get (56).

As for the upper bound on ψ_T , since the $(\Delta Y)_t$, $1 \leq t \leq T$, are distributed as ΔY , we can apply Lemma 5 (with $\beta = \sqrt{\ln 2}/2$), Lemma 6, and Lemma 7 in the appendix (Section 6.4) to bound ψ_T from above under the assumptions $(\text{SG}(\sigma^2))$, $(\text{BEM}(\alpha, M))$, and $(\text{BM}(\alpha, M))$ respectively (the upper bound under $(\text{BD}(\text{B}))$ is straightforward) :

$$\mathbb{E} \left[\max_{1 \leq t \leq T} (\Delta Y)_t^2 \right] \leq \begin{cases} B^2 & \text{if Assumption } (\text{BD}(\text{B})) \text{ holds,} \\ \sigma^2 + 2 C_T \sigma^2 \ln T & \text{if Assumption } (\text{SG}(\sigma^2)) \text{ holds,} \\ \frac{\ln^2((M+e)T)}{\alpha^2} & \text{if Assumption } (\text{BEM}(\alpha, M)) \text{ holds,} \\ (MT)^{2/\alpha} & \text{if Assumption } (\text{BM}(\alpha, M)) \text{ holds,} \end{cases}$$

where we chose $C_T \triangleq \gamma_T(\sqrt{\ln 2}/2)$ as given in Lemma 5. In particular, $\sup_{T \geq 1} C_T \leq 8$ and $\lim_{T \rightarrow +\infty} C_T = 1$. This concludes the proof. \square

Remark 7. The two terms $\mathbb{E}[Y]^2/T$ appearing in Corollary 5 are not necessary. They can be avoided via a slightly more sophisticated way of truncating the predictions. Indeed, at each time round t , instead of truncating the predictions to $[-B_t, B_t]$, which is best suited to the case $\mathbb{E}[Y] = 0$, we could have truncated them to the smallest dyadic interval I_t containing

$$\left[\min_{1 \leq s \leq t-1} Y_s, \max_{1 \leq s \leq t-1} Y_s \right].$$

If η_t is changed accordingly, namely, $\eta_t = 1/(8|I_t|^2)$, the resulting upper bound on $\mathbb{E}[\|f - \hat{f}_T\|_{L^2}^2]$ is the same bound as the one of Corollary 5, but without the two terms $\mathbb{E}[Y]^2/T$. This is essentially due to the fact that, with this more sophisticated truncation, we can replace $\mathbb{E}[\max_{1 \leq t \leq T} Y_t^2]$ in Theorem 2 with

$$\begin{aligned} \frac{1}{2} \mathbb{E} \left[\left(\max_{1 \leq t \leq T} Y_t - \min_{1 \leq t \leq T} Y_t \right)^2 \right] &= \frac{1}{2} \mathbb{E} \left[\left(\max_{1 \leq t \leq T} (Y_t - \mathbb{E}[Y]) - \min_{1 \leq t \leq T} (Y_t - \mathbb{E}[Y]) \right)^2 \right] \\ &\leq 2 \mathbb{E} \left[\max_{1 \leq t \leq T} (\Delta Y)_t^2 \right], \end{aligned}$$

which is the right-hand side of (57) without $2 \mathbb{E}[Y]^2$.

Remark 8. We mentioned after Corollary 5 that each of the four assumptions on ΔY is fulfilled as soon as both the distribution of $f(X) - \mathbb{E}[f(X)]$ and the conditional distribution of ε (conditionally on X) satisfy the same type of assumption. It actually extends to the more general case when the conditional distribution of ε given X is replaced with the distribution of ε itself (without conditioning). This relies on the elementary upper bound

$$\begin{aligned} \mathbb{E} \left[\max_{1 \leq t \leq T} (\Delta Y)_t^2 \right] &= \mathbb{E} \left[\max_{1 \leq t \leq T} (f(X_t) - \mathbb{E}[f(X)] + \varepsilon_t)^2 \right] \\ &\leq 2 \mathbb{E} \left[\max_{1 \leq t \leq T} (f(X_t) - \mathbb{E}[f(X)])^2 \right] + 2 \mathbb{E} \left[\max_{1 \leq t \leq T} \varepsilon_t^2 \right]. \end{aligned}$$

From the last inequality, we can also see that assumptions of different nature can be made on $f(X) - \mathbb{E}[f(X)]$ and ε , such as the assumptions given in (54) or in (55).

Technical explanation of Remark 5. We recall below Proposition 1 of [DT10a] with our notations (e.g., β therein corresponds to $1/\eta$ in this paper).

Proposition 4 (Proposition 1 of [DT10a]). *Assume that for some constant $L_\varphi > 0$ we have $\max_{1 \leq j \leq d} \|\varphi_j\|_\infty \leq L_\varphi$. If, in addition, the errors ε_t have a bounded exponential moment:*

$$\exists b, \sigma^2 > 0 \quad \text{such that} \quad \mathbb{E}[e^{\lambda \varepsilon} | X] \leq e^{\sigma^2 \lambda^2 / 2}, \quad \forall |\lambda| \leq b, \quad \text{a.s.},$$

then, for every $R > 0$ and $\eta \leq \min \left\{ \left(2\sigma^2 + 2 \sup_{\|\mathbf{u}\|_1 \leq R} \|\mathbf{u} \cdot \boldsymbol{\varphi} - f\|_\infty^2 \right)^{-1}, b/(4RL_\varphi) \right\}$, the mirror averaging aggregate $\hat{f}_T : \mathcal{X} \rightarrow \mathbb{R}$ defined in [DT10a, Equations (1) and (3)] satisfies

$$\mathbb{E} \left[\left\| f - \hat{f}_T \right\|_{L^2}^2 \right] \leq \inf_{\mathbf{u}} \left\{ \left\| f - \mathbf{u} \cdot \boldsymbol{\varphi} \right\|_{L^2}^2 + \frac{4}{\eta(T+1)} \sum_{j=1}^d \ln(1 + |u_j|/\tau) \right\} + \Delta_T(\eta, \tau),$$

where the infimum is taken over all $\mathbf{u} \in \mathbb{R}^d$ such that $\|\mathbf{u}\|_1 \leq R - 2d\tau$ and where

$$\Delta_T(\eta, \tau) \triangleq 4\tau^2 \sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + \eta^{-1}(T+1)^{-1}.$$

In the next paragraphs we compare the above proposition with a consequence of Corollary 5 under the assumptions

$$\|f\|_\infty < +\infty \quad \text{and} \quad \mathbb{E}[e^{\lambda \varepsilon_1} | X_1] \leq e^{\lambda^2 \sigma^2 / 2} \quad \text{a.s.}, \quad \forall \lambda \in \mathbb{R}.$$

Under the above assumptions, the sparsity oracle inequality stated in Corollary 5 yields

(59)

$$\begin{aligned} \mathbb{E} \left[\left\| f - \hat{f}_T \right\|_{L^2}^2 \right] &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \left\| f - \mathbf{u} \cdot \boldsymbol{\varphi} \right\|_{L^2}^2 + 128 \left(\frac{\mathbb{E}[Y]^2}{T} + \psi_T \right) \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \frac{1}{dT} \sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + 64 \left(\frac{\mathbb{E}[Y]^2}{T} + \psi_T \right), \end{aligned}$$

where $\psi_T \leq 8 \|f\|_\infty^2 / T + 2\sigma^2 / T + 32 \sigma^2 (\ln T) / T$ as shown in Remark 8. Besides, as explained in Remark 7, the two terms $\mathbb{E}[Y]^2 / T$ above can be avoided for a more sophisticated variant of the algorithm SeqSEW, so that we omit them in the following comparison.

First consider the main remainder terms (i.e., the logarithmic quantities inside the curly brackets). Our main remainder term is proportional to the right-hand side of the upper bound

$$\sum_{j=1}^d \ln(1 + \sqrt{dT} |u_j|) \leq \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right),$$

but our sparsity oracle inequality of course also holds with the tighter quantity of the left-hand side (see Lemma 4 in the appendix). The latter quantity also appears in Proposition 4 above with \sqrt{dT} replaced with $1/\tau$. Since τ should be chosen at most of order R/d (so that the condition $\|\mathbf{u}\|_1 \leq R - 2d\tau$ can be satisfied), our quantity $\sum_{j=1}^d \ln(1 + \sqrt{dT} |u_j|)$ is at most of the same

order as $\sum_{j=1}^d \ln(1 + |u_j|/\tau)$ (up to logarithmic factors). Moreover, the multiplicative factor $1/(\eta(T+1))$ in Proposition 4 above is at least of order

$$\frac{\sigma^2}{T} + \frac{1}{T} \sup_{\|\mathbf{u}\|_1 \leq R} \|\mathbf{u} \cdot \boldsymbol{\varphi} - f\|_\infty^2 ,$$

while ψ_T is at most of order $\sigma^2(\ln T)/T + \|f\|_\infty^2/T$ in our bound. The main remainder terms of Proposition 4 above and in our risk bound (59) are thus comparable.

As for the secondary remainder terms (i.e., the ones outside the curly brackets), note that [DT10a] suggest to choose the tuning parameter τ as $\tau = \min \left\{ \left(\eta T \sum_{j=1}^d \|\varphi_j\|_{L^2}^2 \right)^{-1/2}, R/(4d) \right\}$. The corresponding secondary remainder term $\Delta_T(\eta, \tau)$ is at least of order

$$\frac{1}{\eta T} \gtrsim \frac{\sigma^2}{T} + \frac{1}{T} \sup_{\|\mathbf{u}\|_1 \leq R} \|\mathbf{u} \cdot \boldsymbol{\varphi} - f\|_\infty^2 .$$

Our bound (59) was derived with τ of order $(Td)^{-1/2}$; our secondary remainder term is of order at most

$$\frac{\sigma^2 \ln T}{T} + \frac{1}{dT} \sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + \|f\|_\infty^2/T .$$

Our secondary remainder term is thus also comparable to the one of Proposition 4 above. Putting the main and secondary remainder terms together, we just proved that our bound (59) is not worse (up to logarithmic factors) than the one stated in Proposition 4 above. It might even be better if R is large, or if the $\|\varphi_j\|_{L^2}^2$ are much smaller than the $\|\varphi_j\|_\infty^2$, which is the case, e.g., when the φ_j are square integrable but unbounded.

We now discuss the two questions left open in [DT10a]. Despite the similarity of the two bounds, the sparsity oracle inequality stated in Proposition 4 above only holds for vectors \mathbf{u} within ℓ^1 -balls of finite radii. The authors thus asked in [DT10a, Section 4.2] whether it was possible to extend the infimum to the whole \mathbb{R}^d space. Our results show that, thanks to truncation, the answer is positive.

Equivalently, in [DT10a] it is somehow assumed that the forecaster has access to a relatively precise bound R^* on the ℓ^1 -norm of the linear combination \mathbf{u}^* minimizing the regularized risk

$$\|f - \mathbf{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 + \text{pen}_{T,\tau}(\mathbf{u}) ,$$

where $\text{pen}_{T,\tau}(\mathbf{u})$ is of order

$$\text{pen}_{T,\tau}(\mathbf{u}) \sim \frac{\|\mathbf{u}\|_0}{\eta(\|\mathbf{u}\|_1) T} \ln \left(1 + \frac{\|\mathbf{u}\|_1}{\tau \|\mathbf{u}\|_0} \right) + \Delta_T(\eta(\|\mathbf{u}\|_1), \tau) ,$$

where, for a given $R > 0$, $\eta(R)$ is the largest η that satisfies the assumptions of Proposition 4 above. This a priori bound R^* is then used for the choice of the other parameters η and τ . On the contrary, thanks to truncation, we do not need to assume that such a bound R^* is available to the forecaster.

The second open question, which was raised in [DT10a, Remark 7], deals with the adaptivity to the unknown variance factor σ^2 of the noise. The a priori knowledge of the latter is indeed required by their algorithm for the choice of the inverse temperature parameter η . This is not the case for our algorithm SeqSEW, which thus provides adaptivity (up to logarithmic factors) to the unknown variance σ^2 of the noise at least whenever the latter is Gaussian. \square

5.2 Regression model with fixed design

In this section, we consider the regression model with fixed design under the square loss described in Figure 3. In particular, at the beginning of the game, the forecaster has access to the time horizon T and observes the whole sample $(x_1, Y_1), \dots, (x_T, Y_T)$, where

$$Y_t = f(x_t) + \varepsilon_t, \quad 1 \leq t \leq T,$$

the x_t are deterministic elements in \mathcal{X} (picked by the environment), the $\varepsilon_t \in \mathbb{R}$ are i.i.d. random variables with an unknown distribution, and $f : \mathcal{X} \rightarrow \mathbb{R}$ is an unknown function.

In this setting, just like in Section 5.1, our algorithm and the corresponding analysis are a straightforward consequence of the general results on individual sequences developed in Section 4. The resulting sharp sparsity oracle inequality stated in Theorem 3 below is of the same flavour as Theorem 5 of [DT08]. But, as can be seen from Corollary 6 below, our bound is adaptive to the unknown variance σ^2 of the noise at least whenever the latter is Gaussian (weaker bounds are also proved under weaker assumptions). The framework of prediction of individual sequences thus offers a unifying framework to address tuning issues in both the random and the fixed design cases.

Like for the case of random design, the sample $(x_1, Y_1), \dots, (x_T, Y_T)$ is treated in a sequential fashion. We run the algorithm SeqSEW_τ^* defined in Figure 5 from time 1 to time T with the particular choice of $\tau = 1/\sqrt{dT}$. We then define our data-based regressor $\hat{f}_T : \mathcal{X} \mapsto \mathbb{R}$ by

$$\hat{f}_T(x) \triangleq \begin{cases} \frac{1}{n_x} \sum_{\substack{1 \leq t \leq T \\ t: x_t = x}} \tilde{f}_t(x) & \text{if } x \in \{x_1, \dots, x_T\}, \\ 0 & \text{if } x \notin \{x_1, \dots, x_T\}, \end{cases}$$

where

$$n_x \triangleq \left| \{t : x_t = x\} \right| = \sum_{t=1}^T \mathbb{I}_{\{x_t = x\}},$$

and where the regressors $\tilde{f}_t : \mathcal{X} \rightarrow \mathbb{R}$ sequentially built by the algorithm SeqSEW_τ^* are defined by

$$\tilde{f}_t(x) \triangleq \int_{\mathbb{R}^d} [\mathbf{u} \cdot \boldsymbol{\varphi}(x)]_{B_t} p_t(d\mathbf{u}).$$

In the particular case when the x_t are all distinct, \hat{f}_T is simply defined by $\hat{f}_T(x) \triangleq \tilde{f}_T(x)$ if $x \in \{x_1, \dots, x_T\}$ and $\hat{f}_T(x) = 0$ otherwise.

Theorem 3. *Consider the model of fixed design described in Figure 3. Then, the data-based regressor \hat{f}_T defined above satisfies*

$$\begin{aligned} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T (f(x_t) - \hat{f}_T(x_t))^2 \right] &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \frac{1}{T} \sum_{t=1}^T (f(x_t) - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \right. \\ &\quad \left. + 64 \frac{\mathbb{E} [\max_{1 \leq t \leq T} Y_t^2]}{T} \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \frac{1}{dT^2} \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) + 32 \frac{\mathbb{E} [\max_{1 \leq t \leq T} Y_t^2]}{T}. \end{aligned}$$

The previous theorem can be used in the following particular cases.

Corollary 6. *Consider the model of fixed design described in Figure 3. Assume that one of the following assumptions holds on the distribution of ε .*

(BD(B)) : $|\varepsilon| \leq B$ almost surely for a given constant $B > 0$;

(SG(σ^2)) : ε is subgaussian with variance factor $\sigma^2 > 0$, that is, $\mathbb{E}[e^{\lambda\varepsilon}] \leq e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$;

(BEM(α, M)) : ε has a bounded exponential moment, that is, $\mathbb{E}[e^{\alpha|\varepsilon|}] \leq M$ for some given constants $\alpha > 0$ and $M > 0$;

(BM(α, M)) : ε has a bounded moment, that is, $\mathbb{E}[|\varepsilon|^\alpha] \leq M$ for some given constants $\alpha > 2$ and $M > 0$.

Then, the data-based regressor \hat{f}_T defined above satisfies

$$\begin{aligned} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T (f(x_t) - \hat{f}_T(x_t))^2\right] &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \frac{1}{T} \sum_{t=1}^T (f(x_t) - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \right. \\ &\quad \left. + 128 \left(\frac{\max_{1 \leq t \leq T} f^2(x_t)}{T} + \psi_T \right) \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \frac{1}{dT^2} \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) + 64 \left(\frac{\max_{1 \leq t \leq T} f^2(x_t)}{T} + \psi_T \right), \end{aligned}$$

where

$$\psi_T \triangleq \frac{1}{T} \mathbb{E} \left[\max_{1 \leq t \leq T} \varepsilon_t^2 \right] \leq \begin{cases} \frac{B^2}{T} & \text{if Assumption (BD(B)) holds,} \\ \frac{\sigma^2}{T} + 2C_T \sigma^2 \frac{\ln T}{T} & \text{if Assumption (SG(\sigma^2)) holds,} \\ \frac{\ln^2(T(M+e))}{\alpha^2 T} & \text{if Assumption (BEM(\alpha, M)) holds,} \\ \frac{M^{2/\alpha}}{T^{(\alpha-2)/\alpha}} & \text{if Assumption (BM(\alpha, M)) holds,} \end{cases}$$

for some sequence $(C_T)_{T \geq 1}$ of positive real numbers depending only on T and such that

$$\sup_{T \geq 1} C_T \leq 8 \quad \text{and} \quad \lim_{T \rightarrow +\infty} C_T = 1.$$

Proof (of Theorem 3). The proof follows the same lines as the proof of Theorem 2. We thus only sketch the main arguments.

Applying Corollary 2 we have, almost surely,

$$\sum_{t=1}^T (Y_t - \tilde{f}_t(x_t))^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (Y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + 64 \left(\max_{1 \leq t \leq T} Y_t^2 \right) \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\}$$

$$+ \frac{1}{dT} \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) + 32 \max_{1 \leq t \leq T} Y_t^2 .$$

Taking the expectations of both sides, expanding the squares $(Y_t - \tilde{f}_t(x_t))^2$ and $(Y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2$, noting that two terms $T\sigma^2$ cancel out, and then dividing both sides by T , we get

$$\begin{aligned} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T (f(x_t) - \tilde{f}_t(x_t))^2 \right] &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \frac{1}{T} \sum_{t=1}^T (f(x_t) - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \right. \\ &\quad \left. + 64 \frac{\mathbb{E}[\max_{1 \leq t \leq T} Y_t^2]}{T} \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \frac{1}{dT^2} \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) + 32 \frac{\mathbb{E}[\max_{1 \leq t \leq T} Y_t^2]}{T} . \end{aligned}$$

The right-hand side is exactly the upper bound stated in Theorem 3. We thus only need to check that

$$(60) \quad \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T (f(x_t) - \hat{f}_T(x_t))^2 \right] \leq \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T (f(x_t) - \tilde{f}_t(x_t))^2 \right] .$$

This is an equality if the x_t are all distinct. In general we get an inequality which follows from the convexity of the square loss. Indeed, by definition of n_x , we have, almost surely,

$$\begin{aligned} \sum_{t=1}^T (f(x_t) - \hat{f}_T(x_t))^2 &= \sum_{x \in \{x_1, \dots, x_T\}} \sum_{\substack{1 \leq t \leq T \\ t: x_t = x}} (f(x_t) - \hat{f}_T(x_t))^2 = \sum_{x \in \{x_1, \dots, x_T\}} n_x (f(x) - \hat{f}_T(x))^2 \\ &= \sum_{x \in \{x_1, \dots, x_T\}} n_x \left(f(x) - \frac{1}{n_x} \sum_{\substack{1 \leq t \leq T \\ t: x_t = x}} \tilde{f}_t(x) \right)^2 \\ &\leq \sum_{x \in \{x_1, \dots, x_T\}} n_x \frac{1}{n_x} \sum_{\substack{1 \leq t \leq T \\ t: x_t = x}} (f(x) - \tilde{f}_t(x))^2 = \sum_{t=1}^T (f(x_t) - \tilde{f}_t(x_t))^2 , \end{aligned}$$

where the second line is by definition of \hat{f}_T and where the last line follows from Jensen's inequality. Dividing both sides by T and taking their expectations, we get (60), which concludes the proof. \square

Proof (of Corollary 6). First note that

$$\begin{aligned} \mathbb{E} \left[\max_{1 \leq t \leq T} Y_t^2 \right] &\triangleq \mathbb{E} \left[\max_{1 \leq t \leq T} (f(x_t) + \varepsilon_t)^2 \right] \\ &\leq 2 \left(\max_{1 \leq t \leq T} f^2(x_t) + \mathbb{E} \left[\max_{1 \leq t \leq T} \varepsilon_t^2 \right] \right) . \end{aligned}$$

The proof then follows the exact same lines as for Corollary 5 with the sequence (ε_t) instead of the sequence $((\Delta Y)_t)$. \square

6 Appendix

6.1 A convex duality argument for the Kullback-Leibler divergence

In this section, we recall a convex duality inequality satisfied by the Kullback-Leibler divergence and whose proof can be found, e.g., in [DZ98, p. 264] or [Cat04, p. 159]. For the sake of simplicity, and because it is sufficient for our applications, we only state it for *non-positive* measurable functions (whose expectation with respect to any probability distribution is always well defined, be it finite or not).

Proposition 5. *For any measurable space (E, \mathcal{B}) , any probability distribution π on (E, \mathcal{B}) , and any non-positive measurable function $h : E \rightarrow \mathbb{R}_-$, the Legendre transform of the Kullback-Leibler divergence can be expressed as*

$$\ln \int_E e^h d\pi = \sup_{\rho \in \mathcal{M}_1^+(E)} \left\{ \int_E h d\rho - \mathcal{K}(\rho, \pi) \right\}.$$

6.2 Exp-concavity of the square loss

In this section, we recall the notion of exp-concavity and the (elementary) fact that the square loss is $1/(8B^2)$ -exp-concave on $[-B, B]$. See, e.g., [KW99] or [CBL06] for a reference on exp-concave losses.

Definition 1. *Let $I \subset \mathbb{R}$ be an interval of the real line. A function $h : I \rightarrow \mathbb{R}$ is said to be exp-concave for a given $\eta > 0$ (or simply η -exp-concave) if the function $H_\eta \triangleq e^{-\eta h}$ is concave on I .*

Noting that $H_{\eta'} = H_\eta^{\eta'/\eta}$ and $h = -\frac{1}{\eta} \ln H_\eta$, we can see that if $h : I \rightarrow \mathbb{R}$ is η -exp-concave, then

- h is η' -exp-concave for all $0 < \eta' \leq \eta$ (since $x \mapsto x^{\eta'/\eta}$ is concave and non-decreasing);
- h is convex (since $x \mapsto -\frac{1}{\eta} \ln x$ is convex and non-increasing).

A proof of the next proposition can be found, e.g., in [Vov01].

Proposition 6. *The square loss is $1/(8B^2)$ -exp-concave on $[-B, B]$ in the sense that, for all $y \in [-B, B]$, the function*

$$x \in [-B, B] \mapsto (y - x)^2$$

is $1/(8B^2)$ -exp-concave.

(moreover, the constant $1/(8B^2)$ is not improvable)

6.3 Some tools to exploit our PAC-Bayesian inequalities

In this section, we recall two results needed for the derivation of Proposition 2 and Proposition 3 from the PAC-Bayesian inequalities (15) and (31). The proofs are due to [DT07, DT08] and we

only reproduce² them for the convenience of the reader.

For any $\mathbf{u}^* \in \mathbb{R}^d$ and $\tau > 0$, define $\rho_{\mathbf{u}^*, \tau}$ as the translated of π_τ at \mathbf{u}^* , namely,

$$(61) \quad \rho_{\mathbf{u}^*, \tau} \triangleq \frac{d\pi_\tau}{d\mathbf{u}}(\mathbf{u} - \mathbf{u}^*) d\mathbf{u} = \prod_{j=1}^d \frac{(3/\tau) du_j}{2(1 + |u_j - u_j^*|/\tau)^4}.$$

Lemma 3. For all $\mathbf{u}^* \in \mathbb{R}^d$ and $\tau > 0$, the probability distribution $\rho_{\mathbf{u}^*, \tau}$ satisfies

$$\int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \rho_{\mathbf{u}^*, \tau}(d\mathbf{u}) = \sum_{t=1}^T (y_t - \mathbf{u}^* \cdot \boldsymbol{\varphi}(x_t))^2 + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t).$$

Lemma 4. For all $\mathbf{u}^* \in \mathbb{R}^d$ and $\tau > 0$, the probability distribution $\rho_{\mathbf{u}^*, \tau}$ satisfies

$$\mathcal{K}(\rho_{\mathbf{u}^*, \tau}, \pi_\tau) \leq 4 \|\mathbf{u}^*\|_0 \ln \left(1 + \frac{\|\mathbf{u}^*\|_1}{\|\mathbf{u}^*\|_0 \tau} \right).$$

Proof (of Lemma 3). For all $t \in \{1, \dots, T\}$ we expand the square $(y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 = (y_t - \mathbf{u}^* \cdot \boldsymbol{\varphi}(x_t) + (\mathbf{u}^* - \mathbf{u}) \cdot \boldsymbol{\varphi}(x_t))^2$ and use the linearity of the integral to get

$$(62) \quad \begin{aligned} \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \rho_{\mathbf{u}^*, \tau}(d\mathbf{u}) \\ = \sum_{t=1}^T (y_t - \mathbf{u}^* \cdot \boldsymbol{\varphi}(x_t))^2 + \sum_{t=1}^T \int_{\mathbb{R}^d} ((\mathbf{u}^* - \mathbf{u}) \cdot \boldsymbol{\varphi}(x_t))^2 \rho_{\mathbf{u}^*, \tau}(d\mathbf{u}) \\ + \underbrace{\sum_{t=1}^T 2(y_t - \mathbf{u}^* \cdot \boldsymbol{\varphi}(x_t)) \int_{\mathbb{R}^d} (\mathbf{u}^* - \mathbf{u}) \cdot \boldsymbol{\varphi}(x_t) \rho_{\mathbf{u}^*, \tau}(d\mathbf{u})}_{=0} \end{aligned}$$

The last sum is equal to zero by symmetry of $\rho_{\mathbf{u}^*, \tau}$ around \mathbf{u}^* , which entails that $\int_{\mathbb{R}^d} \mathbf{u} \rho_{\mathbf{u}^*, \tau}(d\mathbf{u}) = \mathbf{u}^*$.

As for the second sum of the right-hand side, it can be bounded from above similarly. Indeed, expanding the inner product and then the square $((\mathbf{u}^* - \mathbf{u}) \cdot \boldsymbol{\varphi}(x_t))^2$ we have, for all $t = 1, \dots, T$,

$$((\mathbf{u}^* - \mathbf{u}) \cdot \boldsymbol{\varphi}(x_t))^2 = \sum_{j=1}^d (u_j^* - u_j)^2 \varphi_j^2(x_t) + \sum_{1 \leq j \neq k \leq d} (u_j^* - u_j)(u_k^* - u_k) \varphi_j(x_t) \varphi_k(x_t).$$

By symmetry of $\rho_{\mathbf{u}^*, \tau}$ around \mathbf{u}^* and the fact that $\rho_{\mathbf{u}^*, \tau}$ is a product-distribution, we get

$$\sum_{t=1}^T \int_{\mathbb{R}^d} ((\mathbf{u}^* - \mathbf{u}) \cdot \boldsymbol{\varphi}(x_t))^2 \rho_{\mathbf{u}^*, \tau}(d\mathbf{u}) = \sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t) \int_{\mathbb{R}^d} (u_j^* - u_j)^2 \rho_{\mathbf{u}^*, \tau}(d\mathbf{u}) + 0$$

²The notations are however slightly modified because of the change in the statistical setting and goal. The target predictions $(f(x_1), \dots, f(x_T))$ are indeed replaced with the observations (y_1, \dots, y_T) and the prediction loss $\|f - f_{\mathbf{u}}\|_n^2$ is replaced with the cumulative loss $\sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2$. Moreover, the analysis of the present proof is slightly simpler since we just need to consider the case $L_0 = +\infty$ according to the notations of Theorem 5 in [DT08].

$$(63) \quad = \sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t) \int_{\mathbb{R}} (u_j^* - u_j)^2 \frac{(3/\tau) du_j}{2(1 + |u_j - u_j^*|/\tau)^4}$$

$$(64) \quad = \tau^2 \sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t) \int_{\mathbb{R}} \frac{3t^2 dt}{2(1 + |t|)^4}$$

$$(65) \quad = \tau^2 \sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t) .$$

Equation (63) follows from the definition of $\rho_{\mathbf{u}^*, \tau}$. Equation (64) is obtained by the change of variables $t = (u_j - u_j^*)/\tau$. As for Equation (65), it follows from the equality $\int_{\mathbb{R}} \frac{3t^2 dt}{2(1 + |t|)^4} = 1$ that can be proved by integrating by parts.

Combined with (65), Equation (62) yields the desired equality. \square

Proof (of Lemma 4). By definition of $\rho_{\mathbf{u}^*, \tau}$ and π_τ , we have

$$(66) \quad \begin{aligned} \mathcal{K}(\rho_{\mathbf{u}^*, \tau}, \pi_\tau) &\triangleq \int_{\mathbb{R}^d} \left(\ln \frac{d\rho_{\mathbf{u}^*, \tau}(\mathbf{u})}{d\pi_\tau} \right) \rho_{\mathbf{u}^*, \tau}(\mathbf{d}\mathbf{u}) = \int_{\mathbb{R}^d} \left(\ln \prod_{j=1}^d \frac{(1 + |u_j|/\tau)^4}{(1 + |u_j - u_j^*|/\tau)^4} \right) \rho_{\mathbf{u}^*, \tau}(\mathbf{d}\mathbf{u}) \\ &= 4 \int_{\mathbb{R}^d} \left(\sum_{j=1}^d \ln \frac{1 + |u_j|/\tau}{1 + |u_j - u_j^*|/\tau} \right) \rho_{\mathbf{u}^*, \tau}(\mathbf{d}\mathbf{u}) . \end{aligned}$$

But, for all $\mathbf{u} \in \mathbb{R}^d$, by the triangle inequality,

$$1 + |u_j|/\tau \leq 1 + |u_j^*|/\tau + |u_j - u_j^*|/\tau \leq (1 + |u_j^*|/\tau)(1 + |u_j - u_j^*|/\tau) ,$$

so that Equation (66) yields the upper bound

$$\mathcal{K}(\rho_{\mathbf{u}^*, \tau}, \pi_\tau) \leq 4 \sum_{j=1}^d \ln(1 + |u_j^*|/\tau) = 4 \sum_{j: u_j^* \neq 0} \ln(1 + |u_j^*|/\tau) .$$

We now recall that $\|\mathbf{u}^*\|_0 \triangleq |\{j : u_j^* \neq 0\}|$ and apply Jensen's inequality to the concave function $x \in (-1, +\infty) \mapsto \ln(1 + x)$ to get

$$\begin{aligned} \sum_{j: u_j^* \neq 0} \ln(1 + |u_j^*|/\tau) &= \|\mathbf{u}^*\|_0 \frac{1}{\|\mathbf{u}^*\|_0} \sum_{j: u_j^* \neq 0} \ln(1 + |u_j^*|/\tau) \leq \|\mathbf{u}^*\|_0 \ln \left(1 + \frac{\sum_{j: u_j^* \neq 0} |u_j^*|}{\|\mathbf{u}^*\|_0 \tau} \right) \\ &\leq \|\mathbf{u}^*\|_0 \ln \left(1 + \frac{\|\mathbf{u}^*\|_1}{\|\mathbf{u}^*\|_0 \tau} \right) . \end{aligned}$$

This concludes the proof. \square

6.4 Some maximal inequalities

In this section, we prove three maximal inequalities needed for the derivation of Corollaries 5 and 6 from Theorems 2 and 3 respectively. All these inequalities follow from a Pisier-type argument. Their proofs are quite standard but we provide them for the convenience of the reader.

Lemma 5. *Let Z_1, \dots, Z_T be $T \geq 1$ (centered) real random variables such that, for a given constant $\nu > 0$, we have*

$$\forall t \in \{1, \dots, T\}, \quad \forall \lambda \in \mathbb{R}, \quad \mathbb{E}\left[e^{\lambda Z_t}\right] \leq e^{\lambda^2 \nu / 2}.$$

Then, for all $0 < \beta < \ln 2$,

$$\mathbb{E}\left[\max_{1 \leq t \leq T} Z_t^2\right] \leq \nu + \gamma_T(\beta) 2\nu \ln T,$$

where

$$\gamma_T(\beta) \triangleq \begin{cases} 1 & \text{if } T = 1, \\ \frac{1}{1 - \beta/\sqrt{\ln T}} + \frac{2(1 - \beta/\sqrt{\ln T})}{\beta\sqrt{\ln T}} & \text{if } T \geq 2, \end{cases}$$

so that for the particular choice of $\beta = \sqrt{\ln 2}/2$ we have

$$\sup_{T \geq 1} \gamma_T(\sqrt{\ln 2}/2) \leq 8, \quad \text{and} \quad \lim_{T \rightarrow +\infty} \gamma_T(\sqrt{\ln 2}/2) = 1.$$

Lemma 6. *Let Z_1, \dots, Z_T be $T \geq 1$ real random variables such that, for some given constants $\alpha > 0$ and $M > 0$, we have*

$$\forall t \in \{1, \dots, T\}, \quad \mathbb{E}\left[e^{\alpha |Z_t|}\right] \leq M.$$

Then,

$$\mathbb{E}\left[\max_{1 \leq t \leq T} Z_t^2\right] \leq \frac{\ln^2((M + e)T)}{\alpha^2}.$$

Lemma 7. *Let Z_1, \dots, Z_T be $T \geq 1$ real random variables such that, for some given constants $\alpha > 2$ and $M > 0$, we have*

$$\forall t \in \{1, \dots, T\}, \quad \mathbb{E}[|Z_t|^\alpha] \leq M.$$

Then,

$$\mathbb{E}\left[\max_{1 \leq t \leq T} Z_t^2\right] \leq (MT)^{2/\alpha}.$$

Proof (of Lemma 5). First note that, if $T = 1$, then

$$(67) \quad \mathbb{E} \left[\max_{1 \leq t \leq T} Z_t^2 \right] = \mathbb{E} [Z_1^2] \leq \nu,$$

since Z_1 is subgaussian with variance factor ν (this upper bound is well-known and can be obtained via a second-order Taylor expansion of $\ln \mathbb{E}[e^{\lambda Z_1}] = \lambda^2 \mathbb{E}[Z_1^2]/2 + o(\lambda^2)$ as $\lambda \rightarrow 0$, see, e.g., [BLM]). Since in this case, we also have $\ln T = 1$, this proves the desired inequality.

We can now assume that $T \geq 2$. The rest of the proof is based on a Pisier-type argument as is done, e.g., in Lemma 2.3 of [Mas07] to prove, under the same assumptions as in the present Lemma, the maximal inequality

$$\mathbb{E} \left[\max_{1 \leq t \leq T} Z_t \right] \leq \sqrt{2\nu \ln T}.$$

Fix $\lambda > 0$, whose value will be chosen by the analysis. First note that

$$\mathbb{E} \left[\max_{1 \leq t \leq T} Z_t^2 \right] = \frac{1}{\lambda} \ln \exp \left(\lambda \mathbb{E} \left[\max_{1 \leq t \leq T} Z_t^2 \right] \right) \leq \frac{1}{\lambda} \ln \mathbb{E} \left[\exp \left(\lambda \max_{1 \leq t \leq T} Z_t^2 \right) \right]$$

by Jensen's inequality. Since $x \mapsto e^{\lambda x}$ is nondecreasing and $e^{\lambda Z_t^2} \geq 0$ for all $t = 1, \dots, T$, the last inequality yields

$$(68) \quad \begin{aligned} \mathbb{E} \left[\max_{1 \leq t \leq T} Z_t^2 \right] &\leq \frac{1}{\lambda} \ln \mathbb{E} \left[\max_{1 \leq t \leq T} e^{\lambda Z_t^2} \right] \leq \frac{1}{\lambda} \ln \mathbb{E} \left[\sum_{t=1}^T e^{\lambda Z_t^2} \right] \\ &\leq \frac{1}{\lambda} \ln \left(\sum_{t=1}^T e^{\lambda \mathbb{E}[Z_t^2]} \mathbb{E} \left[e^{\lambda (Z_t^2 - \mathbb{E}[Z_t^2])} \right] \right) \\ &\leq \nu + \frac{1}{\lambda} \ln \left(\sum_{t=1}^T \mathbb{E} \left[e^{\lambda (Z_t^2 - \mathbb{E}[Z_t^2])} \right] \right), \end{aligned}$$

where the last inequality follows from $\mathbb{E}[Z_t^2] \leq \nu$ for all $t = 1, \dots, T$ proved in (67).

Let $t \in \{1, \dots, T\}$. The rest of the proof is dedicated to bounding $\mathbb{E} \left[e^{\lambda (Z_t^2 - \mathbb{E}[Z_t^2])} \right]$ from above by means of Bernstein's inequality.

Since Z_t is subgaussian with variance factor ν , we have, from Theorem 1.2 of [BLM],

$$\mathbb{E} \left[(Z_t^2)^2 \right] \leq 2 \cdot 2! (2\nu)^2 = 16\nu^2$$

and

$$\forall k \geq 3, \quad \mathbb{E} \left[(Z_t^2)_+^k \right] \leq \mathbb{E} \left[Z_t^{2k} \right] \leq 2 k! (2\nu)^k = \frac{k!}{2} (16\nu^2) (2\nu)^{k-2}.$$

We can thus apply the version of Bernstein's inequality stated in Proposition 2.9 of [Mas07] or, e.g., in Theorem 1.16 of [BLM] to get

$$\forall \lambda \in (0, 1/(2\nu)), \quad \mathbb{E} \left[e^{\lambda (Z_t^2 - \mathbb{E}[Z_t^2])} \right] \leq \exp \left(\frac{\lambda^2 (16\nu^2)}{2(1 - (2\nu)\lambda)} \right).$$

Equation (68) thus yields, for all $\lambda \in (0, 1/(2\nu))$,

$$\begin{aligned} \mathbb{E} \left[\max_{1 \leq t \leq T} Z_t^2 \right] &\leq \nu + \frac{\ln T}{\lambda} + \frac{1}{\lambda} \ln \exp \left(\frac{\lambda^2 (16\nu^2)}{2(1 - (2\nu)\lambda)} \right) \\ &\leq \nu + \frac{\ln T}{\lambda} + \frac{8\lambda\nu^2}{1 - 2\nu\lambda}. \end{aligned}$$

In particular, choosing $\lambda \triangleq \frac{1}{2\nu} \left(1 - \frac{\beta}{\sqrt{\ln T}} \right)$ which can be seen to approximately minimize the last upper bound, we get

$$\begin{aligned} \mathbb{E} \left[\max_{1 \leq t \leq T} Z_t^2 \right] &\leq \nu + \frac{2\nu \ln T}{1 - \frac{\beta}{\sqrt{\ln T}}} + \frac{4\nu\sqrt{\ln T}}{\beta \left(1 - \frac{\beta}{\sqrt{\ln T}} \right)} \\ &\leq \nu + \underbrace{\left(\frac{1}{1 - \beta/\sqrt{\ln T}} + \frac{2(1 - \beta/\sqrt{\ln T})}{\beta\sqrt{\ln T}} \right)}_{\triangleq \gamma_T(\beta)} 2\nu \ln T. \end{aligned}$$

This concludes the proof of the maximal inequality. The fact that

$$\sup_{T \geq 1} \gamma_T(\sqrt{\ln 2/2}) \leq 2 + \frac{4}{\ln 2} \leq 8 \quad \text{and} \quad \lim_{T \rightarrow +\infty} \gamma_T(\sqrt{\ln 2/2}) = 1$$

follows from a direct calculation. □

Proof (of Lemma 6). We first need the following definitions. Let $\psi_\alpha : \mathbb{R}_+ \rightarrow \mathbb{R}$ be a convex majorant of $x \mapsto e^{\alpha\sqrt{x}}$ on \mathbb{R}_+ defined by

$$\psi_\alpha(x) \triangleq \begin{cases} e & \text{if } x < 1/\alpha^2, \\ e^{\alpha\sqrt{x}} & \text{if } x \geq 1/\alpha^2. \end{cases}$$

We associate with ψ_α its generalized inverse $\psi_\alpha^{-1} : \mathbb{R} \rightarrow \mathbb{R}_+$ defined by

$$\psi_\alpha^{-1}(y) = \begin{cases} 1/\alpha^2 & \text{if } y < e, \\ (\ln y)^2/\alpha^2 & \text{if } y \geq e. \end{cases}$$

Elementary manipulations show that

- ψ_α is nondecreasing and convex on \mathbb{R}_+ ;
- ψ_α^{-1} is nondecreasing on \mathbb{R} ;
- $x \leq \psi_\alpha^{-1}(\psi_\alpha(x))$ for all $x \in \mathbb{R}_+$.

We can now use the same Pisier-type argument as in the previous proof. From the inequality $x \leq \psi_\alpha^{-1}(\psi_\alpha(x))$ for all $x \in \mathbb{R}_+$ we have

$$\mathbb{E} \left[\max_{1 \leq t \leq T} Z_t^2 \right] \leq \psi_\alpha^{-1} \left(\psi_\alpha \left(\mathbb{E} \left[\max_{1 \leq t \leq T} Z_t^2 \right] \right) \right)$$

$$\leq \psi_\alpha^{-1} \left(\mathbb{E} \left[\psi_\alpha \left(\max_{1 \leq t \leq T} Z_t^2 \right) \right] \right) = \psi_\alpha^{-1} \left(\mathbb{E} \left[\max_{1 \leq t \leq T} \psi_\alpha(Z_t^2) \right] \right),$$

where the last two inequalities follow by Jensen's inequality (since ψ_α is convex) and the fact that both ψ_α^{-1} and ψ_α are nondecreasing.

Since $\psi_\alpha \geq 0$ and ψ_α^{-1} is nondecreasing we get

$$\begin{aligned} \mathbb{E} \left[\max_{1 \leq t \leq T} Z_t^2 \right] &\leq \psi_\alpha^{-1} \left(\mathbb{E} \left[\sum_{t=1}^T \psi_\alpha(Z_t^2) \right] \right) = \psi_\alpha^{-1} \left(\sum_{t=1}^T \mathbb{E} [\psi_\alpha(Z_t^2)] \right) \\ &\leq \psi_\alpha^{-1} \left(\sum_{t=1}^T \mathbb{E} [e^{\alpha|Z_t|} + e] \right) \\ &\leq \psi_\alpha^{-1} (MT + eT) = \frac{\ln^2(MT + eT)}{\alpha^2}, \end{aligned}$$

where the second line follows from the inequality $\psi_\alpha(x) \leq e + e^{\alpha\sqrt{x}}$ for all $x \in \mathbb{R}_+$, and where the last line follows from the bounded exponential moment assumption and the definition of ψ_α^{-1} . It concludes the proof. \square

Proof (of Lemma 7). As in the previous proofs, we have, by Jensen's inequality and the fact that $x \mapsto x^{\alpha/2}$ is convex and nondecreasing on \mathbb{R}_+ (since $\alpha > 2$),

$$\begin{aligned} \mathbb{E} \left[\max_{1 \leq t \leq T} Z_t^2 \right] &\leq \mathbb{E} \left[\left(\max_{1 \leq t \leq T} Z_t^2 \right)^{\alpha/2} \right]^{2/\alpha} = \mathbb{E} \left[\max_{1 \leq t \leq T} |Z_t|^\alpha \right]^{2/\alpha} \\ &\leq \mathbb{E} \left[\sum_{t=1}^T |Z_t|^\alpha \right]^{2/\alpha} \leq (MT)^{2/\alpha} \end{aligned}$$

by the bounded moment assumption, which concludes the proof. \square

7 Acknowledgments

The author would like to thank Arnak Dalalyan, Gilles Stoltz, and Pascal Massart for their helpful comments and suggestions. The author acknowledges the support of the French Agence Nationale de la Recherche (ANR), under grant PARCIMONIE (<http://www.proba.jussieu.fr/ANR/Parcimonie>), and of the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886.

References

- [AC09] J.-Y. Audibert and O. Catoni. Risk bounds in linear regression through PAC-Bayesian truncation. 2009. Preprint.
- [ACBG02] P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *J. Comp. Sys. Sci.*, 64:48–75, 2002.
- [Aka71] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281, 1971.

-
- [AL10] P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. Technical report, 2010. See <http://arxiv.org/abs/1009.2707>.
 - [Aud04] J. Y. Audibert. *PAC-Bayesian Statistical Learning Theory*. PhD thesis, Université Paris VI, 2004.
 - [AW01] K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, 43(3):211–246, 2001.
 - [BBGO10] G. Biau, K. Bleakley, L. Györfi, and G. Ottucsák. Nonparametric sequential prediction of time series. *J. Nonparametr. Stat.*, 22(3–4):297–317, 2010.
 - [BLM] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities with applications*. Book in progress.
 - [BN08] F. Bunea and A. Nobel. Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Trans. Inform. Theory*, 54(4):1725–1735, 2008.
 - [BRT09] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.
 - [BTW04] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for regression learning. Technical report, 2004. See <http://arxiv.org/abs/math/0410214>.
 - [BTW06] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation and sparsity via ℓ_1 penalized least squares. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT'06)*, pages 379–391, 2006.
 - [BTW07a] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.
 - [BTW07b] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194, 2007.
 - [Cat99] O. Catoni. Universal aggregation rules with exact bias bounds. Technical Report PMA-510, Laboratoire de Probabilités et Modèles Aléatoires, CNRS, Paris, 1999.
 - [Cat04] O. Catoni. Statistical learning theory and stochastic optimization. In *Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI - 2001*, volume 1851 of *Lecture Notes in Mathematics*, pages 1–269, New York, 2004. Springer.
 - [CBCG04] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Trans. Inform. Theory*, 50(9):2050–2057, 2004.
 - [CBG08] N. Cesa-Bianchi and C. Gentile. Improved risk tail bounds for on-line algorithms. *IEEE Trans. Inform. Theory*, 54(1):386–390, 2008.
 - [CBL06] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
 - [CBMS07] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66(2/3):321–352, 2007.
 - [CML07] P.A. Cornillon and E. Matzner-Løber. *Régression: théorie et applications*. Statistiques et probabilités appliquées. Springer, Paris, 2007.
 - [DJ94] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.
 - [DSSST10] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT'10)*, pages 14–26, 2010.
 - [DT07] A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT'07)*, pages 97–111, 2007.

- [DT08] A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2):39–61, 2008.
- [DT09] A. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT'09)*, pages 83–92, 2009.
- [DT10a] A. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 2010. Submitted.
- [DT10b] A. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 2010. Submitted.
- [DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*. Springer, New York, 1998.
- [FG94] D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. *Ann. Statist.*, 22(4):1947–1975, 1994.
- [FSSW97] Y. Freund, R. E. Schapire, Y. Singer, and M. K. Warmuth. Using and combining predictors that specialize. In *Proceedings of the 29th annual ACM Symposium on Theory of Computing (STOC'97)*, pages 334–343, 1997.
- [Gir08] C. Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14(4):1089–1107, 2008.
- [GKKW02] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.
- [GO07] L. Györfi and G. Ottucsák. Sequential prediction of unbounded stationary time series. *IEEE Trans. Inform. Theory*, 53(5):1866–1872, 2007.
- [HK70] A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- [JRT08] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5):2183–2206, 2008.
- [Kol09a] V. Koltchinskii. Sparse recovery in convex hulls via entropy penalization. *Ann. Statist.*, 37(3):1332–1359, 2009.
- [Kol09b] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(1):7–57, 2009.
- [KW99] J. Kivinen and M. K. Warmuth. Averaging expert predictions. In *Proceedings of the 4th European Conference on Computational Learning Theory (EuroCOLT'99)*, pages 153–167, 1999.
- [LB06] G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.
- [LLZ09] J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *J. Mach. Learn. Res.*, 10:777–801, 2009.
- [LW94] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Inform. and Comput.*, 108:212–261, 1994.
- [Mal73] C. L. Mallows. Some Comments on Cp. *Technometrics*, 15(4):661–675, 1973.
- [Mas07] P. Massart. *Concentration Inequalities and Model Selection*. Springer, Berlin/Heidelberg/New York, 2007.
- [MM10] P. Massart and C. Meynet. An ℓ^1 -oracle inequality for the Lasso. Technical Report RR-7356, INRIA, 2010. See <http://arxiv.org/abs/1007.4791>.
- [Nem00] A. Nemirovski. *Topics in Non-Parametric Statistics*. Springer, Berlin/Heidelberg/New York, 2000.

-
- [RT10] P. Rigollet and A. B. Tsybakov. Exponential screening and optimal rates of sparse estimation. Technical report, 2010. See <http://arxiv.org/abs/1003.2654>.
 - [RWY09] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of convergence for high-dimensional regression under ℓ^q -ball sparsity. In *Proceedings of the 47th annual Allerton conference on communication, control, and computing (Allerton'09)*, pages 251–257, 2009.
 - [Sch78] G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6:461–464, 1978.
 - [SKF02] A. C. Singer, S. S. Kozat, and M. Feder. Universal linear least squares prediction: upper and lower bounds. *IEEE Trans. Inform. Theory*, 48(8):2354–2362, 2002.
 - [SST09] S. Shalev-Shwartz and A. Tewari. Stochastic methods for ℓ^1 -regularized loss minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, pages 929–936, 2009.
 - [Tib96] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.
 - [Tsy03] A. B. Tsybakov. Optimal rates of aggregation. In *Proceedings of the 16th Annual Conference on Learning Theory (COLT'03)*, pages 303–313, 2003.
 - [vdG08] S. A. van de Geer. High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, 36(2):614–645, 2008.
 - [vdGB09] S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009.
 - [Ver10] N. Verzelen. Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. Technical report, 2010. See <http://arxiv.org/abs/1008.0526>.
 - [Vov90] V. Vovk. Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory (COLT'90)*, pages 371–383, 1990.
 - [Vov01] V. Vovk. Competitive on-line statistics. *Internat. Statist. Rev.*, 69:213–248, 2001.
 - [Xia10] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, 2010.
 - [Yan01] Y. Yang. Adaptive regression by mixing. *J. Amer. Statist. Assoc.*, 96(454):574–588, 2001.