Sparsity regret bounds for individual sequences in online linear regression

Sébastien Gerchinovitz

Journal of Machine Learning Research 14 (2013) 729-769 Submitted 3/12; Revised 11/12; Published 3/13

Sparsity Regret Bounds for Individual Sequences in Online Linear Regression*

Sébastien Gerchinovitz SEBASTIEN.GERCHINOVITZ@ENS.FR
École Normale Supérieure†
45 rue d'Ulm
Paris, FRANCE

Editor: Nicolò Cesa-Bianchi

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. We introduce the notion of *sparsity regret bound*, which is a deterministic online counterpart of recent risk bounds derived in the stochastic setting under a sparsity scenario. We prove such regret bounds for an online-learning algorithm called SeqSEW and based on exponential weighting and data-driven truncation. In a second part we apply a parameter-free version of this algorithm to the stochastic setting (regression model with random design). This yields risk bounds of the same flavor as in Dalalyan and Tsybakov (2012a) but which solve two questions left open therein. In particular our risk bounds are adaptive (up to a logarithmic factor) to the unknown variance of the noise if the latter is Gaussian. We also address the regression model with fixed design.

Keywords: sparsity, online linear regression, individual sequences, adaptive regret bounds

1. Introduction

Sparsity has been extensively studied in the stochastic setting over the past decade. This notion is key to address statistical problems that are high-dimensional, that is, where the number of unknown parameters is of the same order or even much larger than the number of observations. This is the case in many contemporary applications such as computational biology (e.g., analysis of DNA sequences), collaborative filtering (e.g., Netflix, Amazon), satellite and hyperspectral imaging, and high-dimensional econometrics (e.g., cross-country growth regression problems).

A key message about sparsity is that, although high-dimensional statistical inference is impossible in general (i.e., without further assumptions), it becomes statistically feasible if among the many unknown parameters, only few of them are non-zero. Such a situation is called a *sparsity scenario* and has been the focus of many theoretical, computational, and practical works over the past decade in the stochastic setting. On the theoretical side, most sparsity-related risk bounds take the form of the so-called *sparsity oracle inequalities*, that is, risk bounds expressed in terms of the number of non-zero coordinates of the oracle vector. As of now, such theoretical guarantees have only been proved under stochastic assumptions.[1]

*. A shorter version appeared in the proceedings of COLT 2011 (see Gerchinovitz 2011).

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

1. One could object that most high-probability risk bounds derived for ℓ^1-regularization methods are in fact deterministic inequalities that hold true whenever the noise vector ε belong to some set S (see, e.g., Bickel et al. 2009). However,

In this paper we address the prediction possibilities under a sparsity scenario in both deterministic and stochastic settings. We first prove that theoretical guarantees similar to sparsity oracle inequalities can be obtained in a deterministic online setting, namely, online linear regression on individual sequences. The newly obtained deterministic prediction guarantees are called *sparsity regret bounds*. We prove such bounds for an online-learning algorithm which, in its most sophisticated version, is fully automatic in the sense that no preliminary knowledge is needed for the choice of its tuning parameters. In the second part of this paper, we apply our sparsity regret bounds—of deterministic nature—to the stochastic setting (regression model with random design). One of our key results is that, thanks to our online tuning techniques, these deterministic bounds imply sparsity oracle inequalities that are adaptive to the unknown variance of the noise (up to logarithmic factors) when the latter is Gaussian. In particular, this solves an open question raised by Dalalyan and Tsybakov (2012a).

In the next paragraphs, we introduce our main setting and motivate the notion of sparsity regret bound from an online-learning viewpoint. We then detail our main contributions with respect to the statistical literature and the machine-learning literature.

1.1 Introduction of a Deterministic Counterpart of Sparsity Oracle Inequalities

We consider the problem of online linear regression on arbitrary deterministic sequences. A forecaster has to predict in a sequential fashion the values $y_t \in \mathbb{R}$ of an unknown sequence of observations given some input data $x_t \in \mathcal{X}$ and some base forecasters $\varphi_j : \mathcal{X} \to \mathbb{R}$, $1 \leqslant j \leqslant d$, on the basis of which he outputs a prediction $\widehat{y}_t \in \mathbb{R}$. The quality of the predictions is assessed by the square loss. The goal of the forecaster is to predict almost as well as the best linear forecaster $u \cdot \varphi \triangleq \sum_{j=1}^{d} u_j \varphi_j$, where $u \in \mathbb{R}^d$, that is, to satisfy, uniformly over all individual sequences $(x_t, y_t)_{1 \leqslant t \leqslant T}$, a regret bound of the form

$$\sum_{t=1}^{T} \left(y_t - \widehat{y}_t\right)^2 \leqslant \inf_{u \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \left(y_t - u \cdot \varphi(x_t)\right)^2 + \Delta_{T,d}(u) \right\}$$

for some regret term $\Delta_{T,d}(u)$ that should be as small as possible and, in particular, sublinear in T. (For the sake of introduction, we omit the dependencies of $\Delta_{T,d}(u)$ on the amplitudes $\max_{1 \leqslant t \leqslant T} |y_t|$ and $\max_{1 \leqslant t \leqslant T} \max_{1 \leqslant j \leqslant d} |\varphi_j(x_t)|$.)

In this setting the version of the sequential ridge regression forecaster studied by Azoury and Warmuth (2001) and Vovk (2001) can be tuned to have a regret $\Delta_{T,d}(u)$ of order at most $d \ln\left(T \|u\|_2^2\right)$. When the ambient dimension d is much larger than the number of time rounds T, the latter regret bound may unfortunately be larger than T and is thus somehow trivial. Since the regret bound $d \ln T$ is optimal in a certain sense (see, e.g., the lower bound of Vovk 2001, Theorem 2), additional assumptions are needed to get interesting theoretical guarantees.

A natural assumption, which has already been extensively studied in the stochastic setting, is that there is a sparse vector u^* (i.e., with $s \ll T/(\ln T)$ non-zero coefficients) such that the linear combination $u^* \cdot \varphi$ has a small cumulative square loss. If the forecaster knew in advance the support $J(u^*) \triangleq \{j : u_j^* \neq 0\}$ of u^*, he could apply the same forecaster as above but only to the s-dimensional linear subspace $\left\{u \in \mathbb{R}^d : \forall j \notin J(u^*), u_j = 0\right\}$. The regret bound of this "oracle" would be roughly of order $s \ln T$ and thus sublinear in T. Under this sparsity scenario, a sublinear regret thus seems

the fact that $\varepsilon \in S$ with high-probability is only guaranteed via concentration arguments, so it is a consequence of the underlying statistical assumptions.

possible, though, of course, the aforementioned regret bound $s \ln T$ can only be used as an ideal benchmark (since the support of u^* is unknown).

In this paper, we prove that a regret bound proportional to s is achievable (up to logarithmic factors). In Corollary 2 and its refinements (Corollary 7 and Theorem 10), we indeed derive regret bounds of the form

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{u \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \big(y_t - u \cdot \varphi(x_t)\big)^2 + \big(\|u\|_0 + 1\big)\, g_{T,d}\big(\|u\|_1, \|\varphi\|_\infty\big) \right\}, \tag{1}$$

where $\|u\|_0$ denotes the number of non-zero coordinates of u and where g grows at most logarithmically in T, d, $\|u\|_1 \triangleq \sum_{j=1}^{d} |u_j|$, and $\|\varphi\|_\infty \triangleq \sup_{x \in \mathcal{X}} \max_{1 \leqslant j \leqslant d} |\varphi_j(x)|$. We call regret bounds of the above form *sparsity regret bounds*.

This work is in connection with several papers that belong either to the statistical or to the machine-learning literature. Next we discuss these papers and some related references.

1.2 Related Works in the Stochastic Setting

The above regret bound (1) can be seen as a deterministic online counterpart of the so-called *sparsity oracle inequalities* introduced in the stochastic setting in the past decade. The latter are risk bounds expressed in terms of the number of non-zero coordinates of the oracle vector—see (2) below. More formally, consider the regression model with random of fixed design. The forecaster observes independent random pairs $(X_1, Y_1), \ldots, (X_T, Y_T) \in \mathcal{X} \times \mathbb{R}$ given by

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

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 paragraph, 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). The goal of the forecaster is to construct an estimator $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ of the unknown regression function $f : \mathcal{X} \to \mathbb{R}$ based on the sample $(X_t, Y_t)_{1 \leqslant t \leqslant T}$. Depending on the nature of the design, the performance of $\widehat{f}_T$ is measured through its risk $R\big(\widehat{f}_T\big)$:

$$R\big(\widehat{f}_T\big) \triangleq \begin{cases} \displaystyle\int_{\mathcal{X}} \Big(f(x) - \widehat{f}_T(x)\Big)^2 P^X(\mathrm{d}x) & \text{(random design)} \\ \displaystyle\frac{1}{T} \sum_{t=1}^{T} \big(f(X_t) - \widehat{f}_T(X_t)\big)^2 & \text{(fixed design),} \end{cases}$$

where P^X denotes the common distribution of the X_t if the design is random. With the above notations, and given a dictionary $\varphi = (\varphi_1, \cdots, \varphi_d)$ of base forecasters $\varphi_j : \mathcal{X} \to \mathbb{R}$ as previously, typical examples of *sparsity oracle inequalities* take approximately the form

$$R\big(\widehat{f}_T\big) \leqslant C \inf_{u \in \mathbb{R}^d} \left\{ R\big(u \cdot \varphi\big) + \frac{\|u\|_0 \ln d + 1}{T} \right\} \tag{2}$$

in expectation or with high probability, for some constant $C \geqslant 1$. Thus, sparsity oracle inequalities are risk bounds involving a trade-off between the risk $R(u \cdot \varphi)$ and the number of non-zero coordinates $\|u\|_0$ of any comparison vector $u \in \mathbb{R}^d$. In particular, they indicate that $\widehat{f}_T$ has a small risk

under a sparsity scenario, that is, if f is well approximated by a sparse linear combination $u^* \cdot \varphi$ of the base forecasters φ_j, $1 \leqslant j \leqslant d$.

Sparsity oracle inequalities were first derived by Birgé and Massart (2001) via ℓ^0-regularization methods (through model-selection arguments). Later works in this direction include, among many other papers, those of Birgé and Massart (2007), Abramovich et al. (2006), and Bunea et al. (2007a) in the regression model with fixed design and that of Bunea et al. (2004) in the random design case.

More recently, a large body of research has been dedicated to the analysis of ℓ^1-regularization methods, which are convex and thus computationally tractable variants of ℓ^0-regularization methods. A celebrated example is the Lasso estimator introduced by Tibshirani (1996) and Donoho and Johnstone (1994). Under some assumptions on the design matrix,[2] such methods have been proved to satisfy sparsity oracle inequalities of the form (2) (with $C = 1$ in the recent paper by Koltchinskii et al. 2011). A list of few references—but far from being comprehensive—includes the works of Bunea et al. (2007b), Candes and Tao (2007), van de Geer (2008), Bickel et al. (2009), Koltchinskii (2009a), Koltchinskii (2009b),Hebiri and van de Geer (2011), Koltchinskii et al. (2011) and Lounici et al. (2011). We refer the reader to the monograph by Bühlmann and van de Geer (2011) for a detailed account on ℓ^1-regularization.

A third line of research recently focused on procedures based on exponential weighting. Such methods were proved to satisfy sharp sparsity oracle inequalities (i.e., with leading constant $C = 1$), either in the regression model with fixed design (Dalalyan and Tsybakov, 2007, 2008; Rigollet and Tsybakov, 2011; Alquier and Lounici, 2011) or in the regression model with random design (Dalalyan and Tsybakov, 2012a; Alquier and Lounici, 2011). These papers show that a trade-off can be reached between strong theoretical guarantees (as with ℓ^0-regularization) and computational efficiency (as with ℓ^1-regularization). They indeed propose aggregation algorithms which satisfy sparsity oracle inequalities under almost no assumption on the base forecasters $(\varphi_j)_j$, and which can be approximated numerically at a reasonable computational cost for large values of the ambient dimension d.

Our online-learning algorithm SeqSEW is inspired from a statistical method of Dalalyan and Tsybakov (2008, 2012a). Following the same lines as in Dalalyan and Tsybakov (2012b), it is possible to slightly adapt the statement of our algorithm to make it computationally tractable by means of Langevin Monte-Carlo approximation—without affecting its statistical properties. The technical details are however omitted in this paper, which only focuses on the theoretical guarantees of the algorithm SeqSEW.

1.3 Previous Works on Sparsity in the Framework of Individual Sequences

To the best of our knowledge, Corollary 2 and its refinements (Corollary 7 and Theorem 10) provide the first examples of sparsity regret bounds in the sense of (1). To comment on the optimality of such regret bounds and compare them to related results in the framework of individual sequences, note that (1) can be rewritten in the equivalent form:

2. Despite their computational efficiency, the aforementioned ℓ^1-regularized methods still suffer from a drawback: their ℓ^0-oracle properties hold under rather restrictive assumptions on the design; namely, that the φ_j should be nearly orthogonal (see the detailed discussion in van de Geer and Bühlmann 2009).

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

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 - \inf_{\substack{\|u\|_0 \leqslant s \\ \|u\|_1 \leqslant U}} \sum_{t=1}^{T} \big(y_t - u \cdot \varphi(x_t)\big)^2 \leqslant \big(s+1\big)\, g_{T,d}\big(U, \|\varphi\|_\infty\big) \;,$$

where g grows at most logarithmically in T, d, U, and $\|\varphi\|_\infty$. When $s \ll T$, this upper bound matches (up to logarithmic factors) the lower bound of order $s \ln T$ that follows in a straightforward manner from Theorem 2 of Vovk (2001). Indeed, if $s \ll T$, $\mathcal{X} = \mathbb{R}^d$, and $\varphi_j(x) = x_j$, then for any forecaster, there is an individual sequence $(x_t, y_t)_{1 \leqslant t \leqslant T}$ such that the regret of this forecaster on $\big\{u \in \mathbb{R}^d : \|u\|_0 \leqslant s \text{ and } \|u\|_1 \leqslant d\big\}$ is bounded from below by a quantity of order $s \ln T$. Therefore, up to logarithmic factors, any algorithm satisfying a sparsity regret bound of the form (1) is minimax optimal on intersections of ℓ^0-balls (of radii $s \ll T$) and ℓ^1-balls. This is in particular the case for our algorithm SeqSEW, but this contrasts with related works discussed below.

Recent works in the field of online convex optimization addressed the sparsity issue in the online deterministic setting, but from a quite different angle. They focus on algorithms which output sparse linear combinations, while we are interested in algorithms whose regret is small under a sparsity scenario, that is, on ℓ^0-balls of small radii. See, for example, the papers by Langford et al. (2009), Shalev-Shwartz and Tewari (2011), Xiao (2010), Duchi et al. (2010) 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 propose algorithms which predict as a sparse linear combination $\widehat{y}_t = \widehat{u}_t \cdot \varphi(x_t)$ of the base forecasts (i.e., $\|\widehat{u}_t\|_0$ is small), while no such guarantee can be proved for our algorithm SeqSEW. However they prove bounds on the ℓ^1-regularized regret of the form

$$\sum_{t=1}^{T} \Big((y_t - \widehat{u}_t \cdot x_t)^2 + \lambda \|\widehat{u}_t\|_1 \Big) \leqslant \inf_{u \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \Big((y_t - u \cdot x_t)^2 + \lambda \|u\|_1 \Big) + \widetilde{\Delta}_{T,d}(u) \right\} , \tag{3}$$

for some regret term $\widetilde{\Delta}_{T,d}(u)$ which is suboptimal on intersections of ℓ^0- and ℓ^1-balls as explained below. The truncated gradient algorithm of Langford et al. (2009, Corollary 4.1) satisfies such a regret bound[3] with $\widetilde{\Delta}_{T,d}(u)$ at least of order $\|\varphi\|_\infty \sqrt{dT}$ when the base forecasts $\varphi_j(x_t)$ are dense in the sense that $\max_{1 \leqslant t \leqslant T} \sum_{j=1}^{d} \varphi_j^2(x_t) \approx d \, \|\varphi\|_\infty^2$. This regret bound grows as a power of and not logarithmically in d as is expected for sparsity regret bounds (recall that we are interested in the case when $d \gg T$).

The three other papers mentioned above do prove (some) regret bounds with a logarithmic dependence in d, but these bounds do not have the dependence in $\|u\|_1$ and T we are looking for. For $p - 1 \approx 1/(\ln d)$, the p-norm RDA method of Xiao (2010) and the algorithm SMIDAS of Shalev-Shwartz and Tewari (2011)—the latter being a particular case of the algorithm COMID of Duchi et al. (2010) specialized to the p-norm divergence—satisfy regret bounds of the above form (3) with

3. The bound stated in Langford et al. (2009, Corollary 4.1) differs from (3) in that the constant before the infimum is equal to $C = 1/(1 - 2c_d^2 \eta)$, where $c_d^2 \approx \max_{1 \leqslant t \leqslant T} \sum_{j=1}^{d} \varphi_j^2(x_t) \leqslant d \, \|\varphi\|_\infty^2$, and where a reasonable choice for η can easily be seen to be $\eta \approx 1/\sqrt{2c_d^2 T}$. If the base forecasts $\varphi_j(x_t)$ are dense in the sense that $c_d^2 \approx d \, \|\varphi\|_\infty^2$, then we have $C \approx 1 + \sqrt{2c_d^2/T}$, which yields a regret bound with leading constant 1 as in (3) and with $\widetilde{\Delta}_{T,d}(u)$ at least of order $\sqrt{c_d^2 T} \approx \|\varphi\|_\infty \sqrt{dT}$.

$\widetilde{\Delta}_{T,d}(u) \approx \mu \|u\|_1 \sqrt{T \ln d}$, for some gradient-based constant μ. Therefore, in all three cases, the function $\widetilde{\Delta}$ grows at least linearly in $\|u\|_1$ and as $\sqrt{T}$. This is in contrast with the logarithmic dependence in $\|u\|_1$ and the fast rate $O(\ln T)$ we are looking for and prove, for example, in Corollary 2.

Note that the suboptimality of the aforementioned algorithms is specific to the goal we are pursuing, that is, prediction on ℓ^0-balls (intersected with ℓ^1-balls). On the contrary the rate $\|u\|_1 \sqrt{T \ln d}$ is more suited and actually nearly optimal for learning on ℓ^1-balls (see Gerchinovitz and Yu 2011). Moreover, the predictions output by our algorithm SeqSEW are not necessarily sparse linear combinations of the base forecasts. A question left open is thus whether it is possible to design an algorithm which both ouputs sparse linear combinations (which is statistically useful and sometimes essential for computational issues) and satisfies a sparsity regret bound of the form (1).

1.4 PAC-Bayesian Analysis in the Framework of Individual Sequences

To derive our sparsity regret bounds, we follow a PAC-Bayesian approach combined with the choice of a sparsity-favoring prior. We do not have the space to review the PAC-Bayesian literature in the stochastic setting and only refer the reader to Catoni (2004) for a thorough introduction to the subject. As for the online deterministic setting, PAC-Bayesian-type inequalities were proved in the framework of prediction with expert advice, for example, by Freund et al. (1997) and Kivinen and Warmuth (1999), or in the same setting as ours with a Gaussian prior by Vovk (2001). More recently, Audibert (2009) proved a PAC-Bayesian result on individual sequences for general losses and prediction sets. The latter result relies on a unifying assumption called the online variance inequality, which holds true, for example, when the loss function is exp-concave. In the present paper, we only focus on the particular case of the square loss. We first use Theorem 4.6 of Audibert (2009) to derive a non-adaptive sparsity regret bound. We then provide an adaptive online PAC-Bayesian inequality to automatically adapt to the unknown range of the observations $\max_{1 \leqslant t \leqslant T} |y_t|$.

1.5 Application to the Stochastic Setting When the Noise Level Is Unknown

In Section 4.1 we apply an automatically-tuned version of our algorithm SeqSEW on i.i.d. data. Thanks to the standard online-to-batch conversion, our sparsity regret bounds—of deterministic nature—imply a sparsity oracle inequality of the same flavor as a result of Dalalyan and Tsybakov (2012a). However, our risk bound holds on the whole $\mathbb{R}^d$ space instead of ℓ^1-balls of finite radii, which solves one question left open by Dalalyan and Tsybakov (2012a, Section 4.2). Besides, and more importantly, our algorithm does not need the a priori knowledge of the variance of the noise when the latter is Gaussian. Since the noise level is unknown in practice, adapting to it is important. This solves a second question raised by Dalalyan and Tsybakov (2012a, Section 5.1, Remark 6).

1.6 Outline of the Paper

This paper is organized as follows. In Section 2 we describe our main (deterministic) setting as well as our main notations. In Section 3 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 3.1 and 3.2), and then when no a priori information is available (Section 3.3), which yields a fully automatic algorithm. In Section 4 we apply the algorithm SeqSEW to two stochastic settings: the regression model with random design (Section 4.1) and the regression model with fixed design (Section 4.2). Finally the appendix contains some proofs and several useful inequalities.

2. Setting and Notations

The main setting considered in this paper is an instance of the game of prediction with expert advice called *prediction with side information (under the square loss)* or, more simply, *online linear regression* (see Cesa-Bianchi and Lugosi 2006, Chapter 11 for an introduction to this setting). The data sequence $(x_t, y_t)_{t\geqslant 1}$ at hand is deterministic and arbitrary and we look for theoretical guarantees that hold for every *individual* sequence. We give in Figure 1 a detailed description of our online protocol.

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

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

At each time round $t \in \mathbb{N}^* \triangleq \{1, 2, \ldots\}$,

1. The environment reveals the input data $x_t \in \mathcal{X}$.
2. The forecaster chooses a prediction $\widehat{y}_t \in \mathbb{R}$
   (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 $u \cdot \varphi \triangleq \sum_{j=1}^d u_j \varphi_j$, $u \in \mathbb{R}^d$, incurs the loss $\big(y_t - u \cdot \varphi(x_t)\big)^2$ and the forecaster incurs the loss $(y_t - \widehat{y}_t)^2$.

Figure 1: The online linear regression setting.

Note that our online protocol is described as if the environment were oblivious to the forecaster's predictions. Actually, since we only consider deterministic forecasters, all regret bounds of this paper also hold when $(x_t)_{t\geqslant 1}$ and $(y_t)_{t\geqslant 1}$ are chosen by an adversarial environment.

Two stochastic batch settings are also considered later in this paper. See Section 4.1 for the regression model with random design, and Section 4.2 for the regression model with fixed design.

2.1 Some Notations

We now define some notations. We write $\mathbb{N} \triangleq \{0, 1, \ldots\}$ and $\mathrm{e} \triangleq \exp(1)$. Vectors in $\mathbb{R}^d$ will be denoted by bold letters. For all $u, v \in \mathbb{R}^d$, the standard inner product in $\mathbb{R}^d$ between $u = (u_1, \ldots, u_d)$ and $v = (v_1, \ldots, v_d)$ will be denoted by $u \cdot v = \sum_{i=j}^d u_j v_j$; the ℓ^0-, ℓ^1-, and ℓ^2-norms of $u = (u_1, \ldots, u_d)$ are respectively defined by

$$\|u\|_0 \triangleq \sum_{j=1}^d \mathbb{I}_{\{u_j \neq 0\}} = \big|\{j : u_j \neq 0\}\big| , \qquad \|u\|_1 \triangleq \sum_{j=1}^d |u_j| , \qquad \text{and} \quad \|u\|_2 \triangleq \left(\sum_{j=1}^d u_j^2\right)^{1/2} .$$

The set of all probability distributions on a set Θ (endowed with some σ-algebra, for example, the Borel σ-algebra when $\Theta = \mathbb{R}^d$) 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

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

where $\frac{\mathrm{d}\rho}{\mathrm{d}\pi}$ 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 (or clipped) value:

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

Finally, we will use the (natural) conventions $1/0 = +\infty$, $(+\infty) \times 0 = 0$, and $0 \ln(1 + U/0) = 0$ for all $U \geqslant 0$. Any sum $\sum_{s=1}^{0} a_s$ indexed from 1 up to 0 is by convention equal to 0.

3. Sparsity Regret Bounds for Individual Sequences

In this section we prove sparsity regret bounds for different variants of our algorithm SeqSEW. We first assume in Section 3.1 that the forecaster has access in advance to a bound B_y on the observations $|y_t|$ and a bound B_Φ on the trace of the empirical Gram matrix. We then remove these requirements one by one in Sections 3.2 and 3.3.

3.1 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 $y_1, \ldots, y_T$ and to a bound B_Φ on the trace of the empirical Gram matrix, that is,

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

The first version of the algorithm studied in this paper is defined in Figure 2 (adaptive variants will be introduced later). We name it *SeqSEW* for it is a variant of the Sparse Exponential Weighting algorithm introduced in the stochastic setting by Dalalyan and Tsybakov (2007, 2008) which is tailored for the prediction of individual sequences.

The choice of the heavy-tailed prior π_τ is due to Dalalyan and Tsybakov (2007). The role of heavy-tailed priors to tackle the sparsity issue was already pointed out earlier; see, for example, the discussion by Seeger (2008, Section 2.1). In high dimension, such heavy-tailed priors favor sparsity: sampling from these prior distributions (or posterior distributions based on them) typically results in approximately sparse vectors, that is, vectors having most coordinates almost equal to zero and the few remaining ones with quite large values.

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

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

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

At each time round $t \geqslant 1$,

1. Get the input data x_t and predict[a] as $\widehat{y}_t \triangleq \displaystyle\int_{\mathbb{R}^d} \big[u \cdot \varphi(x_t)\big]_B \, p_t(\mathrm{d}u)$;

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

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

where

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

a. The clipping operator $[\cdot]_B$ is defined in Section 2.

Figure 2: The algorithm $\mathrm{SeqSEW}_\tau^{B,\eta}$.

Proposition 1 *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 $B \geqslant B_y$, all $\eta \leqslant 1/(8B^2)$, and all $\tau > 0$, the algorithm* $\mathrm{SeqSEW}_\tau^{B,\eta}$ *satisfies*

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{u\in\mathbb{R}^d} \left\{ \sum_{t=1}^{T} \big(y_t - u\cdot\varphi(x_t)\big)^2 + \frac{4}{\eta}\,\|u\|_0 \ln\!\left(1 + \frac{\|u\|_1}{\|u\|_0\,\tau}\right) \right\} + \tau^2 \sum_{j=1}^{d}\sum_{t=1}^{T} \varphi_j^2(x_t) \ . \tag{4}$$

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

Then, when used with $B = B_y$, $\eta = \dfrac{1}{8B_y^2}$, and $\tau = \sqrt{\dfrac{16B_y^2}{B_\Phi}}$, the algorithm $\mathrm{SeqSEW}_\tau^{B,\eta}$ *satisfies*

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{u\in\mathbb{R}^d} \left\{ \sum_{t=1}^{T} \big(y_t - u\cdot\varphi(x_t)\big)^2 + 32\,B_y^2\,\|u\|_0 \ln\left(1 + \frac{\sqrt{B_\Phi}\,\|u\|_1}{4\,B_y\,\|u\|_0}\right) \right\} + 16B_y^2 \ . \tag{5}$$

Note that, if $\|\varphi\|_\infty \triangleq \sup_{x\in\mathcal{X}} \max_{1\leqslant j\leqslant d} |\varphi_j(x)|$ is finite, then the last corollary provides a *sparsity regret bound* in the sense of (1). Indeed, in this case, we can take $B_\Phi = d\,T\,\|\varphi\|_\infty^2$, which yields a regret bound proportional to $\|u\|_0$ and that grows logarithmically in d, T, $\|u\|_1$, and $\|\varphi\|_\infty$.

To prove Proposition 1, we first need the following deterministic PAC-Bayesian inequality which is at the core of our analysis. It is a straightforward consequence of Theorem 4.6 of Audibert (2009) when applied to the square loss. An adaptive variant of this inequality will be provided in Section 3.2.

Lemma 3 *Assume that for some known constant $B_y > 0$, we have $y_1, \dots, y_T \in [-B_y, B_y]$.*

For all $\tau > 0$, if the algorithm $\mathrm{SeqSEW}_\tau^{B,\eta}$ is used with $B \geqslant B_y$ and $\eta \leqslant 1/(8B^2)$, then

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^{T} \Big(y_t - \big[u \cdot \varphi(x_t)\big]_B \Big)^2 \rho(\mathrm{d}u) + \frac{\mathcal{K}(\rho, \pi_\tau)}{\eta} \right\} \tag{6}$$

$$\leqslant \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^{T} \big(y_t - u \cdot \varphi(x_t) \big)^2 \rho(\mathrm{d}u) + \frac{\mathcal{K}(\rho, \pi_\tau)}{\eta} \right\} . \tag{7}$$

Proof (of Lemma 3) Inequality (6) is a straightforward consequence of Theorem 4.6 of Audibert (2009) when applied to the square loss, the set of prediction functions $\mathcal{G} \triangleq \big\{ x \mapsto \big[u \cdot \varphi(x)\big]_B : u \in \mathbb{R}^d \big\}$, and the prior[4] $\widetilde{\pi}_\tau$ on $\mathcal{G}$ induced by the prior π_τ on $\mathbb{R}^d$ via the mapping $u \in \mathbb{R}^d \mapsto \big[u \cdot \varphi(\cdot)\big]_B \in \mathcal{G}$.

To apply the aforementioned theorem, recall from Cesa-Bianchi and Lugosi (2006, Section 3.3) that the square loss is $1/(8B^2)$-exp-concave on $[-B,B]$ and thus η-exp-concave,[5] since $\eta \leqslant 1/(8B^2)$ by assumption. Therefore, by Theorem 4.6 of Audibert (2009) with the variance function $\delta_\eta \equiv 0$ (see the comments following Remark 4.1 therein), we get

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{\mu \in \mathcal{M}_1^+(\mathcal{G})} \left\{ \int_{\mathcal{G}} \sum_{t=1}^{T} \big(y_t - g(x_t) \big)^2 \mu(\mathrm{d}g) + \frac{\mathcal{K}(\mu, \widetilde{\pi}_\tau)}{\eta} \right\}$$

$$\leqslant \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^{T} \Big(y_t - \big[u \cdot \varphi(x_t)\big]_B \Big)^2 \rho(\mathrm{d}u) + \frac{\mathcal{K}(\widetilde{\rho}, \widetilde{\pi}_\tau)}{\eta} \right\} ,$$

where the last inequality follows by restricting the infimum over $\mathcal{M}_1^+\big(\mathcal{G}\big)$ to the subset $\big\{\widetilde{\rho} : \rho \in \mathcal{M}_1^+(\mathbb{R}^d)\big\} \subset \mathcal{M}_1^+\big(\mathcal{G}\big)$, where $\widetilde{\rho} \in \mathcal{M}_1^+\big(\mathcal{G}\big)$ denotes the probability distribution induced by $\rho \in \mathcal{M}_1^+(\mathbb{R}^d)$ via the mapping $u \in \mathbb{R}^d \mapsto \big[u \cdot \varphi(\cdot)\big]_B \in \mathcal{G}$. Inequality (6) then follows from the fact that for all $\rho \in \mathcal{M}_1^+(\mathbb{R}^d)$, we have $\mathcal{K}(\widetilde{\rho}, \widetilde{\pi}_\tau) \leqslant \mathcal{K}(\rho, \pi_\tau)$ by joint convexity of $\mathcal{K}(\cdot,\cdot)$.

As for Inequality (7), it follows from (6) by noting that

$$\forall y \in [-B,B], \quad \forall x \in \mathbb{R}, \qquad \big| y - [x]_B \big| \leqslant |y - x| .$$

Therefore, 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 $t = 1, \dots, T$. ■

Remark 4 *As can be seen from the previous proof, if the prior π_τ used to define the algorithm SeqSEW was replaced with any prior $\pi \in \mathcal{M}_1^+(\mathbb{R}^d)$, then Lemma 3 would still hold true with π instead*

4. The set $\mathcal{G}$ is endowed with the σ-algebra generated by all the coordinate mappings $g \in \mathcal{G} \mapsto g(x) \in \mathbb{R}$, $x \in \mathcal{X}$ (where $\mathbb{R}$ is endowed with its Borel σ-algebra).

5. This means that for all $y \in [-B,B]$, the function $x \mapsto \exp\big(-\eta(y-x)^2\big)$ is concave on $[-B,B]$.

of π_τ. This fact is natural from a PAC-Bayesian perspective (see, e.g., Catoni, 2004; Dalalyan and Tsybakov, 2008). We only—but crucially—use the particular shape of the sparsity-favoring prior π_τ to derive Proposition 1 from the PAC-Bayesian bound (7).

Proof (of Proposition 1) Our proof mimics the proof of Theorem 5 by Dalalyan and Tsybakov (2008). We thus only write the outline of the proof and stress the minor changes that are needed to derive Inequality (4). The key technical tools provided by Dalalyan and Tsybakov (2008) are reproduced in Appendix B.2 for the convenience of the reader.

Let $u^* \in \mathbb{R}^d$. Since $B \geqslant B_y$ and $\eta \leqslant 1/(8B^2)$, we can apply Lemma 3 and get

$$\begin{aligned}\sum_{t=1}^T (y_t - \widehat{y}_t)^2 &\leqslant \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \big(y_t - u \cdot \varphi(x_t)\big)^2 \rho(\mathrm{d}u) + \frac{\mathcal{K}(\rho, \pi_\tau)}{\eta} \right\} \\ &\leqslant \underbrace{\int_{\mathbb{R}^d} \sum_{t=1}^T \big(y_t - u \cdot \varphi(x_t)\big)^2 \rho_{u^*,\tau}(\mathrm{d}u)}_{(1)} + \underbrace{\frac{\mathcal{K}(\rho_{u^*,\tau}, \pi_\tau)}{\eta}}_{(2)} \,. \end{aligned} \tag{8}$$

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

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

The two terms (1) and (2) can be upper bounded as in the proof of Theorem 5 by Dalalyan and Tsybakov (2008). By a symmetry argument recalled in Lemma 22 (Appendix B.2), the first term (1) can be rewritten as

$$\int_{\mathbb{R}^d} \sum_{t=1}^T \big(y_t - u \cdot \varphi(x_t)\big)^2 \rho_{u^*,\tau}(\mathrm{d}u) = \sum_{t=1}^T \big(y_t - u^* \cdot \varphi(x_t)\big)^2 + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \,. \tag{9}$$

As for the term (2), we have, as is recalled in Lemma 23,

$$\frac{\mathcal{K}(\rho_{u^*,\tau}, \pi_\tau)}{\eta} \leqslant \frac{4}{\eta} \|u^*\|_0 \ln\left(1 + \frac{\|u^*\|_1}{\|u^*\|_0 \tau}\right) \,. \tag{10}$$

Combining (8), (9), and (10), which all hold for all $u^* \in \mathbb{R}^d$, we get Inequality (4). ■

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

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

since $\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leqslant B_\Phi$ by assumption. The particular (and nearly optimal) choices of η and τ given in the statement of the corollary then yield the desired inequality (5). ■

We end this subsection with a natural question about approximate sparsity: Proposition 1 ensures a low regret with respect to sparse linear combinations $u \cdot \varphi$, but what can be said for approximately sparse linear combinations, that is, predictors of the form $u \cdot \varphi$ where $u \in \mathbb{R}^d$ is very close to a sparse vector? As can be seen from the proof of Lemma 23 in Appendix B.2, the sparsity-related term

$$\frac{4}{\eta} \|u\|_0 \ln\left(1 + \frac{\|u\|_1}{\|u\|_0 \tau}\right)$$

in the regret bound of Proposition 1 can actually be replaced with the smaller (and continous) term

$$\frac{4}{\eta} \sum_{j=1}^{d} \ln\left(1 + |u_j|/\tau\right) \ .$$

The last term is always smaller than the former and guarantees that the regret is small with respect to any approximately sparse vector $u \in \mathbb{R}^d$.

3.2 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 3 and Proposition 1, we assumed that the forecaster had access to a bound B_y on the observations $|y_t|$ and to a bound B_Φ on the trace of the empirical Gram matrix. In this section, we remove the first requirement and prove a sparsity regret bound for a variant of the algorithm $\mathrm{SeqSEW}_\tau^{B,\eta}$ which is adaptive to the unknown bound $B_y = \max_{1 \leqslant t \leqslant T} |y_t|$; see Proposition 5 and Remark 6 below.

For this purpose we consider the algorithm of Figure 3, which we call SeqSEW_τ^* thereafter. It differs from $\mathrm{SeqSEW}_\tau^{B,\eta}$ 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.

The idea of truncating the base forecasts was used many times in the past; see, for example, the work of Vovk (2001) in the online linear regression setting, that of Györfi et al. (2002, Chapter 10) for the regression problem with random design, and the papers of Györfi and Ottucsák (2007) and Biau et al. (2010) for sequential prediction of unbounded time series under the square loss. A key ingredient in the present paper is to perform truncation with respect to a data-driven threshold.

Proposition 5 *For all $\tau > 0$, the algorithm SeqSEW_τ^* satisfies*

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{u \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \big(y_t - u \cdot \varphi(x_t)\big)^2 + 32 B_{T+1}^2 \|u\|_0 \ln\left(1 + \frac{\|u\|_1}{\|u\|_0 \tau}\right) \right\} + \tau^2 \sum_{j=1}^{d} \sum_{t=1}^{T} \varphi_j^2(x_t) + 5 B_{T+1}^2 \ ,$$

where $B_{T+1}^2 \triangleq \max_{1 \leqslant t \leqslant T} y_t^2$.

Remark 6 In view of Proposition 1, the algorithm SeqSEW_τ^* satisfies a sparsity regret bound which is adaptive to the unknown bound $B_y = \max_{1 \leqslant t \leqslant T} |y_t|$. The price for the automatic tuning with respect to B_y consists only of the additive term $5B_{T+1}^2 = 5B_y^2$.

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

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

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

At each time round $t \geqslant 1$,

1. Get the input data x_t and predict[a] as $\widehat{y}_t \triangleq \displaystyle\int_{\mathbb{R}^d} \big[u \cdot \varphi(x_t)\big]_{B_t} p_t(\mathrm{d}u)$;
2. Get the observation y_t and update:
   - the threshold $B_{t+1} \triangleq \max_{1 \leqslant s \leqslant t} |y_s|$,
   - the inverse temperature $\eta_{t+1} \triangleq 1/\big(8B_{t+1}^2\big)$,
   - and the posterior distribution $p_{t+1} \in \mathcal{M}_1^+(\mathbb{R}^d)$ as

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

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

a. The clipping operator $[\cdot]_B$ is defined in Section 2.

Figure 3: The algorithm SeqSEW_τ^*.

As in the previous section, several corollaries can be derived from Proposition 5. If the forecaster has access beforehand to a quantity $B_\Phi > 0$ such that $\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leqslant B_\Phi$, then a suboptimal but reasonable choice of τ is given by $\tau = 1/\sqrt{B_\Phi}$; see Corollary 7 below. The simpler tuning $\tau = 1/\sqrt{dT}$ of Corollary 8 will be useful in the stochastic batch setting (cf., Section 4).[6] The proofs of the next corollaries are immediate.

Corollary 7 *Assume that, for a known constant $B_\Phi > 0$, the $(x_1,y_1),\dots,(x_T,y_T)$ are such that $\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leqslant B_\Phi$. Then, when used with $\tau = 1/\sqrt{B_\Phi}$, the algorithm* SeqSEW_τ^* *satisfies*

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{u\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t - u\cdot\varphi(x_t)\big)^2 + 32B_{T+1}^2\,\|u\|_0 \ln\left(1+\frac{\sqrt{B_\Phi}\,\|u\|_1}{\|u\|_0}\right)\right\} + 5B_{T+1}^2 + 1 \ ,$$

6. The tuning $\tau = 1/\sqrt{dT}$ only uses the knowledge of T, which is known by the forecaster in the stochastic batch setting. In that framework, another simple and easy-to-analyse tuning is given by $\tau = 1/(\|\varphi\|_\infty\sqrt{dT})$—which corresponds to $B_\Phi = dT\,\|\varphi\|_\infty^2$—but it requires that $\|\varphi\|_\infty \triangleq \sup_{x\in\mathcal{X}} \max_{1\leqslant j\leqslant d} |\varphi_j(x)|$ be finite. Note that the last tuning satisfies the scale-invariant property pointed out by Dalalyan and Tsybakov (2012a, Remark 4).

where $B_{T+1}^2 \triangleq \max_{1\leqslant t\leqslant T} y_t^2$.

Corollary 8 *Assume that T is known to the forecaster at the beginning of the prediction game. Then, when used with $\tau = 1/\sqrt{dT}$, the algorithm* SeqSEW$_\tau^*$ *satisfies*

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{u\in\mathbb{R}^d} \left\{ \sum_{t=1}^T \big(y_t - u\cdot\varphi(x_t)\big)^2 + 32 B_{T+1}^2 \, \|u\|_0 \ln\left(1 + \frac{\sqrt{dT}\,\|u\|_1}{\|u\|_0}\right) \right\} + \frac{1}{dT}\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) + 5B_{T+1}^2 \,,$$

where $B_{T+1}^2 \triangleq \max_{1\leqslant t\leqslant T} y_t^2$.

As in the previous section, to prove Proposition 5, we first need a key PAC-Bayesian inequality. The next lemma is an adaptive variant of Lemma 3.

Lemma 9 *For all $\tau > 0$, the algorithm* SeqSEW$_\tau^*$ *satisfies*

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\rho\in\mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \Big(y_t - \big[u\cdot\varphi(x_t)\big]_{B_t}\Big)^2 \rho(\mathrm{d}u) + 8B_{T+1}^2\,\mathcal{K}(\rho,\pi_\tau) \right\} + 4B_{T+1}^2 \tag{11}$$

$$\leqslant \inf_{\rho\in\mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \big(y_t - u\cdot\varphi(x_t)\big)^2 \rho(\mathrm{d}u) + 8B_{T+1}^2\,\mathcal{K}(\rho,\pi_\tau) \right\} + 5B_{T+1}^2 \,, \tag{12}$$

where $B_{T+1}^2 \triangleq \max_{1\leqslant t\leqslant T} y_t^2$.

Proof (of Lemma 9) The proof is based on arguments that are similar to those underlying Lemma 3, except that we now need to deal with B and η changing over time. In the same spirit as in Auer et al. (2002), Cesa-Bianchi et al. (2007) and Györfi and Ottucsák (2007), our analysis relies on the control of $(\ln W_{t+1})/\eta_{t+1} - (\ln W_t)/\eta_t$ where $W_1 \triangleq 1$ and, for all $t \geqslant 2$,

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

Before controlling $(\ln W_{t+1})/\eta_{t+1} - (\ln W_t)/\eta_t$, we first need a little comment. Note that all η_t's such that $\eta_t = +\infty$ (i.e., $B_t = 0$) can be replaced with any finite value without changing the predictions of the algorithm (since the sum $\sum_{s=1}^{t-1}$ above equals zero). Therefore, we assume in the sequel that $(\eta_t)_{t\geqslant 1}$ is a non-decreasing sequence of *finite* positive real numbers.

First step: On the one hand, we have

$$\frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} = \frac{1}{\eta_{T+1}} \ln \int_{\mathbb{R}^d} \exp\left(-\eta_{T+1}\sum_{t=1}^T \Big(y_t - \big[u\cdot\varphi(x_t)\big]_{B_t}\Big)^2\right) \pi_\tau(\mathrm{d}u) - \frac{1}{\eta_1}\ln 1$$

$$= -\inf_{\rho\in\mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \Big(y_t - \big[u\cdot\varphi(x_t)\big]_{B_t}\Big)^2 \rho(\mathrm{d}u) + \frac{\mathcal{K}(\rho,\pi_\tau)}{\eta_{T+1}} \right\} , \tag{13}$$

where the last equality follows from a convex duality argument for the Kullback-Leibler divergence (cf., e.g., Catoni 2004, p. 159) which we recall in Proposition 21 in Appendix B.1.

Second step: 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

$$\frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} = \sum_{t=1}^{T} \left(\frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W_t}{\eta_t} \right) = \sum_{t=1}^{T} \Bigg(\underbrace{\frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W'_{t+1}}{\eta_t}}_{(1)} + \underbrace{\frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t}}_{(2)} \Bigg) , \tag{14}$$

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

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

Let $t \in \{1, \dots, T\}$. The first term (1) is non-positive by Jensen's inequality (note that $x \mapsto x^{\eta_{t+1}/\eta_t}$ is concave on $\mathbb{R}_+^*$ since $\eta_{t+1} \leqslant \eta_t$ by construction). As for the second term (2), by definition of W'_{t+1},

$$\begin{aligned}
&\frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t} \\
&\quad = \frac{1}{\eta_t} \ln \int_{\mathbb{R}^d} \frac{\exp\left(-\eta_t \Big(y_t - \big[u \cdot \varphi(x_t)\big]_{B_t} \Big)^2 \right) \exp\left(-\eta_t \sum_{s=1}^{t-1} \Big(y_s - \big[u \cdot \varphi(x_s)\big]_{B_s} \Big)^2 \right)}{W_t} \pi_\tau(\mathrm{d}u) \\
&\quad = \frac{1}{\eta_t} \ln \int_{\mathbb{R}^d} \exp\left(-\eta_t \Big(y_t - \big[u \cdot \varphi(x_t)\big]_{B_t} \Big)^2 \right) p_t(\mathrm{d}u) \, .
\end{aligned} \tag{15}$$

where (15) follows by definition of p_t. The next paragraphs are dedicated to upper bounding the last integral above. First note that this is straightforward in the particular case where $y_t \in [-B_t, B_t]$. Indeed, by definition of $\eta_t \triangleq 1/(8B_t^2)$ and by the fact that the square loss is $1/(8B_t^2)$-exp-concave on $[-B_t, B_t]$ (as in Lemma 3),[7] we get from Jensen's inequality that

$$\int_{\mathbb{R}^d} e^{-\eta_t \left(y_t - \left[u \cdot \varphi(x_t)\right]_{B_t} \right)^2} p_t(\mathrm{d}u) \leqslant \exp\left(-\eta_t \left(y_t - \int_{\mathbb{R}^d} \big[u \cdot \varphi(x_t)\big]_{B_t} p_t(\mathrm{d}u) \right)^2 \right) = e^{-\eta_t (y_t - \widehat{y}_t)^2} ,$$

where the last equality follows by definition of $\widehat{y}_t$. Taking the logarithms of both sides of the last inequality and dividing by η_t, we can see that the quantity on the right-hand side of (15) is bounded from above by $-\big(y_t - \widehat{y}_t\big)^2$.

In the general case, we cannot assume that $y_t \in [-B_t, B_t]$, since it may happen that $|y_t| > \max_{1 \leqslant s \leqslant t-1} |y_s| \triangleq B_t$. As shown below, we can still use the exp-concavity of the square loss if we replace y_t with its clipped version $[y_t]_{B_t}$. More precisely, setting $\widehat{y}_{t,u} \triangleq [u \cdot \varphi(x_t)]_{B_t}$ for all $u \in \mathbb{R}^d$, the square loss appearing in the right-hand side of (15) equals

$$\begin{aligned}
\big(y_t - \widehat{y}_{t,u}\big)^2 &= \big([y_t]_{B_t} - \widehat{y}_{t,u}\big)^2 + \big(y_t - [y_t]_{B_t}\big)^2 + 2\big(y_t - [y_t]_{B_t}\big)\big([y_t]_{B_t} - \widehat{y}_{t,u}\big) \\
&= \big([y_t]_{B_t} - \widehat{y}_{t,u}\big)^2 + \big(y_t - [y_t]_{B_t}\big)^2 + 2\big(y_t - [y_t]_{B_t}\big)\big([y_t]_{B_t} - \widehat{y}_t\big) + c_{t,u} \, ,
\end{aligned} \tag{16}$$

7. To be more exact, we assigned some arbitrary finite value to η_t when $B_t = 0$. However, in this case, the square loss is of course η_t-exp-concave on $[-B_t, B_t] = \{0\}$ whatever the value of η_t.

where we set

$$\begin{aligned} c_{t,u} &\triangleq 2\big(y_t - [y_t]_{B_t}\big)\big(\widehat{y}_t - \widehat{y}_{t,u}\big) \\ &\geqslant -4B_t\big|y_t - [y_t]_{B_t}\big| \geqslant -4B_t(B_{t+1} - B_t)\;, \end{aligned} \tag{17}$$

where the last two inequalities follow from the property $\widehat{y}_t, \widehat{y}_{t,u} \in [-B_t, B_t]$ (by construction) and from the elementary[8] yet useful upper bound $\big|y_t - [y_t]_{B_t}\big| \leqslant B_{t+1} - B_t$.

Combining (16) with the lower bound (17) yields that, for all $u \in \mathbb{R}^d$,

$$\big(y_t - \widehat{y}_{t,u}\big)^2 \geqslant \big([y_t]_{B_t} - \widehat{y}_{t,u}\big)^2 + C_t\;, \tag{18}$$

where we set $C_t \triangleq \big(y_t - [y_t]_{B_t}\big)^2 + 2\big(y_t - [y_t]_{B_t}\big)\big([y_t]_{B_t} - \widehat{y}_t\big) - 4B_t(B_{t+1} - B_t)$.

We can now continue the upper bounding of $(1/\eta_t)\ln(W'_{t+1}/W_t)$. Indeed, substituting the lower bound (18) into (15), we get that

$$\begin{aligned} \frac{1}{\eta_t}\ln\frac{W'_{t+1}}{W_t} &\leqslant \frac{1}{\eta_t}\ln\left[\int_{\mathbb{R}^d}\exp\Big(-\eta_t\big([y_t]_{B_t} - \widehat{y}_{t,u}\big)^2\Big)\,p_t(\mathrm{d}u)\right] - C_t \\ &\leqslant \frac{1}{\eta_t}\ln\left[\exp\left(-\eta_t\left([y_t]_{B_t} - \int_{\mathbb{R}^d}\widehat{y}_{t,u}\,p_t(\mathrm{d}u)\right)^2\right)\right] - C_t \qquad (19) \\ &= -\big([y_t]_{B_t} - \widehat{y}_t\big)^2 - C_t \qquad (20) \\ &= -\big(y_t - \widehat{y}_t\big)^2 + 4B_t(B_{t+1} - B_t)\;, \qquad (21) \end{aligned}$$

where (19) follows by Jensen's inequality (recall that $\eta_t \triangleq 1/(8B_t^2)$ and that the square loss is $1/(8B_t^2)$-exp-concave on $[-B_t, B_t]$),[9] where (20) is entailed by definition of $\widehat{y}_{t,u}$ and $\widehat{y}_t$, and where (21) follows by definition of C_t above and by elementary calculations.

Summing (21) over $t = 1, \ldots, T$ and using the upper bound $B_t(B_{t+1} - B_t) \leqslant B_{t+1}^2 - B_t^2$, Equation (14) yields

$$\begin{aligned} \frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} &\leqslant -\sum_{t=1}^T (y_t - \widehat{y}_t)^2 + 4\sum_{t=1}^T \big(B_{t+1}^2 - B_t^2\big) \\ &= -\sum_{t=1}^T (y_t - \widehat{y}_t)^2 + 4B_{T+1}^2\;. \end{aligned} \tag{22}$$

Third step: Putting (13) and (22) together, we get the PAC-Bayesian inequality

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \Big(y_t - \big[u\cdot\varphi(x_t)\big]_{B_t}\Big)^2 \rho(\mathrm{d}u) + \frac{\mathcal{K}(\rho,\pi_\tau)}{\eta_{T+1}} \right\} + 4B_{T+1}^2\;,$$

which yields (11) since $\eta_{T+1} \triangleq 1/(8B_{T+1}^2)$ by definition.[10] The other PAC-Bayesian inequality (12), which is stated for non-truncated base forecasts, is a direct consequence of (11) and of the following two arguments: for all $u \in \mathbb{R}^d$ and all $t = 1, \ldots, T$,

$$\big(y_t - [u\cdot\varphi(x_t)]_{B_t}\big)^2 \leqslant \big(y_t - u\cdot\varphi(x_t)\big)^2 + (B_{t+1} - B_t)^2 \tag{23}$$

8. To see why this is true, it suffices to rewrite $[y_t]_{B_t}$ in the three cases $y_t < -B_t$, $|y_t| \leqslant B_t$, or $y_t > B_t$.
9. Same remark as in Footnote 7.
10. If $B_{T+1} = 0$, then $y_t = \widehat{y}_t = 0$ for all $1 \leqslant t \leqslant T$, which immediately yields (11).

and

$$\sum_{t=1}^{T}(B_{t+1}-B_t)^2 \leqslant B_{T+1}^2 \ . \tag{24}$$

Complement: proof of (23) and (24).
To see why (23) is true, we can distinguish between several cases. First note that this inequality is straightforward when $|y_t| \leqslant B_t$ (indeed, in this case, clipping $u\cdot\varphi(x_t)$ to $[-B_t,B_t]$ can only improve prediction). We can thus assume that $|y_t| > B_t$, or just[11] that $y_t > B_t$. In this case, we can distinguish between three sub-cases:

- if $u\cdot\varphi(x_t) < -B_t$, then clipping improves prediction since $y_t > B_t$;
- if $-B_t \leqslant u\cdot\varphi(x_t) \leqslant B_t$, then the clipping operator $[\cdot]_B$ has no effect on $u\cdot\varphi(x_t)$;
- if $u\cdot\varphi(x_t) > B_t$, then $[u\cdot\varphi(x_t)]_{B_t} = B_t$ so that $(y_t - [u\cdot\varphi(x_t)]_{B_t})^2 = (B_{t+1}-B_t)^2$ since $B_{t+1} = y_t$.

Therefore, in all three sub-cases described above, we have

$$(y_t - [u\cdot\varphi(x_t)]_{B_t})^2 \leqslant \max\left\{(y_t - u\cdot\varphi(x_t))^2, (B_{t+1}-B_t)^2\right\} \ ,$$

which concludes the proof of (23). As for (24), it follows from the inequality

$$\sum_{t=1}^{T}(B_{t+1}-B_t)^2 \leqslant \sup_{\substack{\Delta_1,\dots,\Delta_T\geqslant 0\\ \sum_{t=1}^T\Delta_t = B_{T+1}}} \left\{\sum_{t=1}^{T}\Delta_t^2\right\} = B_{T+1}^2 \ ,$$

where the last equality is entailed by convexity of the function $(\Delta_1,\dots,\Delta_T)\mapsto\sum_{t=1}^T\Delta_t^2$ on the polytope $\left\{(\Delta_1,\dots,\Delta_T)\in\mathbb{R}_+^T : \sum_{t=1}^T\Delta_t = B_{T+1}\right\}$. This concludes the proof. ■

Proof (of Proposition 5) The proof follows exactly the same lines as in Proposition 1 except that we apply Lemma 9 instead of Lemma 3. Indeed, using Lemma 9 and restricting the infimum to the $\rho_{u^*,\tau}$, $u^*\in\mathbb{R}^d$ (cf., (40)), we get that

$$\begin{aligned}
\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 &\leqslant \inf_{u^*\in\mathbb{R}^d}\left\{\int_{\mathbb{R}^d}\sum_{t=1}^{T}\big(y_t - u\cdot\varphi(x_t)\big)^2\rho_{u^*,\tau}(\mathrm{d}u) + 8B_{T+1}^2\,\mathcal{K}(\rho_{u^*,\tau},\pi_\tau)\right\} + 5B_{T+1}^2\\
&\leqslant \inf_{u^*\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t - u^*\cdot\varphi(x_t)\big)^2 + 32B_{T+1}^2\,\|u^*\|_0\ln\left(1+\frac{\|u^*\|_1}{\|u^*\|_0\,\tau}\right)\right\}\\
&\quad + \tau^2\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t) + 5B_{T+1}^2 \ ,
\end{aligned}$$

where the last inequality follows from Lemmas 22 and 23. ■

11. If $y_t < -B_t$, it suffices to apply (23) with $-y_t$ and $-u$.

3.3 A Fully Automatic Algorithm

In the previous section, we proved that adaptation to B_y was possible. If we also no longer assume that a bound B_Φ on the trace of the empirical Gram matrix is available to the forecaster, then we can use a doubling trick on the nondecreasing quantity

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

and repeatedly run the algorithm SeqSEW$^*_\tau$ of the previous section for rapidly-decreasing values of τ. This yields a sparsity regret bound with extra logarithmic multiplicative factors as compared to Proposition 5, but which holds for a fully automatic algorithm; see Theorem 10 below.

More formally, our algorithm SeqSEW$^*_*$ is defined as follows. The set of all time rounds $t = 1,2,\ldots$ is partitioned into regimes $r = 0,1,\ldots$ whose final time instances t_r are data-driven. Let $t_{-1} \triangleq 0$ by convention. We call *regime* r, $r = 0,1,\ldots$, the sequence of time rounds $(t_{r-1}+1,\ldots,t_r)$ where t_r is the first date $t \geqslant t_{r-1}+1$ such that $\gamma_t > 2^r$. At the beginning of regime r, we restart the algorithm SeqSEW$^*_\tau$ defined in Figure 3 with the parameter τ set to $\tau_r \triangleq 1/\big(\exp(2^r)-1\big)$.

In particular, on each regime r, the current instance of the algorithm SeqSEW$^*_{\tau_r}$ only uses the past observations y_s, $s \in \{t_{r-1}+1,\ldots,t-1\}$, to perform the online trunction and to tune the inverse temperature parameter. Therefore, the algorithm SeqSEW$^*_*$ is fully automatic.

Theorem 10 *Without requiring any preliminary knowledge at the beginning of the prediction game,* SeqSEW$^*_*$ *satisfies, for all* $T \geqslant 1$ *and all* $(x_1,y_1),\ldots,(x_T,y_T) \in \mathcal{X}\times\mathbb{R}$,

$$\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 \leqslant \inf_{u\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t-u\cdot\varphi(x_t)\big)^2 + 128\Big(\max_{1\leqslant t\leqslant T}y_t^2\Big)\|u\|_0\ln\left(e+\sqrt{\sum_{t=1}^{T}\sum_{j=1}^{d}\varphi_j^2(x_t)}\right)\right.$$
$$\left. + 32\Big(\max_{1\leqslant t\leqslant T}y_t^2\Big)A_T\|u\|_0\ln\left(1+\frac{\|u\|_1}{\|u\|_0}\right)\right\} + \Big(1+9\max_{1\leqslant t\leqslant T}y_t^2\Big)A_T\ ,$$

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

Though the algorithm SeqSEW$^*_*$ is fully automatic, two possible improvements could be addressed in the future. From a theoretical viewpoint, can we contruct a fully automatic algorithm with a bound similar to Theorem 10 but without the extra logarithmic factor A_T? From a practical viewpoint, is it possible to perform the adaptation to B_Φ without restarting the algorithm repeatedly (just like we did for B_y)? A smoother time-varying tuning $(\tau_t)_{t\geqslant 2}$ might enable to answer both questions. This would be very probably at the price of a more involved analysis (e.g., if we adapt the PAC-Bayesian bound of Lemma 9, then a third approximation term would appear in (14) since π_{τ_t} changes over time).

Proof sketch (of Theorem 10) The proof relies on the use of Corollary 7 on all regimes r visited up to time T. More precisely, note that $\gamma_{t_r-1} \leqslant 2^r$ by definition of t_r (except maybe in the trivial case when $t_r = t_{r-1}+1$), which entails that

$$\sum_{t=t_{r-1}+1}^{t_r-1}\sum_{j=1}^{d}\varphi_j^2(x_t) \leqslant \Big(e^{2^r}-1\Big)^2 \triangleq B_{\Phi,r}\ .$$

Since we tuned the instance of the algorithm SeqSEW$^*_\tau$ on regime r with $\tau = \tau_r \triangleq 1/\sqrt{B_{\Phi,r}}$, we can apply Corollary 7 on regime r for all r. Summing the corresponding regret bounds over r then yields the desired result. See Appendix A.1 for a detailed proof. ■

Theorem 10 yields the following corollary. It upper bounds the regret of the algorithm SeqSEW$^*_*$ uniformly over all $u \in \mathbb{R}^d$ such that $\|u\|_0 \leqslant s$ and $\|u\|_1 \leqslant U$, where the sparsity level $s \in \mathbb{N}$ and the ℓ^1-diameter $U > 0$ are both unknown to the forecaster. The proof is postponed to Appendix A.1.

Corollary 11 *Fix $s \in \mathbb{N}$ and $U > 0$. Then, for all $T \geqslant 1$ and all $(x_1,y_1),\dots,(x_T,y_T) \in \mathcal{X} \times \mathbb{R}$, the regret of the algorithm* SeqSEW$^*_*$ *on $\big\{u \in \mathbb{R}^d : \|u\|_0 \leqslant s \text{ and } \|u\|_1 \leqslant U\big\}$ is bounded by*

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \inf_{\substack{\|u\|_0 \leqslant s \\ \|u\|_1 \leqslant U}} \sum_{t=1}^T \big(y_t - u \cdot \varphi(x_t)\big)^2$$
$$\leqslant 128 \Big(\max_{1\leqslant t\leqslant T} y_t^2\Big)\, s \ln\left(e + \sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)}\right) + 32\Big(\max_{1\leqslant t\leqslant T} y_t^2\Big) A_T\, s \ln\left(1 + \frac{U}{s}\right)$$
$$+ \Big(1 + 9 \max_{1\leqslant t\leqslant T} y_t^2\Big) A_T \,,$$

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

4. Adaptivity to the Unknown Variance in the Stochastic Setting

In this section, we apply the online algorithm SeqSEW$^*_\tau$ of Section 3.2 to two related stochastic settings: the regression model with random design (Section 4.1) and the regression model with fixed design (Section 4.2). The sparsity regret bounds proved for this algorithm on individual sequences imply in both settings sparsity oracle inequalities with leading constant 1. These risk bounds are of the same flavor as in Dalalyan and Tsybakov (2008, 2012a) but they are adaptive (up to a logarithmic factor) to the unknown variance σ^2 of the noise if the latter is Gaussian. In particular, we solve two questions left open by Dalalyan and Tsybakov (2012a) in the random design case.

In the sequel, just like in the online deterministic setting, we assume that the forecaster has access to a dictionary $\varphi = (\varphi_1,\dots,\varphi_d)$ of measurable base forecasters $\varphi_j : \mathcal{X} \to \mathbb{R}$, $j = 1,\dots,d$.

4.1 Regression Model With Random Design

In this section we apply the algorithm SeqSEW$^*_\tau$ to the regression model with random design. In this batch setting the forecaster is given at the beginning of the game T independent random copies $(X_1,Y_1),\dots,(X_T,Y_T)$ of $(X,Y) \in \mathcal{X} \times \mathbb{R}$ whose common distribution is unknown. We assume thereafter that $\mathbb{E}[Y^2] < \infty$; the goal of the forecaster is to estimate the regression function $f : \mathcal{X} \to \mathbb{R}$ defined by $f(x) \triangleq \mathbb{E}[Y|X = x]$ for all $x \in \mathcal{X}$. Setting $\varepsilon_t \triangleq Y_t - f(X_t)$ for all $t = 1,\dots,T$, note that

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

and that the pairs $(X_1,\varepsilon_1),\dots,(X_T,\varepsilon_T)$ are i.i.d. and such that $\mathbb{E}[\varepsilon_1^2] < \infty$ and $\mathbb{E}[\varepsilon_1|X_1] = 0$ almost surely. In the sequel, we denote the distribution of X by P^X and we set, for all measurable functions

$h : \mathcal{X} \to \mathbb{R}$,

$$\|h\|_{L^2} \triangleq \left(\int_{\mathcal{X}} h(x)^2 P^X(\mathrm{d}x) \right)^{1/2} = \left(\mathbb{E}\big[h(X)^2\big] \right)^{1/2} .$$

Next we construct an estimator $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ based on the sample $(X_1,Y_1),\dots,(X_T,Y_T)$ that satisfies a sparsity oracle inequality, that is, its expected L^2-risk $\mathbb{E}\big[\,\big\|\, f - \widehat{f}_T \,\big\|_{L^2}^2\big]$ is almost as small as the smallest L^2-risk $\|f - u\cdot\varphi\|_{L^2}^2$, $u \in \mathbb{R}^d$, up to some additive term proportional to $\|u\|_0$.

4.1.1 Algorithm and Main Result

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$^*_\tau$ of Section 3.2 from time 1 to time T with $\tau = 1/\sqrt{dT}$ (note that T is known in this setting). Using the standard online-to-batch conversion (see, e.g., Littlestone 1989; Cesa-Bianchi et al. 2004), we define our estimator $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ as the uniform average

$$\widehat{f}_T \triangleq \frac{1}{T} \sum_{t=1}^{T} \widetilde{f}_t \tag{25}$$

of the estimators $\widetilde{f}_t : \mathcal{X} \to \mathbb{R}$ sequentially built by the algorithm SeqSEW$^*_\tau$ as

$$\widetilde{f}_t(x) \triangleq \int_{\mathbb{R}^d} \big[u\cdot\varphi(x)\big]_{B_t} \, p_t(\mathrm{d}u) \; . \tag{26}$$

Note that, contrary to much prior work from the statistics community such as those of Catoni (2004), Bunea and Nobel (2008) and Dalalyan and Tsybakov (2012a), the estimators $\widetilde{f}_t : \mathcal{X} \to \mathbb{R}$ are tuned online. Therefore, $\widehat{f}_T$ does not depend on any prior knowledge on the unknown distribution of the (X_t,Y_t), $1 \leqslant t \leqslant T$, such as the unknown variance $\mathbb{E}\big[(Y - f(X))^2\big]$ of the noise, the norms $\big\|\varphi_j\big\|_\infty$, or the norms $\big\|\, f - \varphi_j \,\big\|_\infty$ (actually, the functions φ_j and $f - \varphi_j$ do not even need to be bounded in ℓ^∞-norm).

In this respect, this work improves on that of Bunea and Nobel (2008) who tune their online forecasters as a function of $\|f\|_\infty$ and $\sup_{u\in\mathcal{U}} \|u\cdot\varphi\|_\infty$, where $\mathcal{U} \subset \mathbb{R}^d$ is a bounded comparison set.[12] Their technique is not appropriate when $\|f\|_\infty$ is unknown and it cannot be extended to the case where $\mathcal{U} = \mathbb{R}^d$ (since $\sup_{u\in\mathbb{R}^d} \|u\cdot\varphi\|_\infty = +\infty$ if $\varphi \neq \mathbf{0}$). The major technical difference is that we truncate the base forecasts $u\cdot\varphi(X_t)$ instead of truncating the observations Y_t. In particular, this enables us to aggregate the base forecasters $u\cdot\varphi$ for all $u \in \mathbb{R}^d$, that is, over the whole $\mathbb{R}^d$ space.

The next sparsity oracle inequality is the main result of this section. It follows from the deterministic regret bound of Corollory 8 and from Jensen's inequality. Two corollaries are to be derived later.

Theorem 12 *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 $\big\|\varphi_j\big\|_{L^2}^2 \triangleq \mathbb{E}[\varphi_j(X)^2] < +\infty$ for all $j = 1,\dots,d$. Then, the estimator*

12. Bunea and Nobel (2008) study the case where $\mathcal{U}$ is the (scaled) simplex in $\mathbb{R}^d$ or the set of its vertices.

$\widehat{f}_T$ defined in (25)-(26) *satisfies*

$$\mathbb{E}\left[\left\|f-\widehat{f}_T\right\|_{L^2}^2\right] \leqslant \inf_{u\in\mathbb{R}^d}\left\{\|f-u\cdot\varphi\|_{L^2}^2 + 32\,\frac{\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]}{T}\,\|u\|_0 \ln\left(1+\frac{\sqrt{dT}\,\|u\|_1}{\|u\|_0}\right)\right\} + \frac{1}{dT}\sum_{j=1}^d \left\|\varphi_j\right\|_{L^2}^2 + 5\,\frac{\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]}{T}\,.$$

Proof sketch (of Theorem 12) By Corollary 8 and by definition of $\widetilde{f}_t$ above and $\widehat{y}_t \triangleq \widetilde{f}_t(X_t)$ in Figure 3, we have, *almost surely*,

$$\sum_{t=1}^T (Y_t - \widetilde{f}_t(X_t))^2 \leqslant \inf_{u\in\mathbb{R}^d}\left\{\sum_{t=1}^T \big(Y_t - u\cdot\varphi(X_t)\big)^2 + 32\left(\max_{1\leqslant t\leqslant T} Y_t^2\right)\|u\|_0 \ln\left(1+\frac{\sqrt{dT}\,\|u\|_1}{\|u\|_0}\right)\right\} + \frac{1}{dT}\sum_{j=1}^d\sum_{t=1}^T \varphi_j^2(X_t) + 5\max_{1\leqslant t\leqslant T} Y_t^2\,.$$

Taking the expectations of both sides and applying Jensen's inequality yields the desired result. For a detailed proof, see Appendix A.2. ■

Theorem 12 above can be used under several assumptions on the distribution of the output Y. In all cases, it suffices to upper bound the amplitude $\mathbb{E}\big[\max_{1\leqslant t\leqslant T} Y_t^2\big]$. We present below a general corollary and explain later why our fully automatic procedure $\widehat{f}_T$ solves two questions left open by Dalalyan and Tsybakov (2012a) (see Corollary 14 below).

4.1.2 A GENERAL COROLLARY

The next sparsity oracle inequality follows from Theorem 12 and from the upper bounds on $\mathbb{E}\big[\max_{1\leqslant t\leqslant T} Y_t^2\big]$ entailed by Lemmas 24–26 in Appendix B. The proof is postponed to Appendix A.2.

Corollary 13 *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 $\sup_{1\leqslant j\leqslant d}\left\|\varphi_j\right\|_{L^2}^2 < +\infty$, that $\mathbb{E}|Y| < +\infty$, and that one of the following assumptions holds on the distribution of $\Delta Y \triangleq Y - \mathbb{E}[Y]$.*

- $\big(\mathrm{BD}(B)\big)$: *$|\Delta Y| \leqslant B$ almost surely for a given constant $B > 0$;*
- $\big(\mathrm{SG}(\sigma^2)\big)$: *ΔY is subgaussian with variance factor $\sigma^2 > 0$, that is, $\mathbb{E}\left[e^{\lambda\Delta Y}\right] \leqslant e^{\lambda^2\sigma^2/2}$ for all $\lambda\in\mathbb{R}$;*
- $\big(\mathrm{BEM}(\alpha,M)\big)$: *ΔY has a bounded exponential moment, that is, $\mathbb{E}\left[e^{\alpha|\Delta Y|}\right] \leqslant M$ for some given constants $\alpha > 0$ and $M > 0$;*
- $\big(\mathrm{BM}(\alpha,M)\big)$: *ΔY has a bounded moment, that is, $\mathbb{E}\big[|\Delta Y|^\alpha\big] \leqslant M$ for some given constants $\alpha > 2$ and $M > 0$.*

Then, the estimator $\widehat{f}_T$ defined above satisfies

$$\mathbb{E}\left[\left\|f-\widehat{f}_T\right\|_{L^2}^2\right] \leqslant \inf_{u\in\mathbb{R}^d}\left\{\|f-u\cdot\varphi\|_{L^2}^2 + 64\left(\frac{\mathbb{E}[Y]^2}{T}+\psi_T\right)\|u\|_0 \ln\left(1+\frac{\sqrt{dT}\,\|u\|_1}{\|u\|_0}\right)\right\} + \frac{1}{dT}\sum_{j=1}^d \left\|\varphi_j\right\|_{L^2}^2 + 10\left(\frac{\mathbb{E}[Y]^2}{T}+\psi_T\right),$$

where

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

Several comments can be made about Corollary 13. We first stress that, if $T \geqslant 2$, then the two "bias" terms $\mathbb{E}[Y]^2/T$ above can be avoided, at least at the price of a multiplicative factor of $2T/(T-1) \leqslant 4$. This can be achieved via a slightly more sophisticated online clipping—see Remark 19 in Appendix A.2.

Second, under the assumptions $\big(\mathrm{BD}(\mathrm{B})\big)$, $\big(\mathrm{SG}(\sigma^2)\big)$, or $\big(\mathrm{BEM}(\alpha,\mathrm{M})\big)$, the key quantity ψ_T is respectively of the order of $1/T$, $\ln(T)/T$ and $\ln^2(T)/T$. Up to a logarithmic factor, this corresponds to the classical fast rate of convergence $1/T$ obtained in the random design setting for different aggregation problems (see, e.g., Catoni 1999; Juditsky et al. 2008; Audibert 2009 for model-selection-type aggregation and Dalalyan and Tsybakov 2012a for linear aggregation). We were able to get similar rates—with, however, a fully automatic procedure—since our online algorithm SeqSEW^*_τ is well suited for bounded individual sequences with an unknown bound. More precisely, the finite i.i.d. sequence $Y_1,\dots,Y_T$ is almost surely uniformly bounded by the random bound $\max_{1\leqslant t\leqslant T}|Y_t|$. Our individual sequence techniques adapt sequentially to this random bound, yielding a regret bound that scales as $\max_{1\leqslant t\leqslant T} Y_t^2$. As a result, the risk bounds obtained after the online-to-batch conversion scale as $\mathbb{E}\big[\max_{1\leqslant t\leqslant T} Y_t^2\big]/T$. If the distribution of the output Y is sufficiently lightly-tailed—which includes the quite general bounded-exponential-moment assumption—then we can recover the fast rate of convergence $1/T$ up to a logarithmic factor.

We note that there is still a question left open for heavy-tailed output distributions. For example, under the bounded moment assumption $\big(\mathrm{BM}(\alpha,\mathrm{M})\big)$, the rate $T^{-(\alpha-2)/\alpha}$ that we proved does not match the faster rate $T^{-\alpha/(\alpha+2)}$ obtained by Juditsky et al. (2008) and Audibert (2009) under a similar assumption. Their methods use some preliminary knowledge on the output distribution (such as the exponent α). Thus, obtaining the same rate with a procedure tuned in an automatic fashion—just like our method $\widehat{f}_T$—is a challenging task. For this purpose, a different tuning of η_t or a more sophisticated online truncation might be necessary.

Third, several variations on the assumptions are possible. First note that several classical assumptions on Y expressed in terms of $f(X)$ and $\varepsilon \triangleq Y - f(X)$ 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 20 in Appendix A.2 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 20 in Appendix A.2, under the classical assumptions

$$\|f\|_\infty < +\infty \qquad \text{and} \qquad \mathbb{E}\Big[e^{\alpha|\varepsilon|} \,\Big|\, X\Big] \leqslant M \quad \text{a.s.} \tag{27}$$

or

$$\|f\|_\infty < +\infty \qquad \text{and} \qquad \mathbb{E}\Big[e^{\lambda\varepsilon} \,\Big|\, X\Big] \leqslant e^{\lambda^2\sigma^2/2} \quad \text{a.s.}, \quad \forall \lambda \in \mathbb{R}\,, \tag{28}$$

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

$$\psi_T \leqslant \begin{cases} \dfrac{8\,\|f\|_\infty^2}{T} + \dfrac{2\ln^2\big((M+e)T\big)}{\alpha^2\,T} & \text{under the set of assumptions (27),} \\ \dfrac{8\,\|f\|_\infty^2}{T} + \dfrac{4\sigma^2\ln(2eT)}{T} & \text{under the set of assumptions (28).} \end{cases}$$

In particular, under the set of assumptions (28), our procedure $\widehat{f}_T$ solves two questions left open by Dalalyan and Tsybakov (2012a). We discuss below our contributions in this particular case.

4.1.3 QUESTIONS LEFT OPEN BY DALALYAN AND TSYBAKOV

In this subsection we focus on the case when the set of assumptions (28) holds true. Namely, the regression function f is bounded (by an unknown constant) and the noise $\varepsilon \triangleq Y - f(X)$ is subgaussian conditionally on X with an unknown variance factor $\sigma^2 > 0$. An important particular case is when $\|f\|_\infty < +\infty$ and when the noise ε is independent of X and normally distributed $\mathcal{N}(0,\sigma^2)$.

Under the set of assumptions (28), the two terms $\mathbb{E}\big[\max_{1\leqslant t\leqslant T} Y_t^2\big]$ of Theorem 12 can be upper bounded in a simpler and slightly tighter way as compared to the proof of Corollary 13 (we only use the inequality $(x+y)^2 \leqslant 2x^2 + 2y^2$ once, instead of twice). It yields the following sparsity oracle inequality.

Corollary 14 *Assume that $(X_1,Y_1),\dots,(X_T,Y_T) \in \mathcal{X}\times\mathbb{R}$ are independent random copies of $(X,Y) \in \mathcal{X}\times\mathbb{R}$ such that the set of assumptions* (28) *above holds true. Then, the estimator $\widehat{f}_T$ defined in* (25)-(26) *satisfies*

$$\begin{aligned} &\mathbb{E}\bigg[\Big\|f - \widehat{f}_T\Big\|_{L^2}^2\bigg] \\ &\quad\leqslant \inf_{u\in\mathbb{R}^d}\left\{\|f - u\cdot\varphi\|_{L^2}^2 + 64\Big(\|f\|_\infty^2 + 2\sigma^2\ln(2eT)\Big)\,\frac{\|u\|_0}{T}\ln\left(1 + \frac{\sqrt{dT}\,\|u\|_1}{\|u\|_0}\right)\right\} \\ &\qquad + \frac{1}{dT}\sum_{j=1}^d \big\|\varphi_j\big\|_{L^2}^2 + \frac{10}{T}\Big(\|f\|_\infty^2 + 2\sigma^2\ln(2eT)\Big)\,. \end{aligned}$$

Proof We apply Theorem 12 and bound $\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]$ from above. By the elementary inequality $(x+y)^2 \leqslant 2x^2+2y^2$ for all $x,y\in\mathbb{R}$, we get

$$\begin{aligned}\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right] = \mathbb{E}\left[\max_{1\leqslant t\leqslant T} \big(f(X_t)+\varepsilon_t\big)^2\right] &\leqslant 2\left(\|f\|_\infty^2 + \mathbb{E}\left[\max_{1\leqslant t\leqslant T}\varepsilon_t^2\right]\right)\\ &\leqslant 2\left(\|f\|_\infty^2 + 2\sigma^2\ln(2eT)\right)\,,\end{aligned}$$

where the last inequality follows from Lemma 24 in Appendix B and from the fact that, for all $1\leqslant t\leqslant T$ and all $\lambda\in\mathbb{R}$, we have $\mathbb{E}\big[e^{\lambda\varepsilon_t}\big] = \mathbb{E}\big[e^{\lambda\varepsilon}\big] = \mathbb{E}\big[\mathbb{E}\big[e^{\lambda\varepsilon}\,\big|\,X\big]\big] \leqslant e^{\lambda^2\sigma^2/2}$ by (28). (Note that the assumption of conditional subgaussianity in (28) is stronger than what we need, that is, subgaussianity without conditioning.) This concludes the proof. ■

The above bound is of the same order (up to a $\ln T$ factor) as the sparsity oracle inequality proved in Proposition 1 of Dalalyan and Tsybakov (2012a). For the sake of comparison we state below with our notations (e.g., β therein corresponds to $1/\eta$ in this paper) a straightforward consequence of this proposition, which follows by Jensen's inequality and the particular[13] choice $\tau = \min\big\{1/\sqrt{dT}, R/(4d)\big\}$.

Proposition 15 (A consequence of Prop. 1 of Dalalyan and Tsybakov 2012a)
Assume that $\sup_{1\leqslant j\leqslant d}\|\varphi_j\|_\infty < \infty$ *and that the set of assumptions (28) above holds true. Then, for all* $R>0$ *and all* $\eta\leqslant\bar\eta(R)\triangleq\big(2\sigma^2+2\sup_{\|u\|_1\leqslant R}\|u\cdot\varphi-f\|_\infty^2\big)^{-1}$*, the mirror averaging aggregate* $\widehat{f}_T:\mathcal{X}\to\mathbb{R}$ *defined by Dalalyan and Tsybakov (2012a, Equations (1) and (3)) with* $\tau = \min\big\{1/\sqrt{dT}, R/(4d)\big\}$ *satisfies*

$$\begin{aligned}\mathbb{E}\left[\left\|f-\widehat{f}_T\right\|_{L^2}^2\right] \leqslant \inf_{\|u\|_1\leqslant R/2}&\left\{\|f-u\cdot\varphi\|_{L^2}^2 + \frac{4}{\eta}\frac{\|u\|_0}{T+1}\ln\left(1+\frac{\sqrt{dT}\,\|u\|_1+2d}{\|u\|_0}\right)\right\}\\ &+\frac{4}{dT}\sum_{j=1}^d\|\varphi_j\|_{L^2}^2 + \frac{1}{(T+1)\eta}\,.\end{aligned}$$

We can now discuss the two questions left open by Dalalyan and Tsybakov (2012a).
Risk bound on the whole $\mathbb{R}^d$ *space.* Despite the similarity of the two bounds, the sparsity oracle inequality stated in Proposition 15 above only holds for vectors u within an ℓ^1-ball of finite radius $R/2$, while our bound holds over the whole $\mathbb{R}^d$ space. Moreover, the parameter R above has to be chosen in advance, but it cannot be chosen too large since $1/\eta\geqslant 1/\bar\eta(R)$, which grows as R^2 when $R\to+\infty$ (if $\varphi\neq\mathbf{0}$). Dalalyan and Tsybakov (2012a, Section 4.2) thus asked whether it was possible to get a bound with $1/\eta<+\infty$ such that the infimum in Proposition 15 extends to the whole $\mathbb{R}^d$ space. Our results show that, thanks to data-driven truncation, the answer is positive.

Note that it is still possible to transform the bound of Proposition 15 into a bound over the whole $\mathbb{R}^d$ space if the parameter R is chosen (illegally) as $R=2\|u^*\|_1$ (or as a tight upper bound of the last

13. Proposition 1 of Dalalyan and Tsybakov (2012a) may seem more general than Corollary 14 at first sight since it holds for all $\tau>0$, but this is actually also the case for Corollary 14. The proof of the latter would indeed have remained true had we replaced $\tau=1/\sqrt{dT}$ with any value of $\tau>0$ (see Proposition 5). We however chose the reasonable value $\tau=1/\sqrt{dT}$ to make our algorithm parameter-free. As noted earlier, if $\|\varphi\|_\infty\triangleq\sup_{x\in\mathcal{X}}\max_{1\leqslant j\leqslant d}|\varphi_j(x)|$ is finite and known by the forecaster, another simple and easy-to-analyse tuning is given by $\tau=1/(\|\varphi\|_\infty\sqrt{dT})$.

quantity), where $u^* \in \mathbb{R}^d$ minimizes over $\mathbb{R}^d$ the regularized risk

$$\|f - u\cdot\varphi\|_{L^2}^2 + \frac{4}{\bar\eta(2\|u\|_1)}\,\frac{\|u\|_0}{T+1}\ln\left(1+\frac{\sqrt{dT}\,\|u\|_1+2d}{\|u\|_0}\right) + \frac{4}{dT}\sum_{j=1}^d \big\|\varphi_j\big\|_{L^2}^2 + \frac{1}{(T+1)\,\bar\eta(2\|u\|_1)}\,.$$

For instance, choosing $R = 2\|u^*\|_1$ and $\eta = \bar\eta(R)$, we get from Proposition 15 that the expected L^2-risk $\mathbb{E}\big[\|f-\widehat{f}_T\|_{L^2}^2\big]$ of the corresponding procedure is upper bounded by the infimum of the above regularized risk over all $u\in\mathbb{R}^d$. However, this parameter tuning is illegal since $\|u^*\|_1$ is not known in practice. On the contrary, thanks to data-driven truncation, the prior knowledge of $\|u^*\|_1$ is not required by our procedure.

Adaptivity to the unknown variance of the noise. The second open question, which was raised by Dalalyan and Tsybakov (2012a, Section 5.1, Remark 6), 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. Since the noise level σ^2 is unknown in practice, the authors asked the important question whether adaptivity to σ^2 was possible. Up to a $\ln T$ factor, Corollary 14 above provides a positive answer.

4.2 Regression Model With Fixed Design

In this section, we consider the regression model with fixed design. In this batch setting the forecaster is given at the beginning of the game a T-sample $(x_1,Y_1),\dots,(x_T,Y_T)\in\mathcal{X}\times\mathbb{R}$, where the x_t are deterministic elements in $\mathcal{X}$ and where

$$Y_t = f(x_t)+\varepsilon_t\,,\quad 1\leqslant t\leqslant T, \tag{29}$$

for some i.i.d. sequence $\varepsilon_1,\dots,\varepsilon_T\in\mathbb{R}$ (with unknown distribution) and some unknown function $f:\mathcal{X}\to\mathbb{R}$. Next we construct an estimator $\widehat{f}_T:\mathcal{X}\to\mathbb{R}$ of f based on the sample $(x_1,Y_1),\dots,(x_T,Y_T)$ that satisfies a sparsity oracle inequality, that is, its expected mean squared error $\mathbb{E}\big[\frac{1}{T}\sum_{t=1}^T(f(x_t)-\widehat{f}_T(x_t))^2\big]$ is almost as small as the smallest mean squared error $\frac{1}{T}\sum_{t=1}^T(f(x_t)-u\cdot\varphi(x_t))^2$, $u\in\mathbb{R}^d$, up to some additive term proportional to $\|u\|_0$.

In this setting, just like in Section 4.1, our algorithm and the corresponding analysis are a straightforward consequence of the general results on individual sequences developed in Section 3. As in the random design setting, the sample $(x_1,Y_1),\dots,(x_T,Y_T)$ is treated in a sequential fashion. We run the algorithm SeqSEW$_\tau^*$ defined in Figure 3 from time 1 to time T with the particular choice of $\tau = 1/\sqrt{dT}$. We then define our estimator $\widehat{f}_T:\mathcal{X}\to\mathbb{R}$ by

$$\widehat{f}_T(x) \triangleq \begin{cases} \dfrac{1}{n_x}\displaystyle\sum_{\substack{1\leqslant t\leqslant T\\ t:x_t=x}} \widetilde{f}_t(x) & \text{if } x\in\{x_1,\dots,x_T\}\,,\\[2ex] 0 & \text{if } x\notin\{x_1,\dots,x_T\}\,, \end{cases} \tag{30}$$

where $n_x \triangleq \big|\big\{t : x_t = x\big\}\big| = \sum_{t=1}^T \mathbb{I}_{\{x_t=x\}}$, and where the estimators $\widetilde{f}_t:\mathcal{X}\to\mathbb{R}$ sequentially built by the algorithm SeqSEW$_\tau^*$ are defined by

$$\widetilde{f}_t(x) \triangleq \int_{\mathbb{R}^d} \big[u\cdot\varphi(x)\big]_{B_t}\,p_t(\mathrm{d}u)\,. \tag{31}$$

In the particular case when the x_t are all distinct, $\widehat{f}_T$ is simply defined by $\widehat{f}_T(x_t) \triangleq \widetilde{f}_t(x_t)$ for all $t \in \{1,\dots,T\}$ and by $\widehat{f}_T(x) = 0$ otherwise. Therefore, in this case, $\widehat{f}_T$ only uses the observations $y_1,\dots,y_{t-1}$ to estimate $f(x_t)$ (in particular, $\widehat{f}_T(x_1)$ is deterministic).

The next theorem is the main result of this subsection. It follows as in the random design setting from the deterministic regret bound of Corollory 8 and from Jensen's inequality. The proof is postponed to Appendix A.3.

Theorem 16 *Consider the regression model with fixed design described in* (29)*. Then, the estimator* $\widehat{f}_T$ *defined in* (30)–(31) *satisfies*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t)-\widehat{f}_T(x_t)\big)^2\right] \leqslant \inf_{u\in\mathbb{R}^d}\left\{\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t)-u\cdot\varphi(x_t)\big)^2 + 32\,\frac{\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]}{T}\,\|u\|_0\ln\left(1+\frac{\sqrt{dT}\,\|u\|_1}{\|u\|_0}\right)\right\} + \frac{1}{dT^2}\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t) + 5\,\frac{\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]}{T}\,.$$

As in Section 4.1, the amplitude $\mathbb{E}\big[\max_{1\leqslant t\leqslant T} Y_t^2\big]$ can be upper bounded under various assumptions. The proof of the following corollary is postponed to Appendix A.3.

Corollary 17 *Consider the regression model with fixed design described in* (29)*. Assume that one of the following assumptions holds on the distribution of* ε_1*.*

- $\big(\mathrm{BD}(B)\big)$: $|\varepsilon_1| \leqslant B$ *almost surely for a given constant* $B > 0$*;*
- $\big(\mathrm{SG}(\sigma^2)\big)$: ε_1 *is subgaussian with variance factor* $\sigma^2 > 0$*, that is,* $\mathbb{E}\left[e^{\lambda\varepsilon_1}\right] \leqslant e^{\lambda^2\sigma^2/2}$ *for all* $\lambda \in \mathbb{R}$*;*
- $\big(\mathrm{BEM}(\alpha,M)\big)$: ε *has a bounded exponential moment, that is,* $\mathbb{E}\left[e^{\alpha|\varepsilon|}\right] \leqslant M$ *for some given constants* $\alpha > 0$ *and* $M > 0$*;*
- $\big(\mathrm{BM}(\alpha,M)\big)$: ε *has a bounded moment, that is,* $\mathbb{E}\big[|\varepsilon|^\alpha\big] \leqslant M$ *for some given constants* $\alpha > 2$ *and* $M > 0$*.*

Then, the estimator $\widehat{f}_T$ *defined in* (30)–(31) *satisfies*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t)-\widehat{f}_T(x_t)\big)^2\right] \leqslant \inf_{u\in\mathbb{R}^d}\left\{\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t)-u\cdot\varphi(x_t)\big)^2 + 64\left(\frac{\max_{1\leqslant t\leqslant T} f^2(x_t)}{T}+\psi_T\right)\|u\|_0\ln\left(1+\frac{\sqrt{dT}\,\|u\|_1}{\|u\|_0}\right)\right\} + \frac{1}{dT^2}\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t) + 10\left(\frac{\max_{1\leqslant t\leqslant T} f^2(x_t)}{T}+\psi_T\right)\,,$$

where

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

The above bound is of the same flavor as that of Dalalyan and Tsybakov (2008, Theorem 5). It has one advantage and one drawback. On the one hand, we note two additional "bias" terms $\big(\max_{1\leqslant t\leqslant T} f^2(x_t)\big)/T$ as compared to the bound of Dalalyan and Tsybakov (2008, Theorem 5). As of now, we have not been able to remove them using ideas similar to what we did in the random design case (see Remark 19 in Appendix A.2). On the other hand, under Assumption $\big(\mathrm{SG}(\sigma^2)\big)$, contrary to Dalalyan and Tsybakov (2008), our algorithm does not require the prior knowledge of the variance factor σ^2 of the noise.

Acknowledgments

The author would like to thank Arnak Dalalyan, Gilles Stoltz, and Pascal Massart for their helpful feedback and suggestions, as well as two anonymous reviewers for their insightful comments, one of which helped us simplify the online tuning carried out in Section 3.2. 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.

Appendix A. Proofs

In this appendix we provide the proofs of some results stated above.

A.1 Proofs of Theorem 10 and Corollary 11

Before proving Theorem 10, we first need the following comment. Since the algorithm SeqSEW$^*_\tau$ is restarted at the beginning of each regime, the threshold values B_t used on regime r by the algorithm SeqSEW$^*_\tau$ are not computed on the basis of all past observations $y_1,\dots,y_{t-1}$ but only on the basis of the past observations y_t, $t\in\{t_{r-1}+1,\dots,t-1\}$. To avoid any ambiguity, we set $B_{r,t_{r-1}+1}\triangleq 0$ and

$$B_{r,t}\triangleq \max_{t_{r-1}+1\leqslant s\leqslant t-1}|y_s|\,,\quad t\in\{t_{r-1}+2,\dots,t_r\}\,.$$

Proof (of Theorem 10) We denote by $R\triangleq\min\{r\in\mathbb{N}:T\leqslant t_r\}$ the index of the last regime. For notational convenience, we re-define $t_R\triangleq T$ (even if $\gamma_T\leqslant 2^R$).

We upper bound the regret of the algorithm SeqSEW$^*_*$ on $\{1,\dots,T\}$ by the sum of its regrets on each time interval. To do so, first note that[14]

$$\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2=\sum_{r=0}^{R}\sum_{t=t_{r-1}+1}^{t_r}(y_t-\widehat{y}_t)^2=\sum_{r=0}^{R}\left((y_{t_r}-\widehat{y}_{t_r})^2+\sum_{t=t_{r-1}+1}^{t_r-1}(y_t-\widehat{y}_t)^2\right)$$

$$\leqslant\sum_{r=0}^{R}\left(2(y_{t_r}^2+B_{r,t_r}^2)+\sum_{t=t_{r-1}+1}^{t_r-1}(y_t-\widehat{y}_t)^2\right) \tag{32}$$

$$\leqslant\sum_{r=0}^{R}\left(\sum_{t=t_{r-1}+1}^{t_r-1}(y_t-\widehat{y}_t)^2\right)+4(R+1){y_T^*}^2\,, \tag{33}$$

where we set $y_T^*\triangleq\max_{1\leqslant t\leqslant T}|y_t|$, where (32) follows from the upper bound $(y_{t_r}-\widehat{y}_{t_r})^2\leqslant 2(y_{t_r}^2+\widehat{y}_{t_r}^2)\leqslant 2(y_{t_r}^2+B_{r,t_r}^2)$ (since $|\widehat{y}_{t_r}|\leqslant B_{r,t_r}$ by construction), and where (33) follows from the inequalities $y_{t_r}^2\leqslant {y_T^*}^2$ and

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

But, for every $r=0,\dots,R$, the trace of the empirical Gram matrix on $\{t_{r-1}+1,\dots,t_r-1\}$ is upper bounded by

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

where the last inequality follows from the fact that $\gamma_{t_r-1}\leqslant 2^r$ (by definition of t_r). Since in addition $\tau_r\triangleq 1/\sqrt{(e^{2^r}-1)^2}$, we can apply Corollory 7 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 (33) the upper bound

$$\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2\leqslant\sum_{r=0}^{R}\inf_{u\in\mathbb{R}^d}\left\{\sum_{t=t_{r-1}+1}^{t_r-1}\big(y_t-u\cdot\varphi(x_t)\big)^2+\Delta_r(u)\right\}+4(R+1){y_T^*}^2\,, \tag{34}$$

where

$$\Delta_r(u)\triangleq 32B_{r,t_r}^2\,\|u\|_0\ln\left(1+\frac{\left(e^{2^r}-1\right)\|u\|_1}{\|u\|_0}\right)+5B_{r,t_r}^2+1\,. \tag{35}$$

Since the infimum is superadditive and since $\big(y_{t_r}-u\cdot\varphi(x_{t_r})\big)^2\geqslant 0$ for all $r=0,\dots,R$, we get from (34) that

$$\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2\leqslant\inf_{u\in\mathbb{R}^d}\sum_{r=0}^{R}\left(\sum_{t=t_{r-1}+1}^{t_r}\big(y_t-u\cdot\varphi(x_t)\big)^2+\Delta_r(u)\right)+4(R+1){y_T^*}^2$$

$$=\inf_{u\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t-u\cdot\varphi(x_t)\big)^2+\sum_{r=0}^{R}\Delta_r(u)\right\}+4(R+1){y_T^*}^2\,. \tag{36}$$

Let $u\in\mathbb{R}^d$. Next we bound $\sum_{r=0}^{R}\Delta_r(u)$ and $4(R+1){y_T^*}^2$ from above. First note that, by the upper bound $B_{r,t_r}^2\leqslant {y_T^*}^2$ and by the elementary inequality $\ln(1+xy)\leqslant\ln\big((1+x)(1+y)\big)=\ln(1+$

14. In the trivial cases where $t_r=t_{r-1}+1$ for some r, the sum $\sum_{t=t_{r-1}+1}^{t_r-1}(y_t-\widehat{y}_t)^2$ equals 0 by convention.

$x)+\ln(1+y)$ with $x=e^{2^r}-1$ and $y=\|u\|_1/\|u\|_0$, (35) yields

$$\Delta_r(u) \leqslant 32\,{y_T^*}^2\,\|u\|_0\,2^r + 32\,{y_T^*}^2\,\|u\|_0 \ln\left(1+\frac{\|u\|_1}{\|u\|_0}\right) + 5{y_T^*}^2 + 1\;.$$

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

$$\sum_{r=0}^{R}\Delta_r(u) \leqslant 32\left(2^{R+1}-1\right){y_T^*}^2\,\|u\|_0 + (R+1)\left(32\,{y_T^*}^2\,\|u\|_0\ln\left(1+\frac{\|u\|_1}{\|u\|_0}\right)+5{y_T^*}^2+1\right)\;. \qquad (37)$$

First case: $R=0$

Substituting (37) in (36), we conclude the proof by noting that $A_T \geqslant 2+\log_2 1 \geqslant 1$ and that $\ln\left(e+\sqrt{\sum_{t=1}^T\sum_{j=1}^d \varphi_j^2(x_t)}\right)\geqslant 1$.

Second case: $R\geqslant 1$

Since $R\geqslant 1$, we have, by definition of t_{R-1},

$$2^{R-1} < \gamma_{t_{R-1}} \triangleq \ln\left(1+\sqrt{\sum_{t=1}^{t_{R-1}}\sum_{j=1}^d \varphi_j^2(x_t)}\right) \leqslant \ln\left(e+\sqrt{\sum_{t=1}^{T}\sum_{j=1}^d \varphi_j^2(x_t)}\right)\;.$$

The last inequality entails that $2^{R+1}-1\leqslant 4\cdot 2^{R-1}\leqslant 4\ln\left(e+\sqrt{\sum_{t=1}^T\sum_{j=1}^d\varphi_j^2(x_t)}\right)$ and that $R+1\leqslant 2+\log_2\ln\left(e+\sqrt{\sum_{t=1}^T\sum_{j=1}^d\varphi_j^2(x_t)}\right)\triangleq A_T$. Therefore, one the one hand, via (37),

$$\sum_{r=0}^{R}\Delta_r(u) \leqslant 128\,{y_T^*}^2\,\|u\|_0\ln\left(e+\sqrt{\sum_{t=1}^T\sum_{j=1}^d\varphi_j^2(x_t)}\right) + 32\,{y_T^*}^2 A_T\,\|u\|_0\ln\left(1+\frac{\|u\|_1}{\|u\|_0}\right) + A_T\left(5{y_T^*}^2+1\right)\;,$$

and, on the other hand,

$$4(R+1){y_T^*}^2 \leqslant 4A_T\,{y_T^*}^2\;.$$

Substituting the last two inequalities in (36) and noting that ${y_T^*}^2=\max_{1\leqslant t\leqslant T}y_t^2$ concludes the proof. ■

Proof (of Corollary 11) The proof is straightforward. In view of Theorem 10, we just need to check that the quantity (continuously extended in $s=0$)

$$128\Big(\max_{1\leqslant t\leqslant T}y_t^2\Big)\,s\ln\left(e+\sqrt{\sum_{t=1}^T\sum_{j=1}^d\varphi_j^2(x_t)}\right) + 32\Big(\max_{1\leqslant t\leqslant T}y_t^2\Big)A_T\,s\ln\left(1+\frac{U}{s}\right)$$

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

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

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

From the elementary inequality

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

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

A.2 Proofs of Theorem 12 and Corollary 13

In this subsection, we set $\varepsilon \triangleq Y - f(X)$, so that the pairs $(X_1,\varepsilon_1),\ldots,(X_T,\varepsilon_T)$ are independent copies of $(X,\varepsilon) \in \mathcal{X}\times\mathbb{R}$. We also define $\sigma \geqslant 0$ by

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

Proof (of Theorem 12) By Corollory 8 and the definitions of $\widetilde{f}_t$ in (26) and $\widehat{y}_t \triangleq \widetilde{f}_t(X_t)$ in Figure 3, we have, *almost surely*,

$$\begin{aligned}\sum_{t=1}^T (Y_t - \widetilde{f}_t(X_t))^2 \leqslant \inf_{u\in\mathbb{R}^d}\Bigg\{ &\sum_{t=1}^T \big(Y_t - u\cdot\varphi(X_t)\big)^2 + 32\left(\max_{1\leqslant t\leqslant T} Y_t^2\right)\|u\|_0 \ln\left(1+\frac{\sqrt{dT}\,\|u\|_1}{\|u\|_0}\right)\Bigg\} \\ &+ \frac{1}{dT}\sum_{j=1}^d\sum_{t=1}^T \varphi_j^2(X_t) + 5\max_{1\leqslant t\leqslant T} Y_t^2\,.\end{aligned}$$

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

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

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

In the same way,

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

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

$$\begin{aligned}T\sigma^2 + \sum_{t=1}^T \mathbb{E}\left[\big(f(X_t) - \widetilde{f}_t(X_t)\big)^2\right] \leqslant \inf_{u\in\mathbb{R}^d}\Bigg\{ &T\sigma^2 + \sum_{t=1}^T \mathbb{E}\left[\big(f(X_t) - u\cdot\varphi(X_t)\big)^2\right] \\ &+ 32\,\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]\|u\|_0 \ln\left(1+\frac{\sqrt{dT}\,\|u\|_1}{\|u\|_0}\right)\Bigg\} \\ &+ \frac{1}{dT}\sum_{j=1}^d\sum_{t=1}^T \mathbb{E}\left[\varphi_j^2(X_t)\right] + 5\,\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]\,.\end{aligned}$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(f(X_t)-\widetilde{f}_t(X_t)\right)^2\right] \leqslant \inf_{u\in\mathbb{R}^d}\left\{\|f-u\cdot\varphi\|_{L^2}^2 + 32\,\frac{\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]}{T}\,\|u\|_0\,\ln\left(1+\frac{\sqrt{dT}\,\|u\|_1}{\|u\|_0}\right)\right\} + \frac{1}{dT}\sum_{j=1}^{d}\|\varphi_j\|_{L^2}^2 + 5\,\frac{\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]}{T}\,.$$

The right-hand side of the last inequality is exactly the upper bound stated in Theorem 12. To conclude the proof, we thus only need to check that $\mathbb{E}\big[\|f-\widehat{f}_T\|_{L^2}^2\big]$ is bounded from above by the left-hand side. But by definition of $\widehat{f}_T$ and by convexity of the square loss,

$$\mathbb{E}\left[\left\|f-\widehat{f}_T\right\|_{L^2}^2\right] \triangleq \mathbb{E}\left[\left(f(X)-\frac{1}{T}\sum_{t=1}^{T}\widetilde{f}_t(X)\right)^2\right] \leqslant \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(f(X)-\widetilde{f}_t(X)\right)^2\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(f(X_t)-\widetilde{f}_t(X_t)\right)^2\right]\,.$$

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

Remark 18 *The fact that the inequality stated in Corollary 8 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* $\mathbb{E}\left[\|f-\widehat{f}_T\|_{L^2}^2\right]$ *and* $\|f-u\cdot\varphi\|_{L^2}^2$. *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 13) We can apply Theorem 12. Then, to prove the upper bound on $\mathbb{E}\left[\|f-\widehat{f}_T\|_{L^2}^2\right]$, it suffices to show that

$$\frac{\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]}{T} \leqslant 2\left(\frac{\mathbb{E}[Y]^2}{T}+\psi_T\right)\,. \tag{38}$$

Recall that

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

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 $(x+y)^2\leqslant 2x^2+2y^2$ for all $x,y\in\mathbb{R}$, we have

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

Dividing both sides by T, we get (38).

As for the upper bound on ψ_T, since the $(\Delta Y)_t$, $1 \leqslant t \leqslant T$, are distributed as ΔY, we can apply Lemmas 24, 25, and 26 in Appendix B.3 to bound ψ_T from above under the assumptions $\big(\mathrm{SG}(\sigma^2)\big)$, $\big(\mathrm{BEM}(\alpha,\mathrm{M})\big)$, and $\big(\mathrm{BM}(\alpha,\mathrm{M})\big)$ respectively (the upper bound under $\big(\mathrm{BD}(\mathrm{B})\big)$ is straightforward):

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

■

Remark 19 *If $T \geqslant 2$, then the two "bias" terms $\mathbb{E}[Y]^2/T$ appearing in Corollary 13 can be avoided, at least at the price of a multiplicative factor of $2T/(T-1) \leqslant 4$. It suffices to use a slightly more sophisticated online clipping defined as follows. The first round $t=1$ is only used to observe Y_1. Then, the algorithm* SeqSEW^*_τ *is run with $\tau = 1/\sqrt{dT}$ from round 2 up to round T with the following important modification: instead of truncating the predictions to $[-B_t, B_t]$, which is best suited to the case $\mathbb{E}[Y]=0$, we truncate them to the interval*

$$\big[Y_1 - B'_t, Y_1 + B'_t\big] \;, \quad \text{where} \quad B'_t \triangleq \max_{1\leqslant s\leqslant t-1} |Y_s - Y_1| \;.$$

If η_t is changed accordingly, that is, if $\eta_t = 1/(8B_t'^2)$, then it easy to see that the resulting procedure $\widehat{f}_T \triangleq \frac{1}{T-1}\sum_{s=2}^T \widetilde{f}_s$ (where $\widetilde{f}_2, \ldots, \widetilde{f}_T$ are the estimators output by SeqSEW^*_τ*) satisfies*

$$\mathbb{E}\left[\left\| f - \widehat{f}_T \right\|_{L^2}^2\right] \leqslant \inf_{u\in\mathbb{R}^d}\left\{ \|f - u\cdot\varphi\|_{L^2}^2 + 64\left(\frac{Var[Y]}{T-1} + \psi_{T-1}\right)\|u\|_0 \ln\left(1 + \frac{\sqrt{dT}\,\|u\|_1}{\|u\|_0}\right)\right\}$$
$$+ \frac{1}{dT}\sum_{j=1}^d \left\|\varphi_j\right\|_{L^2}^2 + 10\left(\frac{Var[Y]}{T-1} + \psi_{T-1}\right) \;,$$

where $Var[Y] \triangleq \mathbb{E}\big[(Y - \mathbb{E}[Y])^2\big]$. Comparing the last bound to that of Corollary 13, we note that the two terms $\mathbb{E}[Y]^2/T$ are absent, and that we loose a multiplicative factor at most of 4 since $Var[Y] \leqslant \mathbb{E}\big[\max_{2\leqslant t\leqslant T}(Y_t - \mathbb{E}[Y_t])^2\big] \triangleq (T-1)\psi_{T-1}$ so that

$$\frac{Var[Y]}{T-1} + \psi_{T-1} \leqslant 2\psi_{T-1} \leqslant 2\left(\frac{T}{T-1}\right)\psi_T \leqslant 4\psi_T \;.$$

Remark 20 *We mentioned after Corollary 13 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*

$$\mathbb{E}\left[\max_{1\leqslant t\leqslant T}(\Delta Y)_t^2\right] = \mathbb{E}\left[\max_{1\leqslant t\leqslant T}\big(f(X_t) - \mathbb{E}[f(X)] + \varepsilon_t\big)^2\right]$$
$$\leqslant 2\,\mathbb{E}\left[\max_{1\leqslant t\leqslant T}\big(f(X_t) - \mathbb{E}[f(X)]\big)^2\right] + 2\,\mathbb{E}\left[\max_{1\leqslant t\leqslant T}\varepsilon_t^2\right] \;.$$

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* (27) *or in* (28).

A.3 Proofs of Theorem 16 and Corollary 17

Proof (of Theorem 16) The proof follows the sames lines as in the proof of Theorem 12. We thus only sketch the main arguments. In the sequel, we set $\sigma^2 \triangleq \mathbb{E}\big[\varepsilon_1^2\big]$.

Applying Corollory 8 we have, *almost surely*,

$$\sum_{t=1}^{T}\big(Y_t - \widetilde{f}_t(x_t)\big)^2 \leqslant \inf_{u\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(Y_t - u\cdot\varphi(x_t)\big)^2 + 32\left(\max_{1\leqslant t\leqslant T} Y_t^2\right)\|u\|_0 \ln\left(1+\frac{\sqrt{dT}\,\|u\|_1}{\|u\|_0}\right)\right\} + \frac{1}{dT}\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t) + 5\max_{1\leqslant t\leqslant T} Y_t^2\ .$$

Taking the expectations of both sides, expanding the squares $\big(Y_t - \widetilde{f}_t(x_t)\big)^2$ and $\big(Y_t - u\cdot\varphi(x_t)\big)^2$, noting that two terms $T\sigma^2$ cancel out,[15] and then dividing both sides by T, we get

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t) - \widetilde{f}_t(x_t)\big)^2\right] \leqslant \inf_{u\in\mathbb{R}^d}\left\{\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t) - u\cdot\varphi(x_t)\big)^2 + 32\,\frac{\mathbb{E}\big[\max_{1\leqslant t\leqslant T} Y_t^2\big]}{T}\,\|u\|_0\ln\left(1+\frac{\sqrt{dT}\,\|u\|_1}{\|u\|_0}\right)\right\} + \frac{1}{dT^2}\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t) + 5\,\frac{\mathbb{E}\big[\max_{1\leqslant t\leqslant T} Y_t^2\big]}{T}\ .$$

The right-hand side is exactly the upper bound stated in Theorem 16. We thus only need to check that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t) - \widehat{f}_T(x_t)\big)^2\right] \leqslant \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t) - \widetilde{f}_t(x_t)\big)^2\right]\ . \tag{39}$$

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,

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

15. Note that $\mathbb{E}\big[(f(x_t) - \widetilde{f}(x_t))\varepsilon_t\big] = 0$ since $\widetilde{f}_t(x_t)$ and ε_t are independent. This is due to the fact that $\widetilde{f}_t$ is built from the past data only. In particular, truncating the predictions to $B = \max_{1\leqslant t\leqslant T}|Y_t|$ might not work. A similar comment could be made in the random design case (Section 4.1).

where the second line is by definition of $\widehat{f}_T$ and where the last line follows from Jensen's inequality. Dividing both sides by T and taking their expectations, we get (39), which concludes the proof. ■

Proof (of Corollary 17) First note that

$$\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right] \triangleq \mathbb{E}\left[\max_{1\leqslant t\leqslant T} \big(f(x_t)+\varepsilon_t\big)^2\right] \leqslant 2\left(\max_{1\leqslant t\leqslant T} f^2(x_t) + \mathbb{E}\left[\max_{1\leqslant t\leqslant T} \varepsilon_t^2\right]\right) .$$

The proof then follows exactly the same lines as for Corollary 13 with the sequence (ε_t) instead of the sequence $\big((\Delta Y)_t\big)$. ■

Appendix B. Tools

Next we provide several (in)equalities that prove to be useful throughout the paper.

B.1 A Duality Formula for the Kullback-Leibler Divergence

We recall below a key duality formula satisfied by the Kullback-Leibler divergence and whose proof can be found, for example, in the monograph by Catoni (2004, pp. 159–160). We use the notations of Section 2.

Proposition 21 *For any measurable space $(\Theta,\mathcal{B})$, any probability distribution π on $(\Theta,\mathcal{B})$, and any measurable function $h:\Theta\to[a,+\infty)$ bounded from below (by some $a\in\mathbb{R}$), we have*

$$-\ln\int_\Theta e^{-h}\mathrm{d}\pi = \inf_{\rho\in\mathcal{M}_1^+(\Theta)}\left\{\int_\Theta h\,\mathrm{d}\rho + \mathcal{K}(\rho,\pi)\right\} ,$$

where $\mathcal{M}_1^+(\Theta)$ denotes the set of all probability distributions on $(\Theta,\mathcal{B})$, and where the expectations $\int_\Theta h\,\mathrm{d}\rho\in[a,+\infty]$ are always well defined since h is bounded from below.

B.2 Some Tools to Exploit Our PAC-Bayesian Inequalities

In this subsection we recall two results needed for the derivation of Proposition 1 and Proposition 5 from the PAC-Bayesian inequalities (7) and (12). The proofs are due to Dalalyan and Tsybakov (2007, 2008) and we only reproduce them for the convenience of the reader.[16]

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

$$\rho_{u^*,\tau} \triangleq \frac{\mathrm{d}\pi_\tau}{\mathrm{d}u}(u-u^*)\,\mathrm{d}u = \prod_{j=1}^d \frac{(3/\tau)\,\mathrm{d}u_j}{2\big(1+|u_j-u_j^*|/\tau\big)^4} . \tag{40}$$

16. The 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 $\|f-f_u\|_n^2$ is replaced with the cumulative loss $\sum_{t=1}^T\big(y_t-u\cdot\varphi(x_t)\big)^2$. 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 by Dalalyan and Tsybakov (2008).

Lemma 22 *For all $u^* \in \mathbb{R}^d$ and $\tau > 0$, the probability distribution $\rho_{u^*,\tau}$ satisfies*

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

Lemma 23 *For all $u^* \in \mathbb{R}^d$ and $\tau > 0$, the probability distribution $\rho_{u^*,\tau}$ satisfies*

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

Proof (of Lemma 22) For all $t \in \{1, \dots, T\}$ we expand the square $\big(y_t - u \cdot \varphi(x_t)\big)^2 = \big(y_t - u^* \cdot \varphi(x_t) + (u^* - u) \cdot \varphi(x_t)\big)^2$ and use the linearity of the integral to get

$$\begin{aligned}
&\int_{\mathbb{R}^d} \sum_{t=1}^{T} \big(y_t - u \cdot \varphi(x_t)\big)^2 \rho_{u^*,\tau}(\mathrm{d}u) \qquad (41)\\
&\quad = \sum_{t=1}^{T} \big(y_t - u^* \cdot \varphi(x_t)\big)^2 + \sum_{t=1}^{T} \int_{\mathbb{R}^d} \big((u^* - u) \cdot \varphi(x_t)\big)^2 \rho_{u^*,\tau}(\mathrm{d}u)\\
&\quad\quad + \underbrace{\sum_{t=1}^{T} 2\big(y_t - u^* \cdot \varphi(x_t)\big) \int_{\mathbb{R}^d} (u^* - u) \cdot \varphi(x_t)\, \rho_{u^*,\tau}(\mathrm{d}u)}_{=0}
\end{aligned}$$

The last sum equals zero by symmetry of $\rho_{u^*,\tau}$ around u^*, which yields $\int_{\mathbb{R}} u \, \rho_{u^*,\tau}(\mathrm{d}u) = u^*$. 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 $\big((u^* - u) \cdot \varphi(x_t)\big)^2$ we have, for all $t = 1, \dots, T$,

$$\big((u^* - u) \cdot \varphi(x_t)\big)^2 = \sum_{j=1}^{d} (u_j^* - u_j)^2 \varphi_j^2(x_t) + \sum_{1 \leqslant j \neq k \leqslant d} (u_j^* - u_j)(u_k^* - u_k)\, \varphi_j(x_t)\, \varphi_k(x_t) \; .$$

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

$$\begin{aligned}
\sum_{t=1}^{T} \int_{\mathbb{R}^d} \big((u^* - u) \cdot \varphi(x_t)\big)^2 \rho_{u^*,\tau}(\mathrm{d}u) &= \sum_{t=1}^{T} \sum_{j=1}^{d} \varphi_j^2(x_t) \int_{\mathbb{R}^d} (u_j^* - u_j)^2 \rho_{u^*,\tau}(\mathrm{d}u) + 0\\
&= \sum_{t=1}^{T} \sum_{j=1}^{d} \varphi_j^2(x_t) \int_{\mathbb{R}} (u_j^* - u_j)^2 \frac{(3/\tau)\,\mathrm{d}u_j}{2\big(1 + |u_j - u_j^*|/\tau\big)^4} \qquad (42)\\
&= \tau^2 \sum_{t=1}^{T} \sum_{j=1}^{d} \varphi_j^2(x_t) \int_{\mathbb{R}} \frac{3t^2 \mathrm{d}t}{2(1 + |t|)^4} \qquad (43)\\
&= \tau^2 \sum_{t=1}^{T} \sum_{j=1}^{d} \varphi_j^2(x_t) \; , \qquad (44)
\end{aligned}$$

where (42) follows by definition of $\rho_{u^*,\tau}$, where (43) is obtained by the change of variables $t = (u_j - u_j^*)/\tau$, and where (44) follows from the equality $\displaystyle\int_{\mathbb{R}} \frac{3t^2 \mathrm{d}t}{2\big(1 + |t|\big)^4} = 1$ that can be proved by integrating by parts. Substituting (44) into (41) concludes the proof. ∎

Proof (of Lemma 23) By definition of $\rho_{u^*,\tau}$ and π_τ, we have

$$\mathcal{K}(\rho_{u^*,\tau},\pi_\tau) \triangleq \int_{\mathbb{R}^d}\left(\ln\frac{\mathrm{d}\rho_{u^*,\tau}}{\mathrm{d}\pi_\tau}(u)\right)\rho_{u^*,\tau}(\mathrm{d}u) = \int_{\mathbb{R}^d}\left(\ln\prod_{j=1}^d\frac{\left(1+|u_j|/\tau\right)^4}{\left(1+|u_j-u_j^*|/\tau\right)^4}\right)\rho_{u^*,\tau}(\mathrm{d}u)$$
$$= 4\int_{\mathbb{R}^d}\left(\sum_{j=1}^d\ln\frac{1+|u_j|/\tau}{1+|u_j-u_j^*|/\tau}\right)\rho_{u^*,\tau}(\mathrm{d}u)\ . \tag{45}$$

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

$$1+|u_j|/\tau \leqslant 1+|u_j^*|/\tau+|u_j-u_j^*|/\tau \leqslant \left(1+|u_j^*|/\tau\right)\left(1+|u_j-u_j^*|/\tau\right)\ ,$$

so that Equation (45) yields the upper bound

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

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

$$\sum_{j:u_j^*\neq 0}\ln\left(1+|u_j^*|/\tau\right) = \|u^*\|_0\,\frac{1}{\|u^*\|_0}\sum_{j:u_j^*\neq 0}\ln\left(1+|u_j^*|/\tau\right) \leqslant \|u^*\|_0\ln\left(1+\frac{\sum_{j:u_j^*\neq 0}|u_j^*|}{\|u^*\|_0\,\tau}\right)$$
$$\leqslant \|u^*\|_0\ln\left(1+\frac{\|u^*\|_1}{\|u^*\|_0\,\tau}\right)\ .$$

This concludes the proof. ∎

B.3 Some Maximal Inequalities

Next we prove three maximal inequalities needed for the derivation of Corollaries 13 and 17 from Theorems 12 and 16 respectively. Their proofs are quite standard but we provide them for the convenience of the reader.

Lemma 24 *Let $Z_1,\ldots,Z_T$ be $T\geqslant 1$ (centered) real random variables such that, for a given constant $\nu\geqslant 0$, we have*

$$\forall t\in\{1,\ldots,T\},\quad \forall\lambda\in\mathbb{R},\quad \mathbb{E}\left[e^{\lambda Z_t}\right]\leqslant e^{\lambda^2\nu/2}\ . \tag{46}$$

Then,

$$\mathbb{E}\left[\max_{1\leqslant t\leqslant T}Z_t^2\right]\leqslant 2\nu\ln(2eT)\ .$$

Lemma 25 *Let $Z_1,\ldots,Z_T$ be $T\geqslant 1$ real random variables such that, for some given constants $\alpha>0$ and $M>0$, we have*

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

Then,

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

Lemma 26 *Let $Z_1, \dots, Z_T$ be $T \geqslant 1$ real random variables such that, for some given constants $\alpha > 2$ and $M > 0$, we have*

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

Then,

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

Proof (of Lemma 24) Let $t \in \{1, \dots, T\}$. From the subgaussian assumption (46) it is well known (see, e.g., Massart 2007, Chapter 2) that for all $x \geqslant 0$, we have

$$\forall t \in \{1, \dots, T\} \,, \quad \mathbb{P}\big(|Z_t| > x\big) \leqslant 2e^{-x^2/(2\nu)} \,.$$

Let $\delta \in (0,1)$. By the change of variables $x = \sqrt{2\nu \ln(2T/\delta)}$, the last inequality entails that, for all $t = 1, \dots, T$, we have $|Z_t| \leqslant \sqrt{2\nu \ln(2T/\delta)}$ with probability at least $1 - \delta/T$. Therefore, by a union bound, we get, with probability at least $1 - \delta$,

$$\forall t \in \{1, \dots, T\} \,, \quad |Z_t| \leqslant \sqrt{2\nu \ln(2T/\delta)} \,.$$

As a consequence, with probability at least $1 - \delta$,

$$\max_{1 \leqslant t \leqslant T} Z_t^2 \leqslant 2\nu \ln(2T/\delta) \leqslant 2\nu \ln(1/\delta) + 2\nu \ln(2T) \,.$$

It now just remains to integrate the last inequality over $\delta \in (0,1)$ as is made precise below. By the change of variables $\delta = e^{-z}$, the latter inequality yields

$$\forall z > 0 \,, \quad \mathbb{P}\left[\left(\frac{\max_{1 \leqslant t \leqslant T} Z_t^2 - 2\nu \ln(2T)}{2\nu}\right)_+ > z\right] \leqslant e^{-z} \,, \tag{47}$$

where for all $x \in \mathbb{R}$, $x_+ \triangleq \max\{x, 0\}$ denotes the positive part of x. Using the well-known fact that $\mathbb{E}[\xi] = \int_0^{+\infty} \mathbb{P}(\xi > z) \mathrm{d}z$ for all nonnegative real random variable ξ, we get

$$\begin{aligned}
\mathbb{E}\left[\frac{\max_{1 \leqslant t \leqslant T} Z_t^2 - 2\nu \ln(2T)}{2\nu}\right] &\leqslant \mathbb{E}\left[\left(\frac{\max_{1 \leqslant t \leqslant T} Z_t^2 - 2\nu \ln(2T)}{2\nu}\right)_+\right] \\
&= \int_0^{+\infty} \mathbb{P}\left[\left(\frac{\max_{1 \leqslant t \leqslant T} Z_t^2 - 2\nu \ln(2T)}{2\nu}\right)_+ > z\right] \mathrm{d}z \\
&\leqslant \int_0^{+\infty} e^{-z} \mathrm{d}z \;= 1 \,,
\end{aligned}$$

where the last line follows from (47) above. Rearranging terms, we get $\mathbb{E}\big[\max_{1 \leqslant t \leqslant T} Z_t^2\big] \leqslant 2\nu + 2\nu \ln(2T)$, which concludes the proof. ∎

Proof (of Lemma 25) We first need the following definitions. Let $\psi_\alpha : \mathbb{R}_+ \to \mathbb{R}$ be a convex majorant of $x \mapsto e^{\alpha\sqrt{x}}$ on $\mathbb{R}_+$ defined by

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

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

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

Elementary manipulations show that:

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

The proof is based on a Pisier-type argument as is done, for example, by Massart (2007, Lemma 2.3) to prove the maximal inequality $\mathbb{E}[\max_{1\leqslant t\leqslant T} \xi_t] \leqslant \sqrt{2\nu \ln T}$ for all subgaussian real random variables ξ_t, $1 \leqslant t \leqslant T$, with common variance factor $\nu \geqslant 0$.

From the inequality $x \leqslant \psi_\alpha^{-1}\big(\psi_\alpha(x)\big)$ for all $x \in \mathbb{R}_+$ we have

$$\begin{aligned} \mathbb{E}\left[\max_{1\leqslant t\leqslant T} Z_t^2\right] &\leqslant \psi_\alpha^{-1}\left(\psi_\alpha\left(\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Z_t^2\right]\right)\right) \\ &\leqslant \psi_\alpha^{-1}\left(\mathbb{E}\left[\psi_\alpha\Big(\max_{1\leqslant t\leqslant T} Z_t^2\Big)\right]\right) = \psi_\alpha^{-1}\left(\mathbb{E}\left[\max_{1\leqslant t\leqslant T} \psi_\alpha\big(Z_t^2\big)\right]\right)\,, \end{aligned}$$

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

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

where the second line follows from the inequality $\psi_\alpha(x) \leqslant e + e^{\alpha\sqrt{x}}$ 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. ■

Proof (of Lemma 26) As in the previous proof, we have, by Jensen's inequality and the fact that $x \mapsto x^{\alpha/2}$ is convex and nondecreasing on $\mathbb{R}_+$ (since $\alpha > 2$),

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

by the bounded-moment assumption, which concludes the proof. ■

References

F. Abramovich, Y. Benjamini, D.L. Donoho, and I.M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653, 2006.

P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Stat.*, 5:127–145, 2011.

J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37(4): 1591–1646, 2009. ISSN 0090-5364.

P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *J. Comp. Sys. Sci.*, 64:48–75, 2002.

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. ISSN 0885-6125.

G. Biau, K. Bleakley, L. Györfi, and G. Ottucsák. Nonparametric sequential prediction of time series. *J. Nonparametr. Stat.*, 22(3–4):297–317, 2010.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. ISSN 0090-5364.

L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203–268, 2001.

L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields*, 138:33–73, 2007.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, Heidelberg, 2011.

F. Bunea and A. Nobel. Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Trans. Inform. Theory*, 54(4):1725–1735, 2008. ISSN 0018-9448.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for regression learning. Technical report, 2004. Available at `http://arxiv.org/abs/math/0410214`.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007a. ISSN 0090-5364.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194, 2007b. ISSN 1935-7524.

E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

O. Catoni. Universal aggregation rules with exact bias bounds. Technical Report PMA-510, Laboratoire de Probabilités et Modèles Aléatoires, CNRS, Paris, 1999.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Springer, New York, 2004.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

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. ISSN 0018-9448.

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.

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. ISBN 978-3-540-72925-9.

A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2):39–61, 2008. ISSN 0885-6125.

A. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 2012a.

A. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78:1423–1443, 2012b.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81 (3):425–455, 1994. ISSN 0006-3444.

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.

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.

S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *JMLR Workshop and Conference Proceedings*, 19 (COLT 2011 Proceedings):377–396, 2011.

S. Gerchinovitz and J.Y. Yu. Adaptive and optimal online linear regression on ℓ^1-balls. In J. Kivinen, C. Szepesvári, E. Ukkonen, and T. Zeugmann, editors, *Algorithmic Learning Theory*, volume 6925 of *Lecture Notes in Computer Science*, pages 99–113. Springer Berlin/Heidelberg, 2011.

L. Györfi and G. Ottucsák. Sequential prediction of unbounded stationary time series. *IEEE Trans. Inform. Theory*, 53(5):1866–1872, 2007.

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. ISBN 0-387-95441-4.

M. Hebiri and S. van de Geer. The Smooth-Lasso and other $\ell^1 + \ell^2$-penalized methods. *Electron. J. Stat.*, 5:1184–1226, 2011.

A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5): 2183–2206, 2008. ISSN 0090-5364.

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.

V. Koltchinskii. Sparse recovery in convex hulls via entropy penalization. *Ann. Statist.*, 37(3): 1332–1359, 2009a. ISSN 0090-5364.

V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(1):7–57, 2009b. ISSN 0246-0203.

V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *J. Mach. Learn. Res.*, 10:777–801, 2009. ISSN 1532–4435.

N. Littlestone. From on-line to batch learning. In *Proceedings of the 2nd Annual Conference on Learning Theory (COLT'89)*, pages 269–284, 1989.

K. Lounici, M. Pontil, S. van de Geer, and A.B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204, 2011.

P. Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

P. Rigollet and A. B. Tsybakov. Exponential Screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011.

M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *J. Mach. Learn. Res.*, 9:759–813, 2008.

S. Shalev-Shwartz and A. Tewari. Stochastic methods for ℓ^1-regularized loss minimization. *J. Mach. Learn. Res.*, 12:1865–1892, 2011.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1): 267–288, 1996.

S. A. van de Geer. High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, 36(2): 614–645, 2008. ISSN 0090-5364.

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. ISSN 1935-7524.

V. Vovk. Competitive on-line statistics. *Internat. Statist. Rev.*, 69:213–248, 2001.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, 2010.