Sparsity Regret Bounds for Individual Sequences in Online Linear Regression

Sébastien Gerchinovitz SEBASTIEN.GERCHINOVITZ@ENS.FR
École Normale Supérieure[*]
Paris, France

Editor: Sham Kakade, Ulrike von Luxburg

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 on i.i.d. data and derive risk bounds of the same flavor as in Dalalyan and Tsybakov (2008, 2011) 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.

Keywords: Individual Sequences, Sparsity, Online Linear Regression, Regret Bounds

1. Introduction

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 $\boldsymbol{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^d u_j \varphi_j$, where $\boldsymbol{u} \in \mathbb{R}^d$, i.e., to satisfy, uniformly over all individual sequences $(x_t, y_t)_{1 \leqslant t \leqslant T}$, a regret bound of the form

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

for some regret term $\Delta_{T,d}(\boldsymbol{u})$ that should be as small as possible and, in particular, sublinear in T.

In this setting the variant of the sequential Ridge regression forecaster studied by Azoury and Warmuth (2001) and Vovk (2001) has a regret of order at most $d \ln T$. 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

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

$d \ln T$ is optimal in a certain sense (see 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 linear combination $\boldsymbol{u}^*$ (i.e., with $s \ll T/(\ln T)$ non-zero coefficients) which has a small cumulative square loss. If the forecaster knew in advance the support $J(\boldsymbol{u}^*) \triangleq \{j : u_j^* \neq 0\}$ of $\boldsymbol{u}^*$, he could apply the same forecaster as above but only to the s-dimensional linear subspace $\left\{\boldsymbol{u} \in \mathbb{R}^d : \forall j \notin J(\boldsymbol{u}^*), u_j = 0\right\}$. 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 possible, though, of course, the aforementioned regret bound $s \ln T$ can only be used as an ideal benchmark (since the support of $\boldsymbol{u}^*$ 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 (Proposition 4 combined with Remark 6, and Theorem 8) we indeed derive regret bounds of the form

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

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

This work is in connection with several papers that appeared at previous COLT conferences, either in the stochastic setting (Bunea et al., 2006; Dalalyan and Tsybakov, 2007, 2009) or in online convex optimization (Duchi et al., 2010). Next we discuss these papers and some related references.

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 coefficients of the oracle vector. Such inequalities were introduced by Bunea et al. (2004, 2006) for the regression model with random design. The same authors prove similar results for the case of a fixed design in Bunea et al. (2007) through general model selection arguments of Birgé and Massart (2001). As we do not have the space to thoroughly review the extensive literature related to sparsity oracle inequalities, we refer the reader to the full version of this paper (Gerchinovitz, 2011) for further references.

We only mention that, recently, sparsity oracle inequalities with leading constant equal to 1 have been proved for procedures based on exponential weighting; see Dalalyan and Tsybakov (2007) and the other references given in Gerchinovitz (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 Dalalyan and Tsybakov (2008, 2011). Following the same lines as in Dalalyan and Tsybakov (2009), it is possible to slightly adapt its statement to make it computationally tractable by means of Langevin Monte-Carlo approximation while not affecting its statistical properties. The technical details are however omitted in this paper, which only focuses on the theoretical guarantees of the algorithm SeqSEW.

Previous works on sparsity in the framework of individual sequences

To the best of our knowledge, Corollary 2 and its refinements (Proposition 4 combined with Remark 6, and Theorem 8) 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:

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

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

where g grows at most logarithmically in T, d, U, and $\|\boldsymbol{\varphi}\|_\infty$. 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 Vovk (2001, Theorem 2) or Cesa-Bianchi and Lugosi (2006, Chapter 11). 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 $\bigl\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_0 \leqslant s \text{ and } \|\boldsymbol{u}\|_1 \leqslant d\bigr\}$ 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, i.e., on ℓ^0-balls of small radii. See, e.g., Langford et al. (2009); Shalev-Shwartz and Tewari (2009); 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 $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{\varphi}(x_t)$ of the base forecasts (i.e., $\|\widehat{\boldsymbol{u}}_t\|_0$ 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

$$\sum_{t=1}^{T} \Bigl((y_t - \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t)^2 + \lambda \,\|\widehat{\boldsymbol{u}}_t\|_1 \Bigr) \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \Bigl((y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 + \lambda \,\|\boldsymbol{u}\|_1 \Bigr) + \widetilde{\Delta}_{T,d}(\boldsymbol{u}) \right\} \;, \qquad (2)$$

for some regret term $\widetilde{\Delta}_{T,d}(\boldsymbol{u})$ 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[1] such a regret bound with $\widetilde{\Delta}_{T,d}(\boldsymbol{u})$ at least of order $\|\boldsymbol{\varphi}\|_\infty \sqrt{dT}$ when the base forecasts $\varphi_j(x_t)$ are dense in the sense that $\max_{1 \leqslant t \leqslant T} \sum_{j=1}^d \varphi_j^2(x_t) \approx d\, \|\boldsymbol{\varphi}\|_\infty^2$. 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 $\|\boldsymbol{u}\|_1$ 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 (2009) – 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 (2) with $\widetilde{\Delta}_{T,d}(\boldsymbol{u}) \approx \mu \, \|\boldsymbol{u}\|_1 \sqrt{T \ln d}$, for some gradient-based constant μ. Therefore, in all three cases, the function $\widetilde{\Delta}$ grows at least linearly in $\|\boldsymbol{u}\|_1$ and as $\sqrt{T}$. This is in contrast with the logarithmic dependence in $\|\boldsymbol{u}\|_1$ and the fast rate $\mathcal{O}(\ln T)$ we are looking for and prove, e.g., in Corollary 2.

Note that the suboptimality of the aforementioned algorithms is specific to the goal we are pursuing, i.e., prediction on ℓ^0-balls (intersected with ℓ^1-balls). On the contrary the rate $\|\boldsymbol{u}\|_1 \sqrt{T \ln d}$ is more suited and actually optimal for learning on ℓ^1-balls (see Raskutti et al. 2009). 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).

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 inequalities were proved in the framework of prediction with expert advice, e.g., in Freund et al. (1997) and Kivinen and Warmuth (1999), or in the same setting as ours with a Gaussian prior in 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, e.g., 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] The bound stated in Langford et al. (2009, Corollary 4.1) differs from (2) in that the constant before the infimum is equal to $C = 1/(1 - 2c_d^2 \eta)$, where $c_d^2 \approx \max_{1 \leqslant t \leqslant T} \sum_{j=1}^d \varphi_j^2(x_t) \leqslant d\, \|\boldsymbol{\varphi}\|_\infty^2$, 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 $c_d^2 \approx d\, \|\boldsymbol{\varphi}\|_\infty^2$, then we have $C \approx 1 + \sqrt{2c_d^2/T}$, which yields a regret bound with leading constant 1 as in (2) and with $\widetilde{\Delta}_{T,d}(\boldsymbol{u})$ at least of order $\sqrt{c_d^2 T} \approx \|\boldsymbol{\varphi}\|_\infty \sqrt{dT}$.

Open questions by Dalalyan and Tsybakov

In Section 4 we apply a parameter-free version of our algorithm SeqSEW on i.i.d. data and derive a risk bound of the same flavor as in Dalalyan and Tsybakov (2008, 2011). 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 (2011, Section 4.2). Besides, our algorithm does not need the a priori knowledge of the variance factor of the noise when the latter is subgaussian, which solves a second question raised in Dalalyan and Tsybakov (2011, Section 5.1, Remark 6).

Outline of the paper

This paper is organized as follows. In Section 2 we describe our deterministic setting and 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. Finally, in Section 4, we apply one version of the algorithm SeqSEW on i.i.d. data and provide positive answers to two questions left open by Dalalyan and Tsybakov (2011).

2. Setting and notations

The main setting considered in this paper is an equivalent variant of an extension 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 references on this setting. We give in Figure 1 a detailed description of our repeated game.

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

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

The set of all probability distributions on a set Θ (endowed with some σ-algebra, e.g., 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.

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

Initial step: the environment chooses[a] 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 $t \in \mathbb{N}^*$,

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 $\boldsymbol{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^{d} u_j \varphi_j$, $\boldsymbol{u} \in \mathbb{R}^d$, incurs the loss $\big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2$ and the forecaster incurs the loss $(y_t - \widehat{y}_t)^2$.

[a] The game is described as if the environment were oblivious to the forecaster's predictions. Actually, since we only consider deterministic forecasters, our results also hold when $(x_t)_{t \geqslant 1}$ and $(y_t)_{t \geqslant 1}$ are chosen by an adversarial environment.

Figure 1: Description of the repeated game of online linear regression.

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

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

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

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, i.e.,

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

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, e.g., the discussion in 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, i.e., 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

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

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

At each time round $t \geqslant 1$,

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

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

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

where

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

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

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 $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$ *satisfies*

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T \bigl(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\bigr)^2 + \frac{4}{\eta} \left\Vert \boldsymbol{u} \right\Vert_0 \ln\!\left(1 + \frac{\left\Vert \boldsymbol{u} \right\Vert_1}{\left\Vert \boldsymbol{u} \right\Vert_0 \tau}\right) \right\} + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \ . \tag{4}$$

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 $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$ *satisfies*

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T \bigl(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\bigr)^2 + 32\, B_y^2 \left\Vert \boldsymbol{u} \right\Vert_0 \ln\left(1 + \frac{\sqrt{B_\Phi}\, \left\Vert \boldsymbol{u} \right\Vert_1}{4\, B_y \left\Vert \boldsymbol{u} \right\Vert_0}\right) \right\} + 16 B_y^2 \ . \tag{5}$$

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* $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$ *is used with $B \geqslant B_y$ and $\eta \leqslant 1/(8B^2)$, then*

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

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

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 $\mathcal{G} \triangleq \bigl\{x \mapsto \bigl[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x)\bigr]_B : \boldsymbol{u} \in \mathbb{R}^d\bigr\}$, and the prior[2] π on $\mathcal{G}$ induced by the prior π_τ on $\mathbb{R}^d$ via the mapping $\boldsymbol{u} \in \mathbb{R}^d \mapsto \bigl[\boldsymbol{u} \cdot \boldsymbol{\varphi}(\cdot)\bigr]_B \in \mathcal{G}$.

To apply the aforementioned theorem, recall from Vovk (2001, Remark 3) that the square loss is $1/(8B^2)$-exp-concave on $[-B, B]$ and thus η-exp-concave[3] (since $\eta \leqslant 1/(8B^2)$

[2] 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).

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

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

$$\begin{aligned}\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 &\leqslant \inf_{\mu \in \mathcal{M}_1^+(\mathcal{G})} \left\{ \int_{\mathcal{G}} \sum_{t=1}^{T} \bigl(y_t - g(x_t)\bigr)^2 \mu(\mathrm{d}g) + \frac{\mathcal{K}(\mu,\pi)}{\eta} \right\} \\ &\leqslant \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^{T} \Bigl(y_t - \bigl[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\bigr]_B\Bigr)^2 \rho(\mathrm{d}\boldsymbol{u}) + \frac{\mathcal{K}(\widetilde{\rho},\pi)}{\eta} \right\} ,\end{aligned}$$

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

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

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

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 $t = 1,\ldots,T$. ■

Proof (of Proposition 1) Our proof mimics the proof of Theorem 5 in 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).

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

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

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

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

The two terms of the right-hand side of (8) can be upper bounded as in the proof of Theorem 5 in Dalalyan and Tsybakov (2008). It is proved therein that, by a symmetry argument,

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

and, by elementary calculations,

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

Combining (8) with the last two equations, which all hold for all $\boldsymbol{u}^* \in \mathbb{R}^d$, we get Inequality (4). ■

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

$$\begin{aligned}\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 &\leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t-\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2+\frac{4}{\eta}\left\|\boldsymbol{u}\right\|_0\ln\left(1+\frac{\left\|\boldsymbol{u}\right\|_1}{\left\|\boldsymbol{u}\right\|_0\tau}\right)\right\} \;+\; \tau^2\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t)\\ &\leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t-\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2+\frac{4}{\eta}\left\|\boldsymbol{u}\right\|_0\ln\left(1+\frac{\left\|\boldsymbol{u}\right\|_1}{\left\|\boldsymbol{u}\right\|_0\tau}\right)\right\} \;+\; \tau^2 B_\Phi \;, \qquad (9)\end{aligned}$$

since $\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t) \leqslant B_\Phi$ by assumption. The particular choices[4] for η and τ given in the statement of the corollary then yield the desired inequality (5). ■

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 $|y_t|$. In this section, we remove this 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 4 and Remark 5 below.

For this purpose we consider the following algorithm called 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. More precisely, the algorithm SeqSEW_τ^* outputs at time t the prediction

$$\widehat{y}_t \triangleq \int_{\mathbb{R}^d}\big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_{B_t}\,p_t(\mathrm{d}\boldsymbol{u}) \;, \qquad (10)$$

where

$$B_t \triangleq \left(2^{\lceil \log_2 \max_{1\leqslant s\leqslant t-1} y_s^2\rceil}\right)^{1/2} \;, \qquad \eta_t \triangleq \frac{1}{8B_t^2} \;,$$

[4] The best choice of (B,η) that satisfies the assumptions of Proposition 1 is $B = B_y$ and $\eta = 1/(8B_y^2)$. As for the choice of τ, it minimizes the function $\tau \mapsto C_1\ln(C_2/\tau) + C_3\tau^2$ with $C_1 = 4/\eta = 32B_y^2$ and $C_3 = B_\Phi$.

and where, for a normalizing constant W_t, the posterior distribution $p_t \in \mathcal{M}_1^+(\mathbb{R}^d)$ is defined by

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

Note that $\max_{1 \leqslant s \leqslant t-1}|y_s| \leqslant B_t \leqslant \sqrt{2}\max_{1 \leqslant s \leqslant t-1}|y_s|$.

The idea of truncating the base forecasts was already used in the past; see, e.g., Györfi et al. (2002) for the case of least squares regression and Györfi and Ottucsák (2007); 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. The online tuning of this threshold is based on a pseudo-doubling-trick technique provided in Cesa-Bianchi et al. (2007) (we use the prefix *pseudo* since the algorithm does not restart at the beginning of each new regime).

Proposition 4 *For all $\tau > 0$, the algorithm* SeqSEW$_\tau^*$ *satisfies*

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 + 32B_{T+1}^2\,\|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0\,\tau}\right)\right\} \qquad (11)$$
$$+\ \tau^2\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t)\ + 16B_{T+1}^2\ ,$$

where

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

Remark 5 In view of Proposition 1, the algorithm SeqSEW$_\tau^*$ 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 a multiplicative factor smaller than 2 and the additive factor $16B_{T+1}^2$ which is smaller than $32B_y^2$.

Remark 6 As in the previous section, several corollaries can be derived from Proposition 4. 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 the full version of this paper (Gerchinovitz, 2011, Corollary 3). We will also use the simpler choice $\tau = 1/\sqrt{dT}$ for the stochastic setting in Section 4.

As in the previous section, to prove Proposition 4, we first need a key PAC-Bayesian inequality. The next lemma is an adaptive variant of Lemma 3.

Lemma 7 *For all $\tau > 0$, the algorithm* SeqSEW$^*_\tau$ *satisfies*

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

$$\leqslant \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t) \big)^2 \rho(\mathrm{d}\boldsymbol{u}) + 8 B_{T+1}^2 \, \mathcal{K}(\rho, \pi_\tau) \right\} + 16 B_{T+1}^2 \ , \tag{13}$$

where

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

Proof (of Lemma 7) The proof is based on similar arguments as for 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); 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 $t \geqslant 2$,

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

On the one hand, we have

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

$$= \frac{1}{\eta_{T+1}} \sup_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \left(-\eta_{T+1} \sum_{t=1}^T \Big(y_t - \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big]_{B_t} \Big)^2 \right) \rho(\mathrm{d}\boldsymbol{u}) - \mathcal{K}(\rho, \pi_\tau) \right\} \tag{14}$$

$$= - \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \Big(y_t - \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big]_{B_t} \Big)^2 \rho(\mathrm{d}\boldsymbol{u}) + \frac{\mathcal{K}(\rho, \pi_\tau)}{\eta_{T+1}} \right\} \ , \tag{15}$$

where (14) follows from the fact that, for any measurable space $(E, \mathcal{B})$, any probability distribution π on $(E, \mathcal{B})$, and any non-positive measurable function $h : E \to (-\infty, 0]$, the Legendre transform of the Kullback-Leibler divergence can be expressed as

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

This convex duality argument for the KL divergence is proved, e.g., in Catoni (2004, p. 159).

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

$$\frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} = \sum_{t=1}^T \left(\frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W_t}{\eta_t} \right) = \sum_{t=1}^T \Bigg(\underbrace{\frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W'_{t+1}}{\eta_t}}_{(1)} + \underbrace{\frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t}}_{(2)} \Bigg) \ , \tag{16}$$

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

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

Let $t \in \{1,\ldots,T\}$. 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},

$$\begin{aligned}
&\frac{1}{\eta_t}\ln\frac{W'_{t+1}}{W_t} \\
&\quad = \frac{1}{\eta_t}\ln\int_{\mathbb{R}^d} \frac{\exp\Big(-\eta_t\Big(y_t - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_{B_t}\Big)^2\Big)\exp\Big(-\eta_t\sum_{s=1}^{t-1}\Big(y_s - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_s)\big]_{B_s}\Big)^2\Big)}{W_t}\,\pi_\tau(\mathrm{d}\boldsymbol{u}) \\
&\quad = \frac{1}{\eta_t}\ln\int_{\mathbb{R}^d}\exp\left(-\eta_t\Big(y_t - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_{B_t}\Big)^2\right) p_t(\mathrm{d}\boldsymbol{u}) &(17)\\
&\quad \leqslant \begin{cases} -(y_t-\widehat{y}_t)^2 & \text{if } B_{t+1}=B_t; \\ -(y_t-\widehat{y}_t)^2 + (2B_{t+1})^2 & \text{if } B_{t+1}>B_t; \end{cases} &(18)
\end{aligned}$$

where (17) follows by definition of p_t. To get Inequality (18) when $B_{t+1}=B_t$, or, equivalently, $|y_t| \leqslant B_t$, we used the fact that the square loss is $1/(8B_t^2)$-exp-concave on $[-B_t,B_t]$ (as in Lemma 3). Indeed, by definition of $\eta_t \triangleq 1/(8B_t^2)$ and by Jensen's inequality, we get

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

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 get (18) when $B_{t+1}=B_t$.

As for the rounds t such that $B_{t+1} > B_t$, the square loss $x \mapsto (y_t - x)^2$ is no longer $1/(8B_t^2)$-exp-concave on $[-B_t,B_t]$. In this case (18) follows from the cruder upper bound $(1/\eta_t)\ln(W'_{t+1}/W_t) \leqslant 0 \leqslant -(y_t-\widehat{y}_t)^2 + (2B_{t+1})^2$ (since $|y_t|, |\widehat{y}_t| \leqslant B_{t+1}$). Summing (18) over $t=1,\ldots,T$, Equation (16) yields

$$\frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} \leqslant -\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 + 4\sum_{\substack{t=1\\ t:B_{t+1}>B_t}}^{T} B_{t+1}^2 \leqslant -\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 + 8B_{T+1}^2 \ , \qquad (19)$$

where, setting $K \triangleq \lceil \log_2 \max_{1\leqslant t\leqslant T} y_t^2 \rceil$, we bounded the geometric sum $\sum_{t:B_{t+1}>B_t}^{T} B_{t+1}^2$ from above by $\sum_{k=-\infty}^{K} 2^k = 2^{K+1} \triangleq 2B_{T+1}^2$ in the same way as in Theorem 6 of Cesa-Bianchi et al. (2007).

Putting Equations (15) and (19) together, we get the PAC-Bayesian inequality

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

which yields (12) by definition of $\eta_{T+1} \triangleq 1/(8B_{T+1}^2)$. The other PAC-Bayesian inequality (13), which is stated for non-truncated base forecasts, follows from (12) by the fact that truncation to B_t can only improve prediction if $|y_t| \leqslant B_t$. The remaining t's such that $|y_t| > B_t$ then just account for an overall additional term at most equal to $\sum_{t:B_{t+1}>B_t}^{T} \big(2B_{t+1}\big)^2 \leqslant 8B_{T+1}^2$, which concludes the proof. ■

Proof (of Proposition 4) The proof follows the exact sames lines as in Proposition 1 except that we apply Lemma 7 instead of Lemma 3. ■

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 one 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_τ^* 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 4, but which holds for a fully automatic algorithm; see Theorem 8 below.

More formally, our algorithm SeqSEW_*^* is defined as follows. The set of 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_τ^* of the previous section with the parameter $\tau = 1/\big(\exp(2^r) - 1\big)$.

Theorem 8 *Without requiring any preliminary knowledge at the beginning of the prediction game, the algorithm* SeqSEW_*^* *satisfies, for all* $T \geqslant 1$ *and all* $(x_1, y_1), \ldots, (x_T, y_T) \in \mathcal{X} \times \mathbb{R}$,

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 + 256\Big(\max_{1\leqslant t\leqslant T} y_t^2\Big)\|\boldsymbol{u}\|_0 \ln\left(\mathrm{e} + \sqrt{\sum_{t=1}^{T}\sum_{j=1}^{d}\varphi_j^2(x_t)}\right)\right. \tag{20}$$

$$\left. + \ 64\Big(\max_{1\leqslant t\leqslant T} y_t^2\Big) A_T \|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\right\} + \ \Big(1 + 38\max_{1\leqslant t\leqslant T} y_t^2\Big) A_T \ ,$$

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

Proof The proof relies on the application of Proposition 4 with $\tau = 1/\big(\exp(2^r) - 1\big)$ on all regimes r visited up to time T. Summing the corresponding inequalities over r then concludes the proof. Due to lack of space, we refer the reader to the full version of this paper (Gerchinovitz, 2011, Theorem 1) for further details. ■

4. Adaptivity to the unknown variance in the stochastic setting

In this section we apply the algorithm SeqSEW 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 $(X_1, Y_1), \ldots, (X_T, Y_T)$ 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 $x \in \mathcal{X}$. We also set $\|h\|_{L^2} \triangleq \big(\mathbb{E}[h(X)^2]\big)^{1/2}$ for all measurable functions $h : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}[h(X)^2] < \infty$.

4.1. Algorithm and main result

Even if the whole sample $(X_1, Y_1), \ldots, (X_T, Y_T)$ 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 $\tau = 1/\sqrt{dT}$. We then define our data-based regressor $\widehat{f}_T$ as the uniform average $\widehat{f}_T \triangleq \frac{1}{T}\sum_{t=1}^T \widetilde{f}_t$ of the regressors $\widetilde{f}_t : \mathcal{X} \to \mathbb{R}$ sequentially built by the algorithm SeqSEW$^*_\tau$ as

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

This technique is now quite standard in the machine learning community. Though we only state our risk bounds in expectation (which already improves on existing results in the stochastic setting), we refer to Kakade and Tewari (2009) to transform our results into risk bounds with large probability.

Note that, contrary to much prior work from the statistics community such as Catoni (2004) and Dalalyan and Tsybakov (2011), the regressors $\widetilde{f}_t : \mathcal{X} \to \mathbb{R}$ 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 $\|\varphi_j\|_\infty$, or the $\|f - \varphi_j\|_\infty$ (actually, the φ_j and the $f - \varphi_j$ do not even need to be bounded in ℓ^∞-norm).

Theorem 9 *Assume that $(X_1, Y_1), \ldots, (X_T, Y_T) \in \mathcal{X} \times \mathbb{R}$ are independent random copies of $(X, Y) \in \mathcal{X} \times \mathbb{R}$, where $\mathbb{E}[Y^2] < +\infty$ and $\|\varphi_j\|^2_{L^2} \triangleq \mathbb{E}[\varphi_j(X)^2] < +\infty$ for all $j = 1, \ldots, d$. Then, the data-based regressor $\widehat{f}_T$ defined above satisfies*

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

Proof By Proposition 4 with $\tau = 1/\sqrt{dT}$ and by definition of $\widetilde{f}_t$ above and $\widehat{y}_t \triangleq \widetilde{f}_t(X_t)$ in Equation (10), we have, *almost surely*,

$$\sum_{t=1}^{T}(Y_t - \widetilde{f}_t(X_t))^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(Y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(X_t)\big)^2 + 64\left(\max_{1\leqslant t\leqslant T} Y_t^2\right)\|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\right\}$$
$$+ \frac{1}{dT}\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(X_t) + 32 \max_{1\leqslant t\leqslant T} Y_t^2 \;.$$

Taking the expectations of both sides and applying Jensen's inequality straightforwardly concludes the proof (we refer the reader to the full version of this paper, Gerchinovitz 2011, Theorem 2 for more details). ■

The above theorem can be used under several assumptions on the distribution of the output Y. We only discuss below its application to one important set of assumptions studied, e.g., in Dalalyan and Tsybakov (2011).

4.2. Questions left open by Dalalyan and Tsybakov

Theorem 9 above provides answers to two questions left open in Dalalyan and Tsybakov (2011) when the regression function f is bounded and when the i.i.d. errors $\varepsilon_t \triangleq Y_t - f(X_t)$ are subgaussian (conditionally on the X_t) in the sense that, for some constant $\sigma^2 > 0$,

$$\|f\|_\infty < +\infty \qquad \text{and} \qquad \mathbb{E}\left[e^{\lambda\varepsilon_1} \,\middle|\, X_1\right] \leqslant e^{\lambda^2\sigma^2/2} \quad \text{a.s.,} \quad \forall \lambda \in \mathbb{R}\;. \tag{21}$$

Under the above assumptions, we prove in Gerchinovitz (2011, Corollary 5 and Remark 8) that Theorem 9 above yields, for some universal constant $C > 0$, that for all $T \geqslant 2$,

$$\mathbb{E}\left[\left\|f - \widehat{f}_T\right\|_{L^2}^2\right] \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\|f - \boldsymbol{u}\cdot\boldsymbol{\varphi}\|_{L^2}^2 + 2C\left(\|f\|_\infty^2 + \sigma^2\ln T\right)\frac{\|\boldsymbol{u}\|_0}{T}\ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\right\} \tag{22}$$
$$+ \frac{1}{dT}\sum_{j=1}^{d}\|\varphi_j\|_{L^2}^2 + \frac{C}{T}\left(\|f\|_\infty^2 + \sigma^2\ln T\right)\;.$$

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 (2011). 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[5] choice $\tau = 1/\sqrt{dT}$.

[5] Proposition 1 of Dalalyan and Tsybakov (2011) may seem more general than Theorem 9 at first sight since it holds for all $\tau > 0$, but this is actually also the case for Theorem 9. The proof of the latter would indeed have remained true had we replaced $\tau = 1/\sqrt{dT}$ with any value of $\tau > 0$. We however chose the reasonable value $\tau = 1/\sqrt{dT}$ to make our algorithm parameter-free.

Proposition 10 (A consequence of Prop. 1 of Dalalyan and Tsybakov 2011) *Assume that* $\sup_{1 \leqslant j \leqslant d} \|\varphi_j\|_\infty < \infty$ *and that the set of assumptions (21) above hold true. Then, for every* $R > 0$ *and* $\eta \leqslant \left(2\sigma^2 + 2\sup_{\|\boldsymbol{u}\|_1 \leqslant R} \|\boldsymbol{u} \cdot \boldsymbol{\varphi} - f\|_\infty^2\right)^{-1}$*, the mirror averaging aggregate* $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ *defined in Dalalyan and Tsybakov (2011, Equations (1) and (3)) satisfies*

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

We can now discuss the two questions left open by Dalalyan and Tsybakov (2011). Despite the similarity of the two bounds, the sparsity oracle inequality stated in Proposition 10 above only holds for vectors $\boldsymbol{u}$ within ℓ^1-balls of finite radii. The authors thus asked in Dalalyan and Tsybakov (2011, Section 4.2) whether it was possible to extend the infimum to the whole $\mathbb{R}^d$ space. Our results show that, thanks to data-driven truncation, the answer is positive.

The second open question, which was raised in Dalalyan and Tsybakov (2011, 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. The authors thus asked whether adaptivity to σ^2 was possible. Our sparsity oracle inequality (22) above provides a positive answer (up to a $\ln T$ factor).

Remark 11 Similar adaptivity results hold in the regression model with fixed design; see the full version of this paper (Gerchinovitz, 2011, Section 5.2). The framework of prediction of individual sequences thus seems to offer a unifying setting to address tuning issues both in the random and in the fixed design regression models.

Acknowledgments

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

References

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

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.

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

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

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 and sparsity via ℓ_1 penalized least squares. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT'06)*, pages 379–391, 2006.

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

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, 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.

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

A. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT'09)*, pages 83–92, 2009.

A. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 2011. To appear. Available at `http://hal.archives-ouvertes.fr/hal-00461580/`.

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. Technical report, 2011. Available at `http://arxiv.org/abs/1101.1057`.

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.

S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems 21 (NIPS'08)*, pages 801–808. 2009.

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.

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

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of convergence for high-dimensional regression under ℓ^q-ball sparsity. In *Proceedings of the 47th annual Allerton conference on communication, control, and computing (Allerton'09)*, pages 251–257, 2009.

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. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, pages 929–936, 2009.

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.