Adaptive and Optimal Online Linear Regression on L1-balls

Sébastien Gerchinovitz, Jia Yuan Yu

Adaptive and Optimal Online Linear Regression on ℓ^1-balls

Sébastien Gerchinovitz[1] and Jia Yuan Yu[1,2]

[1] École Normale Supérieure**, Paris , France
[2] HEC Paris, CNRS, Jouy-en-Josas, France

S. Gerchinovitz is a student author and wishes to be considered for the E.M. Gold Award.

Abstract. We consider the problem of online linear regression on individual sequences. The goal in this paper is for the forecaster to output sequential predictions which are, after T time rounds, almost as good as the ones output by the best linear predictor in a given ℓ^1-ball in $\mathbb{R}^d$. We consider both the cases where the dimension d is small and large relative to the time horizon T. We first present regret bounds with optimal dependencies on the sizes U, X and Y of the ℓ^1-ball, the input data and the observations. The minimax regret is shown to exhibit a regime transition around the point $d = \sqrt{T}UX/(2Y)$. Furthermore, we present efficient algorithms that are adaptive, i.e., they do not require the knowledge of U, X, and Y, but still achieve nearly optimal regret bounds.

1 Introduction

In this paper, we consider the problem of online linear regression against arbitrary sequences of input data and observations, with the objective of being competitive with respect to the best linear predictor in an ℓ^1-ball of arbitrary radius. This extends the task of convex aggregation. We consider both low and high dimensional input data. Indeed, in a large number of contemporary problems, the available data can be high-dimensional—the dimension of each data point is greater than the number of data points. Examples include analysis of DNA sequences, prediction with sparse data (e.g., Netflix problem), times series of seismic activity. In such high-dimensional problems, even linear regression on a small ℓ^1-ball is sufficient if the best predictor is sparse. Our goal is, in both low and high dimensions, to provide online linear regression algorithms along with bounds on ℓ^1-balls that characterize their robustness to worst-case scenarios.

1.1 Setting

We consider the online version of linear regression, which unfolds as follows. First, the environment chooses a sequence of observations $(y_t)_{t\geqslant 1}$

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

in $\mathbb{R}$ and a sequence of input vectors $(\boldsymbol{x}_t)_{t\geqslant 1}$ in $\mathbb{R}^d$, both initially hidden from the forecaster. At each time instant $t \in \mathbb{N}^* = \{1, 2, \ldots\}$, the environment reveals the data $\boldsymbol{x}_t \in \mathbb{R}^d$; the forecaster then gives a prediction $\widehat{y}_t \in \mathbb{R}$; the environment in turn reveals the observation $y_t \in \mathbb{R}$; and finally, the forecaster incurs the square loss $(y_t - \widehat{y}_t)^2$. The dimension d can be either small or large relative to the number T of time steps: we consider both cases.

An ℓ^1-ball of radius U is the following bounded subset of $\mathbb{R}^d$:

$$B_1(U) \triangleq \left\{ \boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U \right\}.$$

Given a fixed radius $U > 0$ and a time horizon $T \geqslant 1$, in this paper, the goal of the forecaster is to predict almost as well as the best linear forecaster in $\left\{ \boldsymbol{x} \in \mathbb{R}^d \mapsto \boldsymbol{u} \cdot \boldsymbol{x} \in \mathbb{R} : \boldsymbol{u} \in B_1(U) \right\}$, where $\cdot$ denotes the standard inner product in $\mathbb{R}^d$, i.e., to minimize the regret on $B_1(U)$ defined by

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 - \min_{\boldsymbol{u} \in B_1(U)} \left\{ \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \right\}.$$

We shall present algorithms along with bounds on their regret that hold uniformly over all sequences[3] $(\boldsymbol{x}_t, y_t)_{1 \leqslant t \leqslant T}$ such that $\|\boldsymbol{x}_t\|_\infty \leqslant X$ and $|y_t| \leqslant Y$ for all $t = 1, \ldots, T$, where $X, Y > 0$. These regret bounds contain three important quantities: X, Y, and U, which may be known or unknown to the forecaster.

1.2 Contributions and related works

The literature on online linear regression is extensive, we can only situate our work with those closest to ours.

Our first contribution is to show a refined regret bound for online linear regression on ℓ^1-balls in the arbitrary sequence setting. This bound is expressed in terms of Y, d, and a quantity $\kappa = \sqrt{T}UX/(2dY)$. This quantity κ is used to distinguish two regimes: we show a distinctive regime transition[4] at $\kappa = 1$ or $d = \sqrt{T}UX/(2Y)$. Namely, for $\kappa < 1$, the regret is of the order of $\sqrt{T}$, whereas it is of the order of $\ln T$ for $\kappa > 1$.

This regret bound matches the optimal risk bounds for stochastic settings[5] [BM01,Tsy03,RWY09]. Hence, linear regression is just as hard in the stochastic setting as in the arbitrary sequence setting. Using the standard online to batch trick, we make the latter statement more precise by establishing a lower bound for all κ at least of order $\sqrt{\ln d}/d$. This lower bound extends those of [CB99,KW97], which only hold for small κ of the order of $1/d$.

[3] Actually our results hold whether $(\boldsymbol{x}_t, y_t)_{t\geqslant 1}$ is generated by an oblivious environment or a non-oblivious opponent since we consider deterministic forecasters.

[4] In high dimensions (i.e., when $d > \omega T$, for some absolute constant $\omega > 0$), we do not observe this transition (cf. Figure 1).

[5] For example, $(\boldsymbol{x}_t, y_t)_{1 \leqslant t \leqslant T}$ may be i.i.d. , or $\boldsymbol{x}_t$ can be deterministic and $y_t = f(\boldsymbol{x}_t) + \varepsilon_t$ for an unknown function f and an i.i.d. sequence $(\varepsilon_t)_{1 \leqslant t \leqslant T}$ of Gaussian noise.

In the individual sequence setting, [CBLW96] presents a gradient descent algorithm with regret bounds relative to predictors in an ℓ^2-ball. For the regret relative to predictors in an ℓ^1-ball, the EG$^\pm$ algorithm of [KW97] achieves a regret bound of $2UXY\sqrt{2T\ln(2d)} + 2U^2X^2\ln(2d)$. This algorithm is efficient, and our lower bound in terms of κ shows that it is optimal up to logarithmic factors in the regime $\kappa \leqslant 1$. However, the EG$^\pm$ algorithm requires prior knowledge of U, X, and Y.

Our second contribution is a generic method, called *loss Lipschitzification*, which enables to adapt automatically to X and Y when U is known. Our method transforms the loss function $\boldsymbol{u} \mapsto (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2$ into a Lipschitz continuous function and adapts to the unknown Lipschitz constant. The LEG algorithm (Section 3) illustrates this technique by modifying the EG$^\pm$ algorithm [KW97] to yield an algorithm of the same computational complexity that also achieves the minimax regret without needing to know X and Y beforehand.

Our third contribution is a simple method to achieve minimax regret uniformly over all ℓ^1-balls $B_1(U)$ for $U > 0$. This robustness property is similar to that of the p-norm algorithms [GL03], but our method guarantees a better regret bound[6]. This method aggregates instances of an algorithm that require prior knowledge of U. For the sake of simplicity, we assume that X and Y are known, but explain in the discussions how to extend the method to a fully adaptive algorithm that does not require X, Y or U.

The SMIDAS algorithm [SST09] and the COMID algorithm [DSSST10], which generalize p-norm algorithms, can be shown to achieve the minimax regret if U, X and Y are known. The LEG algorithm (Section 3) does so without prior knowledge of the problem parameters X and Y. When U is unknown, the Scaling algorithm (Section 4) has a better bound than the SMIDAS algorithm[6].

The paper is organized as follows. In Section 2, we establish our refined upper and lower bounds in terms of the intrinsic quantity κ. In Section 3, we present an efficient and adaptive algorithm that achieves the optimal regret on $B_1(U)$ when U is known. Finally, we use an aggregating strategy to achieve an optimal regret uniformly over all ℓ^1-balls $B_1(U)$, for $U > 0$, when X and Y are known (Section 4). In Section 5, we discuss as an extension a fully automatic algorithm that requires no prior knowledge of U, X or Y.

2 Optimal rates

In this section, we first present a refined upper bound on the minimax regret on $B_1(U)$ for an arbitrary $U > 0$. In Corollary 1, we express this upper bound in terms of an intrinsic quantity $\kappa \triangleq \sqrt{T}UX/(2dY)$. The optimality of the latter bound is shown in Section 2.2.

2.1 Upper bound

Theorem 1 (Upper bound). *Let $d, T \in \mathbb{N}^*$, and $U, X, Y > 0$. The minimax regret on $B_1(U)$ for bounded base predictions and observations*

[6] Our regret bound grows as U instead of U^2.

satisfies

$$\inf_F \sup_{\|\boldsymbol{x}_t\|_\infty \leqslant X,\ |y_t| \leqslant Y} \left\{ \sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \right\}$$
$$\leqslant \begin{cases} 3UXY\sqrt{2T\ln(2d)} & \text{if} \quad U < \frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}} \ , \\ 26\,UXY\sqrt{T\ln\left(1+\frac{2dY}{\sqrt{T}UX}\right)} & \text{if} \quad \frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}} \leqslant U \leqslant \frac{2dY}{\sqrt{T}X} \ , \\ 32\,dY^2\ln\left(1+\frac{\sqrt{T}UX}{dY}\right) + dY^2 & \text{if} \quad U > \frac{2dY}{X\sqrt{T}} \ , \end{cases}$$

where the infimum is taken over all forecasters F and where the supremum extends over all sequences $(\boldsymbol{x}_t, y_t)_{1 \leqslant t \leqslant T} \in (\mathbb{R}^d \times \mathbb{R})^T$ such that $|y_1|, \ldots, |y_T| \leqslant Y$ and $\|\boldsymbol{x}_1\|_\infty, \ldots, \|\boldsymbol{x}_T\|_\infty \leqslant X$.

Theorem 1 improves the bound of [KW97, Theorem 5.11] for the EG$^\pm$ algorithm. First, our bound depends logarithmically—as opposed to linearly—on U for $U > 2dY/(\sqrt{T}X)$. Secondly, it is smaller by a factor ranging from 1 to $\sqrt{\ln d}$ when

$$\frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}} \leqslant U \leqslant \frac{2dY}{\sqrt{T}X} \ . \tag{1}$$

Hence, Theorem 1 answers a question[7] raised in [KW97] about the gap of $\sqrt{\ln(2d)}$ between the upper and lower bounds.

The proof appears in [GY11]. It uses a Maurey-type argument: we randomize over a discretization of $B_1(U)$. Although this argument was used in the stochastic setting (cf. [Nem00,Tsy03,BN08,SSSZ10]), we adapt it to the deterministic setting. This is yet another technique that can be applied to both the stochastic and individual sequence settings.

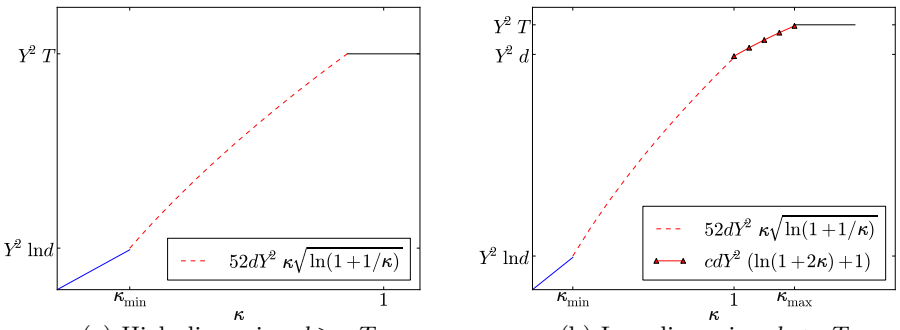
The following corollary expresses the upper bound of Theorem 1 in terms of an intrinsic quantity $\kappa \triangleq \sqrt{T}UX/(2dY)$ that relates $\sqrt{T}UX/(2Y)$ to the ambient dimension d.

Corollary 1 (Upper bound in terms of an intrinsic quantity). *Let $d, T \in \mathbb{N}^*$, and $U, X, Y > 0$. The upper bound of Theorem 1 expressed in terms of d, Y, and the intrinsic quantity $\kappa \triangleq \sqrt{T}UX/(2dY)$ reads:*

$$\inf_F \sup_{\|\boldsymbol{x}_t\|_\infty \leqslant X,\ |y_t| \leqslant Y} \left\{ \sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \right\}$$
$$\leqslant \begin{cases} 6\,dY^2\kappa\sqrt{2\ln(2d)} & \text{if} \quad \kappa < \frac{\sqrt{\ln(1+2d)}}{2d\sqrt{\ln 2}} \ , \\ 52\,dY^2\kappa\sqrt{\ln(1+1/\kappa)} & \text{if} \quad \frac{\sqrt{\ln(1+2d)}}{2d\sqrt{\ln 2}} \leqslant \kappa \leqslant 1 \ , \\ 32\,dY^2\big(\ln(1+2\kappa)+1\big) & \text{if} \quad \kappa > 1 \ . \end{cases}$$

The upper bound of Corollary 1 is shown in Figure 1. Observe that, in low dimension (Figure 1(b)), a clear transition from a regret of the order of $\sqrt{T}$ to one of $\ln T$ occurs at $\kappa = 1$. This transition is absent for high dimensions: for $d \geqslant \omega T$, where $\omega \triangleq \big(32(\ln(3)+1)\big)^{-1}$, the regret bound $32\,dY^2\big(\ln(1+2\kappa)+1\big)$ is worse than a trivial bound of TY^2 when $\kappa \geqslant 1$.

[7] The authors ask "For large d there is a significant gap between the upper and lower bounds. We would like to know if it possible to improve the upper bounds by eliminating the $\ln d$ factors."



(a) High dimension $d \geqslant \omega T$. (b) Low dimension $d < \omega T$.

Fig. 1. The regret bound of Corollary 1 over $B_1(U)$ as a function of $\kappa = \sqrt{T}UX/(2dY)$. The constant c is chosen to ensure continuity at $\kappa = 1$, and $\omega \triangleq \big(32(\ln(3)+1)\big)^{-1}$. We define: $\kappa_{\min} = \sqrt{\ln(1+2d)}/(2d\sqrt{\ln 2})$ and $\kappa_{\max} = (e^{(T/d-1)/c} - 1)/2$.

2.2 Lower bound

Corollary 1 gives an upper bound on the regret in terms of the quantities d, Y, and $\kappa \triangleq \sqrt{T}UX/(2dY)$. We now show that for all $d \in \mathbb{N}^*$, $Y > 0$, and $\kappa \geqslant \sqrt{\ln(1+2d)}/(2d\sqrt{\ln 2})$, the upper bound can not be improved[8] up to logarithmic factors.

Theorem 2 (Lower bound). *For all $d \in \mathbb{N}^*$, $Y > 0$, and $\kappa \geqslant \frac{\sqrt{\ln(1+2d)}}{2d\sqrt{\ln 2}}$, there exist $T \geqslant 1$, $U > 0$, and $X > 0$ such that $\sqrt{T}UX/(2dY) = \kappa$ and*

$$\inf_F \sup_{\|\boldsymbol{x}_t\|_\infty \leqslant X,\ |y_t| \leqslant Y} \left\{ \sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \right\}$$
$$\geqslant \begin{cases} \frac{c_1}{\ln\left(2+16d^2\right)} dY^2 \kappa \sqrt{\ln\left(1+1/\kappa\right)} & \text{if} \quad \frac{\sqrt{\ln(1+2d)}}{2d\sqrt{\ln 2}} \leqslant \kappa \leqslant 1\ , \\ \frac{c_2}{\ln\left(2+16d^2\right)} dY^2 & \text{if} \quad \kappa > 1\ , \end{cases}$$

where $c_1, c_2 > 0$ are absolute constants. The infimum is taken over all forecasters F and the supremum extends over all sequences $(\boldsymbol{x}_t, y_t)_{1\leqslant t\leqslant T} \in (\mathbb{R}^d \times \mathbb{R})^T$ such that $|y_1|, \ldots, |y_T| \leqslant Y$ and $\|\boldsymbol{x}_1\|_\infty, \ldots, \|\boldsymbol{x}_T\|_\infty \leqslant X$.

The above lower bound extends those of [CB99,KW97], which hold for small κ of the order of $1/d$. The proof appears in [GY11]. We perform a reduction to the stochastic batch setting—via a standard online to batch trick, and employ a version of a lower bound of [Tsy03].

[8] For T sufficiently large, we may overlook the case $\kappa < \sqrt{\ln(1+2d)}/(2d\sqrt{\ln 2})$ or $\sqrt{T} < (Y/(UX))\sqrt{\ln(1+2d)/\ln 2}$. Observe that in this case, the minimax regret is already of the order of $Y^2 \ln(1+d)$ (cf. Figure 1).

3 Adaptation to unknown X and Y

Although the proof of Theorem 1 already gives an algorithm that achieves the minimax regret, the latter takes as input X and Y, and it is inefficient in high dimensions. In this section, we present a new method that achieves the minimax regret both efficiently and without prior knowledge of X and Y, provided that U is known. Adaptation to an unknown U is considered in Section 4. Our method consists of modifying an underlying linear regression algorithm such as the EG$^{\pm}$ algorithm [KW97] or the sequential Ridge forecaster [Vov01,AW01]. Next, we show that the EG$^{\pm}$ algorithm with Lipschitzified losses achieves the minimax regret for the regime $d > \sqrt{T}UX/(2Y)$. A simpler modification (without loss Lipschitzification) can be applied to the Ridge forecaster to achieve a nearly optimal regret bound of order $dY^2 \ln\Bigl(1 + d\Bigl(\frac{\sqrt{T}UX}{dY}\Bigr)^2\Bigr)$ in the regime $d < \sqrt{T}UX/(2Y)$. The latter analysis is more technical and omitted.

3.1 Lipschitzification of the loss function

The second algorithm of the proof of Theorem 1 is computationally inefficient because it aggregates approximately $d^{\sqrt{T}}$ experts. In contrast, the EG$^{\pm}$ algorithm has a manageable computational complexity that is linear in d. We now describe a version of the EG$^{\pm}$ algorithm that is minimax optimal but does not require prior knowledge of X and Y—as opposed to the EG$^{\pm}$ algorithm. Our key technique consists of transforming the loss functions $\boldsymbol{u} \mapsto (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$ into functions $\widetilde{\ell}_t$ that are Lipschitz continuous with respect to $\|\cdot\|_1$. Afterward, adaptation to the unknown Lipschitz constants is carried out using the techniques of [CBMS07].

We point out that our Lipschitzification method can be applied to other algorithms, such as the p-norm algorithm and its regularized variants (SMIDAS and COMID) [GL03,SST09,DSSST10]. The method may also apply to loss functions other than the square loss, e.g., convex but non-Lipschitz functions.

The Lipschitzification proceeds as follows. At each time t, we set

$$B_t \triangleq \Bigl(2^{\lceil \log_2(\max_{1 \leqslant s \leqslant t-1} y_s^2) \rceil}\Bigr)^{1/2} \;,$$

so that $y_s \in [-B_t, B_t]$ for all $s = 1, \dots, t$. The modified loss function $\widetilde{\ell}_t : \mathbb{R}^d \to \mathbb{R}$ is constructed as follows:

- if $|y_t| > B_t$, then

$$\widetilde{\ell}_t(\boldsymbol{u}) = 0 \quad \text{for all } \boldsymbol{u} \in \mathbb{R}^d \;;$$

- if $|y_t| \leqslant B_t$, then $\widetilde{\ell}_t$ is the convex function that coincides with the square loss when $\bigl|\boldsymbol{u} \cdot \boldsymbol{x}_t\bigr| \leqslant B_t$ and is linear elsewhere. This function is shown in Figure 2 and can be formally defined as

$$\widetilde{\ell}_t(\boldsymbol{u}) \triangleq \begin{cases} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 & \text{if } \bigl|\boldsymbol{u} \cdot \boldsymbol{x}_t\bigr| \leqslant B_t, \\ (y_t - B_t)^2 + 2(B_t - y_t)(\boldsymbol{u} \cdot \boldsymbol{x}_t - B_t) & \text{if } \boldsymbol{u} \cdot \boldsymbol{x}_t > B_t, \\ (y_t + B_t)^2 + 2(-B_t - y_t)(\boldsymbol{u} \cdot \boldsymbol{x}_t + B_t) & \text{if } \boldsymbol{u} \cdot \boldsymbol{x}_t < -B_t. \end{cases}$$

Observe that in both cases $|y_t| > B_t$ and $|y_t| \leqslant B_t$, the function $\widetilde{\ell}_t$ is continuously differentiable and Lipschitz continuous with respect to $\|\cdot\|_1$ with Lipschitz constant

$$\left\|\nabla \widetilde{\ell}_t\right\|_\infty \leqslant 2\big(|y_t| + B_t\big)\,\|\boldsymbol{x}_t\|_\infty \leqslant 2\big(1+\sqrt{2}\big)\,\|\boldsymbol{x}_t\|_\infty \max_{1\leqslant s\leqslant t} |y_s| \;, \qquad (2)$$

where we used the fact that $B_t \leqslant \sqrt{2}\max_{1\leqslant s\leqslant t-1}|y_s|$. We can also glean from Figure 2 that, when $|y_t| \leqslant B_t$, we have

$$\forall \boldsymbol{u} \in \mathbb{R}^d, \quad \big(y_t - [\boldsymbol{u}\cdot\boldsymbol{x}_t]_{B_t}\big)^2 \leqslant \widetilde{\ell}_t(\boldsymbol{u}) \leqslant \big(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2 \;, \qquad (3)$$

where for all $B > 0$, we define the clipping operator $[\cdot]_B$ by

$$[x]_B \triangleq \min\big\{B, \max\{-B, x\}\big\} \quad \text{for all } x \in \mathbb{R} \;.$$

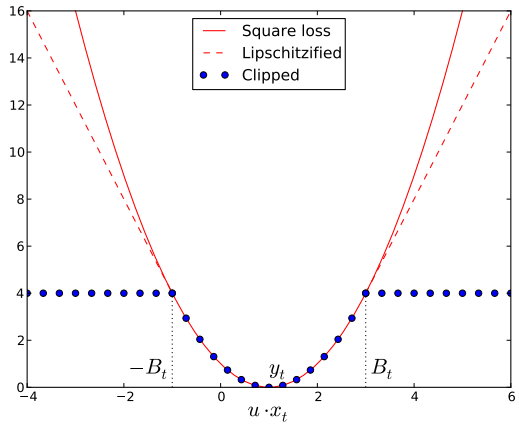


Fig. 2. Example when $|y_t| \leqslant B_t$. The square loss $(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2$, its clipped version $\big(y_t - [\boldsymbol{u}\cdot\boldsymbol{x}_t]_{B_t}\big)^2$ and its Lipschitzified version $\widetilde{\ell}_t(\boldsymbol{u})$ are plotted as a function of $\boldsymbol{u}\cdot\boldsymbol{x}_t$.

3.2 Lipschitzifying Exponentiated Gradient algorithm

Consider the LEG algorithm of Figure 3. Let $(\mathbf{e}_j)_{1\leqslant j\leqslant d}$ denote the canonical basis of $\mathbb{R}^d$ and $\pm U\boldsymbol{e}_j$ denote the vertices of $B_1(U)$. We use as a blackbox the exponentially weighted majority forecaster of [CBMS07] on $2d$ experts—namely, $\{\pm U\boldsymbol{e}_j : j = 1,\dots,d\}$—as in [KW97]. It adapts to the unknown Lipschitz constant $\max_{1\leqslant t\leqslant T}\|\nabla\widetilde{\ell}_t\|_\infty$ by the particular choice of η_t.

We first need some notations. Following the tuning provided by [CBMS07], the parameter η_t of the LEG algorithm (see Figure 3) is defined by

$$\eta_t = \min\left\{\frac{1}{\widehat{E}_{t-1}}, C\sqrt{\frac{\ln K}{V_{t-1}}}\right\} \;, \qquad (5)$$

Parameter: radius $U > 0$.
Initialization: $B_1 \triangleq 0$, $\widehat{\boldsymbol{w}}_0 \triangleq \big(1/(2d), \dots, 1/(2d)\big) \in \mathbb{R}^{2d}$.
At each time round $t \geqslant 1$,

1. Compute the linear combination
$$\widehat{\boldsymbol{u}}_t \triangleq U \sum_{j=1}^{d} \big(\widehat{w}_{2j-1,t} - \widehat{w}_{2j,t}\big)\boldsymbol{e}_j \in B_1(U); \tag{4}$$
2. Get $\boldsymbol{x}_t \in \mathbb{R}^d$ and output the clipped prediction $\widehat{y}_t \triangleq \big[\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t\big]_{B_t}$;
3. Get $y_t \in \mathbb{R}$ and define the modified loss $\widetilde{\ell}_t : \mathbb{R}^d \to \mathbb{R}$ as above;
4. Update the parameter η_{t+1} according to (5);
5. Update the weight vector $\boldsymbol{w}_{t+1} = (w_{1,t+1}, \dots, w_{2d,t+1})$ defined for all $j = 1, \dots, d$ and $\varepsilon \in \{0, 1\}$ by
$$w_{2j-1+\varepsilon,t+1} \triangleq \frac{\exp\left(-\eta_{t+1} \sum_{s=1}^{t} (-1)^{\varepsilon} \nabla \widetilde{\ell}_s\big(\widehat{\boldsymbol{u}}_s\big) \cdot U\boldsymbol{e}_j\right)}{\sum\limits_{\substack{1 \leqslant k \leqslant K \\ \varepsilon' \in \{0,1\}}} \exp\left(-\eta_{t+1} \sum_{s=1}^{t} (-1)^{\varepsilon'} \nabla \widetilde{\ell}_s\big(\widehat{\boldsymbol{u}}_s\big) \cdot U\boldsymbol{e}_k\right)};$$
6. Update the threshold $\quad B_{t+1} \triangleq \left(2^{\lceil \log_2(\max_{1 \leqslant s \leqslant t} y_s^2) \rceil}\right)^{1/2}$.

Fig. 3. The Lipschitzifying Exponentiated Gradient (LEG) algorithm.

where $C \triangleq \sqrt{2(\sqrt{2}-1)/(\mathrm{e}-2)}$ and

$$z_{j,\varepsilon}^{s} \triangleq (-1)^{\varepsilon} \nabla \widetilde{\ell}_s\big(\widehat{\boldsymbol{u}}_s\big) \cdot U\boldsymbol{e}_j, \quad s \in \{1, \dots, T\}, j \in \{1, \dots, d\}, \varepsilon \in \{0, 1\},$$

$$\widehat{E}_{t-1} \triangleq \min_{k=1,2,\dots} \left\{ 2^k : 2^k \geqslant \max_{1 \leqslant s \leqslant t-1} \left| \max_{\substack{1 \leqslant j \leqslant d \\ \varepsilon \in \{0,1\}}} z_{j,\varepsilon}^{s} - \min_{\substack{1 \leqslant j \leqslant d \\ \varepsilon \in \{0,1\}}} z_{j,\varepsilon}^{s} \right| \right\}$$

$$V_{t-1} \triangleq \sum_{s=1}^{t-1} \sum_{j,\varepsilon} w_{2j-1+\varepsilon,s} \left(z_{j,\varepsilon}^{s} - \sum_{k,\gamma} w_{2k-1+\gamma,s} z_{k,\gamma}^{s}\right)^2.$$

Note that $\widehat{E}_{t-1}$ approximates the range of the $z_{j,\varepsilon}^{s}$ up to time $t-1$, while V_{t-1} is the corresponding cumulative variance of the forecaster.

The next theorem bounds the regret of the LEG algorithm on $B_1(U)$. This algorithm is efficient and adaptive in X and Y; it achieves approximately the regret bound of Theorem 1 in the regime $\kappa \leqslant 1$ or $d \geqslant \sqrt{T}UX/(2Y)$.

Theorem 3. *Let $U > 0$ and $T \geqslant 1$. Then, for all individual sequences $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T) \in \mathbb{R}^d \times \mathbb{R}$, the Lipschitzifying Exponentiated Gradient algorithm tuned with U satisfies the regret bound*

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T}(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \leqslant c_1 U X Y \left(\sqrt{T \ln(2d)} + 8 \ln(2d)\right) + c_2 Y^2 \ ,$$

where $c_1 \triangleq 4\big(\sqrt{2}+1\big)$ and $c_2 \triangleq 4\left(1 + 1/\sqrt{2}\right)^2$, and where the quantities $X = \max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty$ and $Y = \max_{1 \leqslant t \leqslant T} |y_t|$ are unknown to the forecaster.

Proof (of Theorem 3). By definition of $\widehat{y}_t$ and $B_{t+1} \geqslant |y_t|$ we have

$$\begin{aligned}
\sum_{t=1}^{T}\big(y_t - \widehat{y}_t\big)^2 &\leqslant \sum_{\substack{t=1 \\ t:|y_t| \leqslant B_t}}^{T} \left(y_t - \left[\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t\right]_{B_t}\right)^2 + \sum_{\substack{t=1 \\ t:|y_t| > B_t}}^{T} (B_{t+1} + B_t)^2 \\
&\leqslant \sum_{\substack{t=1 \\ t:|y_t| \leqslant B_t}}^{T} \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) + \left(1 + \frac{1}{\sqrt{2}}\right)^2 \sum_{\substack{t=1 \\ t:B_{t+1} > B_t}}^{T} B_{t+1}^2 \\
&\leqslant \sum_{t=1}^{T} \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) + 4\left(1 + \frac{1}{\sqrt{2}}\right)^2 Y^2 \ ,
\end{aligned}$$

where the second inequality follows from the fact that:

- if $|y_t| \leqslant B_t$ then $(y_t - [\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t]_{B_t})^2 \leqslant \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)$ by Equation (3);
- if $|y_t| > B_t$, which is equivalent to $B_{t+1} > B_t$ by definition of B_{t+1}, then $B_t \leqslant B_{t+1}/\sqrt{2}$, so that $B_{t+1} + B_t \leqslant \big(1 + 1/\sqrt{2}\big)B_{t+1}$.

As for the third inequality above, we used the non-negativity of $\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)$ and upper bounded the geometric sum $\sum_{t:B_{t+1}>B_t}^{T} B_{t+1}^2$ in the same way as in [CBMS07, Theorem 6], i.e., setting $K \triangleq \lceil \log_2 \max_{1 \leqslant t \leqslant T} y_t^2 \rceil$,

$$\sum_{t:B_{t+1}>B_t}^{T} B_{t+1}^2 \leqslant \sum_{k=-\infty}^{K} 2^k = 2^{K+1} \leqslant 4Y^2 \ .$$

Since $\widetilde{\ell}_t(\boldsymbol{u}) \leqslant (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$ for all $\boldsymbol{u} \in \mathbb{R}^d$ (by Equation (3) if $|y_t| \leqslant B_t$, obvious otherwise), the last inequality yields

$$\begin{aligned}
&\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T}(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \\
&\qquad \leqslant \sum_{t=1}^{T} \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \widetilde{\ell}_t(\boldsymbol{u}) + 4\left(1 + \frac{1}{\sqrt{2}}\right)^2 Y^2 \ . \qquad (6)
\end{aligned}$$

But, by convexity and continuous differentiability of $\widetilde{\ell}_t$,

$$\begin{aligned}
\sum_{t=1}^{T} \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \widetilde{\ell}_t(\boldsymbol{u}) &\leqslant \sup_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \nabla \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) \cdot (\widehat{\boldsymbol{u}}_t - \boldsymbol{u}) \\
&\leqslant \sum_{t=1}^{T} \nabla \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) \cdot \widehat{\boldsymbol{u}}_t - \min_{\substack{1 \leqslant j \leqslant d \\ \gamma \in \{\pm 1\}}} \sum_{t=1}^{T} \nabla \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) \cdot (U\gamma \boldsymbol{e}_j) \ , \qquad (7)
\end{aligned}$$

where the second inequality holds by linearity of $\boldsymbol{u} \mapsto \nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) \cdot (\widehat{\boldsymbol{u}}_t - \boldsymbol{u})$ on the polytope $B_1(U)$.

In view of (4), and applying Lemma 2 in the appendix (a straightforward consequence of Corollary 1 in [CBMS07]), we get

$$\begin{aligned}
\sum_{t=1}^{T} U\,\nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) \cdot \frac{\widehat{\boldsymbol{u}}_t}{U} - \min_{\substack{1\leqslant j\leqslant d \\ \gamma\in\{\pm 1\}}} \sum_{t=1}^{T} \nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) \cdot (U\gamma \boldsymbol{e}_j) \\
\leqslant U \max_{1\leqslant t\leqslant T} \left\|\nabla\widetilde{\ell}_t\right\|_\infty \left(2\sqrt{T\ln(2d)} + 4\ln(2d) + 6\right) \\
\leqslant 4\big(\sqrt{2}+1\big)UXY\left(\sqrt{T\ln(2d)} + 2\ln(2d) + 3\right)\,, \qquad (8)
\end{aligned}$$

where the last inequality follows from $\|\nabla\widetilde{\ell}_t\|_\infty \leqslant 2\big(\sqrt{2}+1\big)XY$ by (2). Putting Equations (6), (7), and (8) together and noting that $3 \leqslant 6\ln(2d)$ concludes the proof. □

4 Adaptation to unknown U

In the previous section, the forecaster is given a radius $U > 0$ and asked to ensure a low worst-case regret on the ℓ^1-ball $B_1(U)$. In this section, U is no longer given: the forecaster is asked to be competitive against all balls $B_1(U)$, for $U > 0$. Namely, its worst-case regret on each $B_1(U)$ should be almost as good as if U were known beforehand. For simplicity, we assume that X and Y are known: we discuss in Section 5 how to simultaneously adapt to all parameters.

Parameters: $X, Y, \eta > 0$, $T \geqslant 1$, and $c > 0$ (a constant).
Initialization: $R = \lceil \log_2(2T/c) \rceil_+$, $\boldsymbol{w}_1 = \mathbf{1}/(R+1) \in \mathbb{R}^{R+1}$.
For time steps $t = 1, \ldots, T$:

1. For experts $r = 0, \ldots, R$:
   - Run the sub-algorithm $\mathcal{A}(U_r)$ on the ball $B_1(U_r)$ and obtain the prediction $\widehat{y}_t^{(r)}$.
2. Output the prediction $\widehat{y}_t = \sum_{r=0}^{R} \frac{w_t^{(r)}}{\sum_{r'=0}^{R} w_t^{(r')}} \big[\widehat{y}_t^{(r)}\big]_Y$.
3. Update $w_{t+1}^{(r)} = w_t^{(r)} \exp\Big(-\eta\big(y_t - \big[\widehat{y}_t^{(r)}\big]_Y\big)^2\Big)$ for $r = 0, \ldots, R$.

Fig. 4. The Scaling algorithm.

We define

$$R \triangleq \lceil \log_2(2T/c) \rceil_+ \quad \text{and} \quad U_r \triangleq \frac{Y}{X}\frac{2^r}{\sqrt{T\ln(2d)}}, \quad \text{for } r = 0, \ldots, R\,, \qquad (9)$$

where $c > 0$ is a known absolute constant and

$$\lceil x \rceil_+ \triangleq \min\bigl\{k \in \mathbb{N} : k \geqslant x\bigr\} \quad \text{for all } x \in \mathbb{R} \ .$$

The Scaling algorithm of Figure 4 works as follows. We have access to a sub-algorithm $\mathcal{A}(U)$ which we run simultaneously for all $U = U_r$, $r = 0, \ldots, R$. Each instance of the sub-algorithm $\mathcal{A}(U_r)$ performs online linear regression on the ℓ^1-ball $B_1(U_r)$. We employ an exponentially weighted forecaster to aggregate these $R+1$ sub-algorithms to perform online linear regression simultaneously on the balls $B_1(U_0), \ldots, B_1(U_R)$.

The following regret bound follows by exp-concavity of the square loss.

Theorem 4. *Suppose that $X, Y > 0$ are known. Let $c, c' > 0$ be two absolute constants. Suppose that for all $U > 0$, we have access to a sub-algorithm $\mathcal{A}(U)$ with regret against $B_1(U)$ of at most*

$$cUXY\sqrt{T\ln(2d)} + c'Y^2 \quad \textit{for } T \geqslant T_0 \ , \tag{10}$$

uniformly over all sequences $(\boldsymbol{x}_t)$ and (y_t) bounded by X and Y. Then, for a known $T \geqslant T_0$, the Scaling algorithm with $\eta = 1/(8Y^2)$ satisfies

$$\begin{aligned}\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 + 2c\left\|\boldsymbol{u}\right\|_1 XY\sqrt{T\ln(2d)} \right\} \\ + 8Y^2\ln\bigl(\lceil\log_2(2T/c)\rceil_+ + 1\bigr) + (c+c')Y^2. \end{aligned} \tag{11}$$

In particular, for every $U > 0$,

$$\begin{aligned}\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in B_1(U)} \left\{ \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \right\} + 2cUXY\sqrt{T\ln(2d)} \\ + 8Y^2\ln\bigl(\lceil\log_2(2T/c)\rceil_+ + 1\bigr) + (c+c')Y^2. \end{aligned}$$

Remark 1. By Theorem 3 the LEG algorithm satisfies assumption (10) with $T_0 = \ln(2d)$, $c \triangleq 9c_1 = 36\bigl(\sqrt{2}+1\bigr)$, and $c' \triangleq c_2 = 4\left(1 + 1/\sqrt{2}\right)^2$.

Proof. Since the Scaling algorithm is an exponentially weighted average forecaster (with clipping) applied to the $R+1$ experts $\mathcal{A}(U_r) = \bigl(\widehat{y}_t^{(r)}\bigr)_{t\geqslant 1}$, $r = 0, \ldots, R$, we have, by Lemma 3 in the appendix,

$$\begin{aligned}&\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \min_{r=0,\ldots,R} \sum_{t=1}^T \left(\widehat{y}_t^{(r)} - \widehat{y}_t\right)^2 + 8Y^2\ln(R+1) \\ &\leqslant \min_{r=0,\ldots,R} \left\{ \inf_{\boldsymbol{u}\in B_1(U_r)} \left\{ \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \right\} + cU_rXY\sqrt{T\ln(2d)} \right\} + z \ , \end{aligned} \tag{12}$$

where the last inequality follows by assumption (10), and where we set

$$z \triangleq 8Y^2\ln(R+1) + c'Y^2 \ .$$

Let $\boldsymbol{u}_T^* \in \arg\min_{\boldsymbol{u}\in\mathbb{R}^d} \left\{\sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 + 2c\left\|\boldsymbol{u}\right\|_1 XY\sqrt{T\ln(2d)}\right\}$. Next, we proceed by considering three cases: $U_0 < \left\|\boldsymbol{u}_T^*\right\|_1 < U_R$, $\left\|\boldsymbol{u}_T^*\right\|_1 \leqslant U_0$, and $\left\|\boldsymbol{u}_T^*\right\|_1 \geqslant U_R$.

Case 1: $U_0 < \left\|\boldsymbol{u}_T^*\right\|_1 < U_R$. Let $r^* \triangleq \min\{r = 0,\dots,R : U_r \geqslant \left\|\boldsymbol{u}_T^*\right\|_1\}$. Note that $r^* \geqslant 1$ since $\left\|\boldsymbol{u}_T^*\right\|_1 > U_0$. By (12) we have

$$\begin{aligned}\sum_{t=1}^T (y_t - \widehat{y}_t)^2 &\leqslant \inf_{\boldsymbol{u}\in B_1(U_{r^*})}\left\{\sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2\right\} + cU_{r^*}XY\sqrt{T\ln(2d)} + z \\ &\leqslant \sum_{t=1}^T (y_t - \boldsymbol{u}_T^*\cdot\boldsymbol{x}_t)^2 + 2c\left\|\boldsymbol{u}_T^*\right\|_1 XY\sqrt{T\ln(2d)} + z\ ,\end{aligned}$$

where the last inequality follows from $\boldsymbol{u}_T^* \in B_1(U_{r^*})$ and from the fact that $U_{r^*} \leqslant 2\left\|\boldsymbol{u}_T^*\right\|_1$ (since, by definition of r^*, $\left\|\boldsymbol{u}_T^*\right\|_1 > U_{r^*-1} = U_{r^*}/2$). Finally, we obtain (11) by definition of $\boldsymbol{u}_T^*$ and $z \triangleq 8Y^2\ln(R+1) + c'Y^2$.

Case 2: $\left\|\boldsymbol{u}_T^*\right\|_1 \leqslant U_0$. By (12) we have

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \left\{\sum_{t=1}^T (y_t - \boldsymbol{u}_T^*\cdot\boldsymbol{x}_t)^2 + cU_0XY\sqrt{T\ln(2d)}\right\} + z\ , \quad (13)$$

which yields (11) since $cU_0XY\sqrt{T\ln(2d)} = cY^2$ (by definition of U_0), by adding $2c\left\|\boldsymbol{u}_T^*\right\|_1 XY\sqrt{T\ln(2d)} \geqslant 0$, and by definition of $\boldsymbol{u}_T^*$ and z.

Case 3: $\left\|\boldsymbol{u}_T^*\right\|_1 \geqslant U_R$. By construction, we have $\widehat{y}_t \in [-Y, Y]$, and by assumption, we have $y_t \in [-Y, Y]$, so that

$$\begin{aligned}\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant 4Y^2T &\leqslant \sum_{t=1}^T (y_t - \boldsymbol{u}_T^*\cdot\boldsymbol{x}_t)^2 + 2cU_RXY\sqrt{T\ln(2d)} \\ &\leqslant \sum_{t=1}^T (y_t - \boldsymbol{u}_T^*\cdot\boldsymbol{x}_t)^2 + 2c\left\|\boldsymbol{u}_T^*\right\|_1 XY\sqrt{T\ln(2d)}\ ,\end{aligned}$$

where the second inequality follows by $2cU_RXY\sqrt{T\ln(2d)} = 2cY^2 2^R \geqslant 4Y^2T$ (since $2^R \geqslant 2T/c$ by definition of R), and the last inequality uses the assumption $\left\|\boldsymbol{u}_T^*\right\|_1 \geqslant U_R$. We finally get (11) by definition of $\boldsymbol{u}_T^*$.

This concludes the proof of the first claim (11). The second claim follows by bounding $\left\|\boldsymbol{u}\right\|_1 \leqslant U$. □

5 Extension to a fully adaptive algorithm and other discussions

The Scaling algorithm of Section 4 uses prior knowledge of Y, Y/X, and T. In order to obtain a fully automatic algorithm, we need to adapt efficiently to these quantities. Adaptation to Y is possible via a technique of the LEG algorithm, i.e., by updating the clipping range B_t based on the past observations $|y_s|$, $s \leqslant t-1$.

In parallel to adapting to Y, adaptation to Y/X can be carried out as follows. We replace the exponential sequence $\{U_0, \dots, U_R\}$ by another exponential sequence $\{U'_0, \dots, U'_{R'}\}$:

$$U'_r \triangleq \frac{1}{T^k} \frac{2^r}{\sqrt{T \ln(2d)}} \ , \quad r = 0, \dots, R' \ , \tag{14}$$

where $R' \triangleq R + \lceil \log_2 T^{2k} \rceil = \lceil \log_2(2T/c) \rceil_+ + \lceil \log_2 T^{2k} \rceil$, and where $k > 1$ is a fixed constant. On the one hand, for $T \geqslant \max\{(X/Y)^{1/k}, (Y/X)^{1/k}\}$, we have (cf. (9) and (14)),

$$[U_0, U_R] \subset [U'_0, U'_{R'}] \ .$$

Therefore, the argument in the proof of Theorem 4 applied to the grid $\{U'_0, \dots, U_{R'}\}$ yields[9] a regret bound of the order of $UXY\sqrt{T \ln d} + Y^2 \ln(R' + 1)$. On the other hand, clipping the predictions ensures a regret of $Y^2 T$. Hence, the overall regret for all $T \geqslant 1$ is of the order of

$$UXY\sqrt{T \ln d} + Y^2 \ln(k \ln T) + Y^2 \max\{(X/Y)^{1/k}, (Y/X)^{1/k}\} \ .$$

Adaptation to an unknown time horizon T can be carried out via a standard doubling trick on T. However, to avoid restarting the algorithm repeatedly, we can use a time-varying exponential sequence $\{U'_{-R'(t)}(t), \dots, U'_{R'(t)}(t)\}$ with a length $R'(t)$ that grows at the rate of $k \ln t$. This gives[10] us an algorithm that is fully automatic in the parameters U, X, Y and T. In this case, we can show that the regret is of the order of

$$UXY\sqrt{T \ln d} + Y^2 k(\ln T) + Y^2 \max\left\{\left(\sqrt{T}X/Y\right)^{1/k}, \left(Y/(\sqrt{T}X)\right)^{1/k}\right\},$$

where the last two terms are negligible when $T \to +\infty$ (since $k > 1$).

There is a logarithmic gap between the upper bound of Theorem 1 and the lower bound of Theorem 2. This gap comes from a concentration argument on a specific sequence of (unbounded) normal random variables in the proof of the lower bound. In the interval $\kappa \geqslant cd$, for some large enough absolute constant c, we can recover the missing $\ln(1 + 2\kappa)$ in our lower bound by using the argument of [Vov01, Theorem 2] instead. Another possible solution for $\kappa \leqslant cd$ is using a different sequence of random variables with bounded support, and the use of, e.g., Assouad's Lemma.

Acknowledgments

This work was supported in part by French National Research Agency (ANR, project EXPLO-RA, ANR-08-COSI-004) and the PASCAL2 Network of Excellence under EC grant no. 216886. J. Y. Yu was partly supported by a fellowship from Le Fonds québécois de la recherche sur la nature et les technologies.

[9] The proof remains the same by replacing $8Y^2 \ln(R + 1)$ with $8Y^2 \ln(R' + 1)$.

[10] Each time the exponential sequence (U'_r) expands, the weights assigned to the existing points U'_r are appropriately reassigned the whole new sequence.

References

AW01. K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, 43(3):211–246, 2001.

BM01. L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203–268, 2001.

BN08. F. Bunea and A. Nobel. Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Trans. Inform. Theory*, 54(4):1725–1735, 2008.

CB99. N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. *J. Comput. System Sci.*, 59(3):392–411, 1999.

CBL06. N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

CBLW96. N. Cesa-Bianchi, P. M. Long, and M. K. Warmuth. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks*, 7(3):604–619, 1996.

CBMS07. N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66(2/3):321–352, 2007.

DSSST10. J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT'10)*, pages 14–26, 2010.

Ger11. S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. Technical report, 2011. See `http://arxiv.org/abs/1101.1057`.

GL03. C. Gentile and N. Littlestone. The robustness of the p-norm algorithms. *Mach. Learn.*, 53(3):265–299, 2003.

GY11. S. Gerchinovitz and J. Y. Yu. Adaptive and optimal online linear regression on ℓ^1-balls. See http://www.math.ens.fr/∼gerchinovitz/docs/GY11.pdf, 2011.

KW97. J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Inform. and Comput.*, 132(1):1–63, 1997.

Mas07. P. Massart. *Concentration Inequalities and Model Selection*. Springer, Berlin/Heidelberg/New York, 2007.

Nem00. A. Nemirovski. *Topics in Non-Parametric Statistics*. Springer, Berlin/Heidelberg/New York, 2000.

RWY09. G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of convergence for high-dimensional regression under ℓ^q-ball sparsity. In *Proceedings of the 47th annual Allerton conference on communication, control, and computing (Allerton'09)*, pages 251–257, 2009.

SSSZ10. S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM J. Optim.*, 20(6):2807–2832, 2010.

SST09. S. Shalev-Shwartz and A. Tewari. Stochastic methods for ℓ^1-regularized loss minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, pages 929–936, 2009.

Tsy03. A. B. Tsybakov. Optimal rates of aggregation. In *Proceedings of the 16th Annual Conference on Learning Theory (COLT'03)*, pages 303–313, 2003.

Vov01. V. Vovk. Competitive on-line statistics. *Internat. Statist. Rev.*, 69:213–248, 2001.

A Proofs

A.1 Proof of Theorem 1

Proof. For each of the three cases distinguished in the statement of the theorem, we exhibit an algorithm whose worst-case regret on $B_1(U)$ is smaller than the stated upper bound.

Case 1: assume that $U < \frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}}$.
First note that if $U \geqslant (Y/X)\sqrt{T/(2\ln(2d))}$, then the upper bound $3UXY\sqrt{2T\ln(2d)} \geqslant 3TY^2 \geqslant TY^2$ is trivial (by choosing the predictor which outputs $\widehat{y}_t = 0$ at each time t).

We can thus assume that $U < (Y/X)\sqrt{T/(2\ln(2d))}$. Consider the EG$^{\pm}$ algorithm as given in [KW97, Theorem 5.11], and denote by $\widehat{\boldsymbol{u}}_t \in B_1(U)$ the linear combination it outputs at each time $t \geqslant 1$. Then, by the aforementioned theorem, this forecaster satisfies, uniformly over all individual sequences bounded by X and Y, that

$$\begin{aligned}
&\sum_{t=1}^{T}(y_t - \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T}(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \\
&\quad \leqslant 2UXY\sqrt{2T\ln(2d)} + 2U^2X^2\ln(2d) \\
&\quad \leqslant 2UXY\sqrt{2T\ln(2d)} + 2\left(Y\sqrt{\frac{T}{2\ln(2d)}}\right)UX\ln(2d) \qquad (15) \\
&\quad \leqslant 3UXY\sqrt{2T\ln(2d)}\ ,
\end{aligned}$$

where (15) follows from the assumption $UX < Y\sqrt{T/(2\ln(2d))}$. This concludes the first part of this proof.

Case 2: assume that $\frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}} \leqslant U \leqslant \frac{2dY}{\sqrt{T}X}$.
This part of the proof is based on a Maurey-type argument used under various forms, e.g., in [Nem00,Tsy03,BN08,SSSZ10]. It consists of discretizing $B_1(U)$ and looking at a sample average of random points in this discretization (see Lemma 4). We clip the prediction to get a regret bound growing as U instead of a naive U^2.

More precisely, we first use the fact that to be competitive against $B_1(U)$, it is sufficient to be competitive against its finite subset

$$\widetilde{B}_{U,m} \triangleq \left\{ \left(\frac{k_1 U}{m}, \dots, \frac{k_d U}{m} \right) : (k_1, \dots, k_d) \in \mathbb{Z}^d, \sum_{j=1}^{d} |k_j| \leqslant m \right\} \subset B_1(U) \,,$$

where $m \triangleq \lfloor \alpha \rfloor$ with $\alpha \triangleq \dfrac{UX}{Y} \sqrt{T(\ln 2) / \ln\left(1 + \dfrac{2dY}{\sqrt{T}UX}\right)}$.

By Lemma 4 in appendix, we indeed have

$$\begin{aligned}
&\inf_{\boldsymbol{u} \in \widetilde{B}_{U,m}} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \\
&\leqslant \inf_{\boldsymbol{u} \in B_1(U)} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 + \frac{TU^2X^2}{m} \\
&\leqslant \inf_{\boldsymbol{u} \in B_1(U)} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 + \frac{2}{\sqrt{\ln 2}} UXY \sqrt{T \ln\left(1 + \frac{2dY}{\sqrt{T}UX}\right)} \,, \quad (16)
\end{aligned}$$

where (16) follows from $m \triangleq \lfloor \alpha \rfloor \geqslant \alpha/2$ since $\alpha \geqslant 1$.

To see why $\alpha \geqslant 1$, note that it suffices to show that $x\sqrt{\ln(1+x)} \leqslant 2d\sqrt{\ln 2}$ where we set $x \triangleq 2dY/(\sqrt{T}UX)$. But from the assumption $U \geqslant (Y/X)\sqrt{\ln(1+2d)/(T \ln 2)}$, we have $x \leqslant 2d\sqrt{\ln(2)/\ln(1+2d)} \triangleq y$, so that, by monotonicity, $x\sqrt{\ln(1+x)} \leqslant y\sqrt{\ln(1+y)} \leqslant y\sqrt{\ln(1+2d)} = 2d\sqrt{\ln 2}$.

Therefore it only remains to exhibit an algorithm which is competitive against $\widetilde{B}_{U,m}$ at an aggregation price of the same order as the last term in (16). This is the case for the standard exponentially weighted average forecaster applied to the clipped predictions

$$\big[\boldsymbol{u} \cdot \boldsymbol{x}_t\big]_Y \triangleq \min\Big\{Y, \max\big\{-Y, \boldsymbol{u} \cdot \boldsymbol{x}_t\big\}\Big\} \,, \quad \boldsymbol{u} \in \widetilde{B}_{U,m} \,,$$

and tuned with the inverse temperature parameter $\eta = 1/(8Y^2)$. More formally, this algorithm predicts at each time $t = 1, \dots, T$ as

$$\widehat{y}_t \triangleq \sum_{\boldsymbol{u} \in \widetilde{B}_{U,m}} p_t(\boldsymbol{u}) \big[\boldsymbol{u} \cdot \boldsymbol{x}_t\big]_Y \,,$$

where $p_1(\boldsymbol{u}) \triangleq 1/\big|\widetilde{B}_{U,m}\big|$ (denoting by $\big|\widetilde{B}_{U,m}\big|$ the cardinality of the set $\widetilde{B}_{U,m}$), and where the weights $p_t(\boldsymbol{u})$ are defined for all $t = 2, \dots, T$ and $\boldsymbol{u} \in \widetilde{B}_{U,m}$ by

$$p_t(\boldsymbol{u}) \triangleq \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \big(y_s - [\boldsymbol{u} \cdot \boldsymbol{x}_s]_Y\big)^2\right)}{\sum_{\boldsymbol{v} \in \widetilde{B}_{U,m}} \exp\left(-\eta \sum_{s=1}^{t-1} \big(y_s - [\boldsymbol{v} \cdot \boldsymbol{x}_s]_Y\big)^2\right)} \,.$$

By Lemma 3 in appendix, the above forecaster tuned with $\eta = 1/(8Y^2)$ satisfies

$$\begin{aligned}
\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 - \inf_{\boldsymbol{u}\in\widetilde{B}_{U,m}} \sum_{t=1}^{T}(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 &\leqslant 8Y^2 \ln\bigl|\widetilde{B}_{U,m}\bigr| \\
&\leqslant 8Y^2 \ln\left(\frac{\mathrm{e}(2d+m)}{m}\right)^m && (17)\\
&= 8Y^2 m\bigl(1+\ln(1+2d/m)\bigr) \leqslant 8Y^2\alpha\bigl(1+\ln(1+2d/\alpha)\bigr) && (18)\\
&= 8Y^2\alpha + 8Y^2\alpha\ln\left(1+\frac{2dY}{\sqrt{T}UX}\sqrt{\frac{\ln\bigl(1+2dY/(\sqrt{T}UX)\bigr)}{\ln 2}}\right)\\
&\leqslant 8Y^2\alpha + 16Y^2\alpha\ln\left(1+\frac{2dY}{\sqrt{T}UX}\right) && (19)\\
&\leqslant \left(\frac{8}{\sqrt{\ln 2}} + 16\sqrt{\ln 2}\right) UXY\sqrt{T\ln\left(1+\frac{2dY}{\sqrt{T}UX}\right)}\,. && (20)
\end{aligned}$$

To get (17) we used Lemma 5 in appendix. Inequality (18) follows by definition of $m \leqslant \alpha$ and the fact that $x \mapsto x\bigl(1+\ln(1+A/x)\bigr)$ is nondecreasing on $\mathbb{R}_+^*$ for all $A > 0$. Inequality (19) follows from the assumption $U \leqslant 2dY/(\sqrt{T}X)$ and the elementary inequality $\ln\bigl(1+x\sqrt{\ln(1+x)/\ln 2}\bigr) \leqslant 2\ln(1+x)$ which holds for all $x \geqslant 1$ and was used, e.g., at the end of [BN08, Theorem 2-a)]. Finally, elementary manipulations combined with the assumption that $2dY/(\sqrt{T}UX) \geqslant 1$ lead to (20).

Putting Equations (16) and (20) together, the previous algorithm has a regret on $B_1(U)$ which is bounded from above by

$$\left(\frac{10}{\sqrt{\ln 2}} + 16\sqrt{\ln 2}\right) UXY\sqrt{T\ln\left(1+\frac{2dY}{\sqrt{T}UX}\right)}\,,$$

which concludes the second part of this proof since $10/\sqrt{\ln 2}+16\sqrt{\ln 2} \leqslant 26$.

Case 3: assume that $U > \frac{dY}{X\sqrt{T}}$.
The stated upper bound follows straightforwardly from the regret bound proved in [Ger11, Proposition 2] for the algorithm SeqSEW tuned with $\eta = 1/(8Y^2)$ and $\tau = Y/(\sqrt{T}X)$. This algorithm has indeed a cumulative square loss upper bounded by

$$\begin{aligned}
&\inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\bigl(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\bigr)^2 + 32\,\|\boldsymbol{u}\|_0\, Y^2\ln\left(1+\frac{\sqrt{T}X\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0\, Y}\right)\right\} + dY^2\\
&\leqslant \inf_{\|\boldsymbol{u}\|_1\leqslant U}\left\{\sum_{t=1}^{T}\bigl(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\bigr)^2\right\} + 32dY^2\ln\left(1+\frac{\sqrt{T}XU}{dY}\right) + dY^2\,,
\end{aligned}$$

where the last inequality follows by monotonicity[11] in $\|\boldsymbol{u}\|_0$ and $\|\boldsymbol{u}\|_1$ of the second term of the left-hand side. This concludes the proof. □

[11] Note that for all $A > 0$, the function $x \mapsto x\ln(1+A/x)$ (continuously extended at $x = 0$) has a nonnegative first derivative and is thus nondecreasing on $\mathbb{R}_+$.

A.2 Proof of Theorem 2

To prove Theorem 2, we perform a reduction to the stochastic batch setting (via the standard online to batch trick), and employ a version of the lower bound proved in [Tsy03] for convex aggregation.

We first need the following notations. Let (S, μ) be a probability space for which we can find an orthonormal family[12] $(\varphi_j)_{1 \leqslant j \leqslant d}$ in the space of square-integrable functions on S, which we denote by $\mathbb{L}^2(S, \mu)$ thereafter. For all $\boldsymbol{u} \in \mathbb{R}^d$ and $\gamma, \sigma > 0$, denote by $\mathbb{P}_{\boldsymbol{u}}^{\gamma,\sigma}$ the joint law of the i.i.d. sequence $(X_t, Y_t)_{1 \leqslant t \leqslant T}$ such that

$$Y_t = \gamma \varphi_{\boldsymbol{u}}(X_t) + \sigma \varepsilon_t \in \mathbb{R} \ ,$$

where $\varphi_{\boldsymbol{u}} \triangleq \sum_{j=1}^d u_j \varphi_j$, where the X_t are i.i.d points in S drawn from μ, and where the ε_t are i.i.d standard Gaussian random variables such that $(X_t)_{1 \leqslant t \leqslant T}$ and $(\varepsilon_t)_{1 \leqslant t \leqslant T}$ are independent.

The next lemma is a direct adaptation of [Tsy03, Theorem 2], which we state with our notations in a slightly more precise form (we make clear how the lower bound depends on the noise level σ and the signal level γ).

Lemma 1 (An extension of Theorem 2 of [Tsy03]).
Let $d \in \mathbb{N}^*$ and $\gamma, \sigma > 0$. Let (S, μ) be a probability space for which we can find an orthonormal family $(\varphi_j)_{1 \leqslant j \leqslant d}$ in $\mathbb{L}^2(S, \mu)$, and consider the Gaussian linear model described above. Then there exist absolute constants $c_4, c_5, c_6, c_7 > 0$ such that

$$\inf_{\widehat{f}_T} \sup_{\substack{\boldsymbol{u} \in \mathbb{R}^d_+ \\ \sum_j u_j \leqslant 1}} \left\{ \mathbb{E}_{\mathbb{P}_{\boldsymbol{u}}^{\gamma,\sigma}} \left\| \widehat{f}_T - \gamma \varphi_{\boldsymbol{u}} \right\|_\mu^2 \right\}$$
$$\geqslant \begin{cases} c_4 \frac{d\sigma^2}{T} & \text{if} \quad \frac{d}{\sqrt{T}} \leqslant c_5 \frac{\gamma}{\sigma} \ , \\ c_6 \gamma \sigma \sqrt{\frac{1}{T} \ln\left(1 + \frac{d\sigma}{\sqrt{T}\gamma}\right)} & \text{if} \quad c_5 \frac{\gamma}{\sigma} < \frac{d}{\sqrt{T}} \leqslant c_7 \frac{\gamma d}{\sigma \sqrt{\ln(1+d)}} \ , \end{cases}$$

where the infimum is taken over all estimators $\widehat{f}_T : S \to \mathbb{R}$, and where $\|f\|_\mu^2 \triangleq \int_S f(x)^2 \mu(dx)$ for all measurable functions $f : S \to \mathbb{R}$.

Note that the lower bound we stated in Theorem 2 is very similar to T times the above lower bound with $\gamma \sim X$ and $\sigma \sim Y$ (recall that $\kappa \triangleq \sqrt{T}UX/(2dY)$). The main difference is that the latter holds for unbounded observations, while we need bounded observations y_t, $1 \leqslant t \leqslant T$. A simple concentration argument will show that these observations lie in $[-Y, Y]$ with high probability, which will yield the desired lower bound. The proof of Theorem 2 thus consists of the following steps:

– step 1: reduction to the stochastic batch setting;

[12] An example is given by $S = [-\pi, \pi]$, $\mu(\mathrm{d}x) = \mathrm{d}x/(2\pi)$, and $\varphi_j(x) = \sqrt{2}\sin(jx)$ for all $1 \leqslant j \leqslant d$ and $x \in [-\pi, \pi]$. We will use this particular case later.

– step 2: application of Lemma 1;
– step 3: concentration argument.

Proof (of Theorem 2). We first assume that $\sqrt{\ln(1+2d)}/\big(2d\sqrt{\ln 2}\big) \leqslant \kappa \leqslant 1$. The case when $\kappa > 1$ will easily follow from the monotonicity of the minimax regret in κ. We set

$$T \triangleq 1 + \big\lceil (4d\kappa)^2 \big\rceil \;, \quad U \triangleq 1 \;, \quad \text{and} \quad X \triangleq \frac{2d\kappa Y}{\sqrt{T}} \;, \tag{21}$$

so that $T \geqslant 2$, $\sqrt{T}UX/(2dY) = \kappa$, and $X \leqslant Y/2$ (since $\sqrt{T} \geqslant 4d\kappa$).

Step 1: reduction to the stochastic batch setting.
First note that by clipping to $[-Y, Y]$, we have

$$\begin{aligned}
&\inf_{(\widetilde{f}_t)_t} \sup_{\substack{\|\boldsymbol{x}_t\|_\infty \leqslant X \\ |y_t| \leqslant Y}} \left\{ \sum_{t=1}^T \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \right\} \\
&= \inf_{\substack{(\widetilde{f}_t)_t \\ |\widetilde{f}_t| \leqslant Y}} \sup_{\substack{\|\boldsymbol{x}_t\|_\infty \leqslant X \\ |y_t| \leqslant Y}} \left\{ \sum_{t=1}^T \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \right\},
\end{aligned} \tag{22}$$

where the first infimum is taken over all online forecasters[13] $(\widetilde{f}_t)_t$, where the second infimum is restricted to online forecasters $(\widetilde{f}_t)_t$ which output predictions in $[-Y, Y]$, and where both suprema are taken over all individual sequences $(\boldsymbol{x}_t, y_t)_{1\leqslant t\leqslant T} \in (\mathbb{R}^d \times \mathbb{R})^T$ such that $|y_1|, \dots, |y_T| \leqslant Y$ and $\|\boldsymbol{x}_1\|_\infty, \dots, \|\boldsymbol{x}_T\|_\infty \leqslant X$.

Next we use the standard online to batch trick to bound from below the right-hand side of (22) by T times the lower bound of Lemma 1, which we apply to the particular case where $S = [-\pi, \pi]$, $\mu(\mathrm{d}x) = \mathrm{d}x/(2\pi)$, and $\varphi_j(x) = \sqrt{2}\sin(jx)$ for all $1 \leqslant j \leqslant d$ and $x \in [-\pi, \pi]$. Let

$$\gamma \triangleq c_8 X \quad \text{and} \quad \sigma \triangleq \frac{c_9 Y}{\sqrt{\ln T}} \;, \tag{23}$$

for some absolute constants $c_8, c_9 > 0$ to be chosen by the analysis.

Let $(\widetilde{f}_t)_{t\geqslant 1}$ be any online forecaster whose predictions lie in $[-Y, Y]$, and consider the estimator $\widehat{f}_T$ defined for each sample $(X_t, Y_t)_{1\leqslant t\leqslant T}$ and each new input X by

$$\widehat{f}_T\Big(X; (X_t, Y_t)_{1\leqslant t\leqslant T}\Big) \triangleq \frac{1}{T}\sum_{t=1}^T \widetilde{f}_t\Big(\gamma\boldsymbol{\varphi}(X); (\gamma\boldsymbol{\varphi}(X_s), Y_s)_{1\leqslant s\leqslant t-1}\Big) \;, \tag{24}$$

where $\boldsymbol{\varphi} \triangleq (\varphi_1, \dots, \varphi_d)$, and where we explicitely wrote all the dependencies[13] of the $\widetilde{f}_t$, $1 \leqslant t \leqslant T$.

[13] An online forecaster is a sequence of functions $(\widetilde{f}_t)_{t\geqslant 1}$, where $\widetilde{f}_t : \mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R})^{t-1} \to \mathbb{R}$ maps at time t the new input $\boldsymbol{x}_t$ and the past data $(\boldsymbol{x}_s, y_s)_{1\leqslant s\leqslant t-1}$ to a prediction $\widetilde{f}_t\big(\boldsymbol{x}_t; (\boldsymbol{x}_s, y_s)_{1\leqslant s\leqslant t-1}\big)$. However, unless mentioned otherwise, we omit the dependence in $(\boldsymbol{x}_s, y_s)_{1\leqslant s\leqslant t-1}$, and only write $\widetilde{f}_t(\boldsymbol{x}_t)$.

Take $\boldsymbol{u}^* \in \mathbb{R}^d_+$ achieving the supremum[14] in Lemma 1 for the estimator $\widehat{f}_T$. Accordingly, consider the random sequence $(\boldsymbol{x}_t, y_t)_{1 \leqslant t \leqslant T}$ in $\mathbb{R}^d \times \mathbb{R}$ defined for all $t = 1, \dots, T$ by

$$\boldsymbol{x}_t \triangleq \big(\gamma\varphi_1(X_t), \dots, \gamma\varphi_d(X_t)\big) \quad \text{and} \quad y_t \triangleq \gamma\varphi_{\boldsymbol{u}^*}(X_t) + \sigma\varepsilon_t \ ,$$

where $\varphi_{\boldsymbol{u}^*} \triangleq \sum_{j=1}^d u_j^* \varphi_j$, where the X_t are i.i.d points in $[-\pi, \pi]$ drawn from the uniform distribution $\mu(\mathrm{d}x) = \mathrm{d}x/(2\pi)$, and where the ε_t are i.i.d standard Gaussian random variables such that $(X_t)_t$ and $(\varepsilon_t)_t$ are independent. All the expectations below should thus be understood as being taken with respect to the probability distribution $\mathbb{P}^{\gamma,\sigma}_{\boldsymbol{u}^*}$.

Let (X', y') be a random copie of (X_1, y_1) independent of $(X_t, y_t)_{1 \leqslant t \leqslant T}$, and set $\boldsymbol{x}' \triangleq \big(\gamma\varphi_1(X'), \dots, \gamma\varphi_d(X')\big)$. By standard manipulations (conditioning on the past and applying Jensen's inequality), we get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant 1} \sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2\right] \\
&\quad \geqslant T\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\Big[\big(y' - \widetilde{f}_t(\boldsymbol{x}')\big)^2\Big] - \inf_{\|\boldsymbol{u}\|_1 \leqslant 1} \mathbb{E}\Big[\big(y' - \boldsymbol{u}\cdot\boldsymbol{x}'\big)^2\Big]\right) \\
&\quad \geqslant T\left(\mathbb{E}\Big[\big(y' - \widehat{f}_T(X')\big)^2\Big] - \inf_{\|\boldsymbol{u}\|_1 \leqslant 1} \mathbb{E}\Big[\big(y' - \boldsymbol{u}\cdot\boldsymbol{x}'\big)^2\Big]\right) \qquad (25) \\
&\quad = T\,\mathbb{E}\left\|\widehat{f}_T - \gamma\varphi_{\boldsymbol{u}^*}\right\|_\mu^2 \ . \qquad (26)
\end{aligned}$$

Inequality (25) follows by definition of $\widehat{f}_T$ (see (24)) and by Jensen's inequality. Inequality (26) follows by expanding the square $\big(y' - \widehat{f}_t(X')\big)^2 = \big(\gamma\varphi_{\boldsymbol{u}^*}(X') - \widehat{f}_T(X') + y' - \gamma\varphi_{\boldsymbol{u}^*}(X')\big)^2$ and by the fact that

$$\inf_{\|\boldsymbol{u}\|_1 \leqslant 1} \mathbb{E}\Big[\big(y' - \boldsymbol{u}\cdot\boldsymbol{x}'\big)^2\Big] = \mathbb{E}\Big[\big(y' - \gamma\varphi_{\boldsymbol{u}^*}(X')\big)^2\Big] \ ,$$

where we used $\|\boldsymbol{u}^*\|_1 \leqslant 1$ (by definition of $\boldsymbol{u}^*$) and $\boldsymbol{u}\cdot\boldsymbol{x}' = \gamma\varphi_{\boldsymbol{u}}(X')$.

Step 2: application of Lemma 1.
Next we use Lemma 1 to bound from below the right-hand side of (26). By Lemma 1 and by definition of $\boldsymbol{u}^*$, we have

$$\begin{aligned}
&\mathbb{E}\left\|\widehat{f}_T - \gamma\varphi_{\boldsymbol{u}^*}\right\|_\mu^2 \\
&\geqslant \begin{cases} c_4 \frac{d\sigma^2}{T} & \text{if} \quad \frac{d}{\sqrt{T}} \leqslant c_5 \frac{\gamma}{\sigma} \ , \\ c_6 \gamma\sigma \sqrt{\frac{1}{T}\ln\left(1 + \frac{d\sigma}{\sqrt{T}\gamma}\right)} & \text{if} \quad c_5\frac{\gamma}{\sigma} < \frac{d}{\sqrt{T}} \leqslant \frac{c_7\gamma d}{\sigma\sqrt{\ln(1+d)}} \ . \end{cases} \\
&\geqslant \begin{cases} \frac{c_4 c_9^2}{T(\ln T)} dY^2 & \text{if} \quad \frac{d}{\sqrt{T}} \leqslant c_5 \frac{\gamma}{\sigma} \ , \\ \frac{c_6 c_8 c_9}{\sqrt{\ln T}} UXY \sqrt{\frac{1}{T}\ln\left(1 + \frac{c_9 dY}{c_8\sqrt{T(\ln T)}UX}\right)} & \text{if} \quad c_5\frac{\gamma}{\sigma} < \frac{d}{\sqrt{T}} \leqslant \frac{c_7\gamma d}{\sigma\sqrt{\ln(1+d)}} \ , \end{cases} \qquad (27)
\end{aligned}$$

[14] If the supremum in Lemma 1 is not achieved, then we can instead take an ε-almost-maximizer for any $\varepsilon > 0$. Letting $\varepsilon \to 0$ in the end will conclude the proof.

where the last inequality follows from (23) and from $U = 1$.

Next we bound the two expressions of the right-hand side of (27) from below by a single quantity. First, by definition of $T \triangleq 1 + \lceil (4d\kappa)^2 \rceil$ and by the assumption $\sqrt{\ln(1+2d)}/\big(2d\sqrt{\ln 2}\big) \leqslant \kappa$, elementary manipulations show that the condition

$$\frac{d}{\sqrt{T}} \leqslant \frac{c_7 \gamma d}{\sigma \sqrt{\ln(1+d)}} \tag{28}$$

holds true whenever[15] $c_9 \leqslant c_7 c_8 c_{10}$, where $c_{10} \triangleq \frac{1}{2} \inf_{x \geqslant 2\sqrt{\frac{\ln 3}{\ln 2}}} \left\{ \frac{x}{\sqrt{1+\lceil x^2 \rceil}} \right\}$ (note that $c_{10} > 0$). Moreover, note that if $c_9 \leqslant c_8 2\sqrt{\ln 2}$, then $c_8 \geqslant c_9/(2\sqrt{\ln 2}) \geqslant c_9/(2\sqrt{\ln T})$. In this case, since $x \mapsto x\sqrt{\ln(1+A/x)}$ is nondecreasing on $\mathbb{R}_+^*$ for all $A > 0$, we can replace c_8 with $c_9/(2\sqrt{\ln T})$ in the next expression and get

$$\begin{aligned}
&\frac{c_6 c_8 c_9}{\sqrt{\ln T}} UXY \sqrt{\frac{1}{T} \ln\left(1 + \frac{c_9 dY}{c_8 \sqrt{T(\ln T)} UX}\right)} \\
&\quad \geqslant \frac{c_6 c_9^2}{2 \ln T} UXY \sqrt{\frac{1}{T} \ln\left(1 + \frac{2dY}{\sqrt{T} UX}\right)} = \frac{c_6 c_9^2}{T(\ln T)} dY^2 \kappa \sqrt{\ln(1+1/\kappa)} \,,
\end{aligned}$$

where we used the definition of $\kappa \triangleq \sqrt{T} UX/(2dY)$.
In the sequel we will choose the absolute constants c_8 and c_9 such that

$$c_9 \leqslant c_7 c_8 c_{10} \quad \text{and} \quad c_9 \leqslant c_8 2\sqrt{\ln 2} \,. \tag{29}$$

Therefore, by the above remarks, by the fact that $\ln T \triangleq \ln\big(1 + \lceil (4d\kappa)^2 \rceil\big) \leqslant \ln\big(2 + 16d^2\big)$ (since $\kappa \leqslant 1$ by assumption), and multiplying both sides of (27) by T, we get

$$\begin{aligned}
&T\, \mathbb{E} \left\| \widehat{f}_T - \gamma \varphi_{\boldsymbol{u}^*} \right\|_\mu^2 \\
&\quad \geqslant \begin{cases} \frac{c_4 c_9^2}{\ln\left(2+16d^2\right)} dY^2 & \text{if} \quad \frac{d}{\sqrt{T}} \leqslant c_5 \frac{\gamma}{\sigma} \,, \\ \frac{c_6 c_9^2}{\ln\left(2+16d^2\right)} dY^2 \kappa \sqrt{\ln(1+1/\kappa)} & \text{if} \quad c_5 \frac{\gamma}{\sigma} < \frac{d}{\sqrt{T}} \,. \end{cases} \\
&\quad \geqslant \frac{c_{11} c_9^2}{\ln\big(2+16d^2\big)} dY^2 \kappa \sqrt{\ln(1+1/\kappa)} \,,
\end{aligned} \tag{30}$$

where we set $c_{11} \triangleq \min\big\{c_4/\sqrt{\ln 2}, c_6\big\}$. The last inequality follows by the fact that the function $x \mapsto x\sqrt{\ln(1+1/x)}$ is nondecreasing on $\mathbb{R}_+^*$, so that its value at $x = \kappa \leqslant 1$ is smaller than $\sqrt{\ln 2}$.

[15] By definition of γ and σ, (28) is equivalent to $T \ln T \geqslant c_9^2/(c_7^2 c_8^2) Y^2/X^2 \ln(1+d)$. But by definition of X and by the assumption $\kappa \geqslant \sqrt{\ln(1+2d)}/(2d\sqrt{\ln 2})$, we have $Y/X \leqslant c_{10}$. Therefore, (28) is implied by $T \ln T \geqslant c_9^2/(c_7^2 c_8^2 c_{10}^2) \ln(1+d)$, which in turns is implied by the condition $c_9 \leqslant c_7 c_8 c_{10}$ (by definition of T).

Combining (26) and (30), we get

$$\mathbb{E}\left[\sum_{t=1}^{T}\big(y_t-\widetilde{f}_t(\boldsymbol{x}_t)\big)^2-\inf_{\|\boldsymbol{u}\|_1\leqslant 1}\sum_{t=1}^{T}\big(y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2\right]$$
$$\geqslant \frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)}dY^2\kappa\sqrt{\ln(1+1/\kappa)}\;. \tag{31}$$

Step 3: concentration argument.
Note that the standard trick consisting in upper bounding the last expectation by the supremum over all individual sequences does not conclude the proof immediately, since the random observations y_t lie outside of $[-Y,Y]$ with positive probability.

Next we prove that this probability is actually small, so that with high probability, the random sequence $(\boldsymbol{x}_t,y_t)_{1\leqslant t\leqslant T}$ lies in $[-X,X]^d\times[-Y,Y]$. The desired lower bound then follows by noting that with high probability the regret on this random sequence is at least as large as (half of) its expectation. More formally, define the event

$$\mathcal{A}\triangleq\bigcap_{t=1}^{T}\big\{|y_t|\leqslant Y\big\}\;.$$

Set $\widehat{L}_T\triangleq\sum_{t=1}^{T}\big(y_t-\widetilde{f}_t(\boldsymbol{x}_t)\big)^2$ and $L_T(\boldsymbol{u})\triangleq\sum_{t=1}^{T}\big(y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2$ for all $\boldsymbol{u}\in\mathbb{R}^d$. Denote by $\mathcal{A}^c$ the complement of $\mathcal{A}$, and by $\mathbb{I}_{\mathcal{A}}$ and $\mathbb{I}_{\mathcal{A}^c}$ the corresponding indicator functions. We have

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}_{\mathcal{A}}\left(\widehat{L}_T-\inf_{\|\boldsymbol{u}\|_1\leqslant 1}L_T(\boldsymbol{u})\right)\right]\\
&\quad=\mathbb{E}\left[\widehat{L}_T-\inf_{\|\boldsymbol{u}\|_1\leqslant 1}L_T(\boldsymbol{u})\right]-\mathbb{E}\left[\mathbb{I}_{\mathcal{A}^c}\left(\widehat{L}_T-\inf_{\|\boldsymbol{u}\|_1\leqslant 1}L_T(\boldsymbol{u})\right)\right]\\
&\quad\geqslant\frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)}dY^2\kappa\sqrt{\ln(1+1/\kappa)}-\mathbb{E}\Big[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\Big]\;,
\end{aligned}\tag{32}$$

where the last inequality follows by (31) and by the fact that $L_T(\boldsymbol{u})\geqslant 0$ for all $\boldsymbol{u}\in\mathbb{R}^d$. The rest of the proof is dedicated to upper bound the above quantity $\mathbb{E}\big[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\big]$ by half the term on his left.

We have

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\Big]&\triangleq\mathbb{E}\left[\mathbb{I}_{\mathcal{A}^c}\sum_{t=1}^{T}\big(y_t-\widetilde{f}_t(\boldsymbol{x}_t)\big)^2\right]\\
&\leqslant\mathbb{E}\left[\mathbb{I}_{\mathcal{A}^c}\sum_{t=1}^{T}\Big(4Y^2\mathbb{I}_{\{|y_t|\leqslant Y\}}+\big(y_t-\widetilde{f}_t(\boldsymbol{x}_t)\big)^2\mathbb{I}_{\{|y_t|>Y\}}\Big)\right] &(33)\\
&\leqslant 4TY^2\mathbb{P}\big(\mathcal{A}^c\big)+\sum_{t=1}^{T}\mathbb{E}\Big[\big(y_t-\widetilde{f}_t(\boldsymbol{x}_t)\big)^2\mathbb{I}_{\left\{|\varepsilon_t|>\frac{Y}{2\sigma}\right\}}\Big]\;, &(34)
\end{aligned}$$

where (33) follows from the fact that the online forecaster $(\widetilde{f}_t)_t$ outputs its predictions in $[-Y,Y]$. As for Inequality (34), note by definition of y_t

that $|y_t| \leqslant \|\boldsymbol{u}^*\|_1 \gamma \|\boldsymbol{\varphi}(X_t)\|_\infty + \sigma|\varepsilon_t| \leqslant \gamma\sqrt{2} + \sigma|\varepsilon_t|$ since $\|\boldsymbol{u}^*\|_1 \leqslant 1$ and $|\varphi_j(x)| \triangleq |\sqrt{2}\sin(jx)| \leqslant \sqrt{2}$ for all $j = 1, \dots, d$ and $x \in \mathbb{R}$. Therefore, by definition of $\gamma \triangleq c_9 X$, and since $X \leqslant Y/2$ (by definition of X), we get $|y_t| \leqslant c_9\sqrt{2}\, Y/2 + \sigma|\varepsilon_t| \leqslant Y/2 + \sigma|\varepsilon_t|$ provided that

$$c_9 \leqslant \frac{1}{\sqrt{2}} \,, \tag{35}$$

which we assume thereafter. The above remarks show that $\{|y_t| > Y\} \subset \{|\varepsilon_t| > Y/(2\sigma)\}$, which entails (34). By the same comments and since $|\widetilde{f}_t| \leqslant Y$, we have, for all $t = 1, \dots, T$,

$$\begin{aligned}
\mathbb{E}\Big[\big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 \mathbb{I}_{\{|\varepsilon_t| > \frac{Y}{2\sigma}\}}\Big] &\leqslant \mathbb{E}\Big[\big(Y/2 + \sigma|\varepsilon_t| + Y\big)^2 \mathbb{I}_{\{|\varepsilon_t| > \frac{Y}{2\sigma}\}}\Big] \\
&\leqslant 2\left(\frac{3Y}{2}\right)^2 \mathbb{P}\left(|\varepsilon_t| > \frac{Y}{2\sigma}\right) + 2\sigma^2 \mathbb{E}\Big[\varepsilon_t^2 \mathbb{I}_{\{|\varepsilon_t| > \frac{Y}{2\sigma}\}}\Big] \qquad (36) \\
&\leqslant \frac{9Y^2}{2}\mathbb{P}\left(|\varepsilon_t| > \frac{Y}{2\sigma}\right) + 2\sigma^2\sqrt{3}\,\mathbb{P}^{1/2}\left(|\varepsilon_t| > \frac{Y}{2\sigma}\right) \qquad (37) \\
&\leqslant 9Y^2 T^{-1/(8c_9^2)} + 2\frac{c_9^2 Y^2}{\ln 2}\sqrt{6}\, T^{-1/(16c_9^2)} \,. \qquad (38)
\end{aligned}$$

Inequality (36) follows by the elementary inequality $(a+b)^2 \leqslant 2(a^2+b^2)$ for all $a, b \in \mathbb{R}$. To get (37) we used the Cauchy-Schwarz inequality and the fact that $\mathbb{E}\big[\varepsilon_t^4\big] = 3$ (since ε_t is a standard Gaussian random variable). Finally, (38) follows by definition of $\sigma \triangleq c_9 Y/\sqrt{\ln T} \leqslant c_9 Y/\sqrt{\ln 2}$ and from the fact that, since ε_t is a standard Gaussian random variable,

$$\mathbb{P}\left(|\varepsilon_t| > \frac{Y}{2\sigma}\right) \leqslant 2e^{-\frac{1}{2}\left(\frac{Y}{2\sigma}\right)^2} = 2e^{-\frac{1}{2}\left(\frac{\sqrt{\ln T}}{2c_9}\right)^2} = 2T^{-1/(8c_9^2)} \,. \tag{39}$$

Using the fact that $\mathbb{P}\big(\mathcal{A}^c\big) \leqslant \sum_{t=1}^T \mathbb{P}\big(|y_t| > Y\big) \leqslant \sum_{t=1}^T \mathbb{P}\big(|\varepsilon_t| > Y/(2\sigma)\big) \leqslant 2T^{1-1/(8c_9^2)}$ by the inequality above, we can substitute (38) in (34) to get

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\Big] \\
&\quad\leqslant 8Y^2 T^{2-1/(8c_9^2)} + 9Y^2 T^{1-1/(8c_9^2)} + \frac{2c_9^2\sqrt{6}}{\ln 2} Y^2 T^{1-1/(16c_9^2)} \\
&\quad\leqslant 8Y^2 2^{2-1/(8c_9^2)} + 9Y^2 2^{1-1/(8c_9^2)} + \frac{2c_9^2\sqrt{6}}{\ln 2} Y^2 2^{1-1/(16c_9^2)} \,, \qquad (40)
\end{aligned}$$

where the last inequality follows from the fact that $T^\alpha \leqslant 2^\alpha$ for all $\alpha < 0$ (since $T \geqslant 2$) and from a choice of c_9 such that $c_9 < 1/4$ (which we assume thereafter).

Next we show that by choosing the absolute constant c_9 small enough, the last upper bound is smaller than half the lower bound of (31). Note that, since $x \mapsto x\sqrt{\ln(1+1/x)}$ is nondecreasing on $\mathbb{R}_+^*$ and since $\kappa \geqslant \kappa_{\min} \triangleq \sqrt{\ln(1+2d)}/(2d\sqrt{\ln 2})$ by assumption, the latter lower bound

can be further bounded from below by

$$\begin{aligned}
&\frac{c_{11}c_9^2}{\ln\bigl(2+16d^2\bigr)}dY^2\kappa\sqrt{\ln(1+1/\kappa)} \\
&\quad\geqslant \frac{c_{11}c_9^2}{\ln\bigl(2+16d^2\bigr)}dY^2\kappa_{\min}\sqrt{\ln(1+1/\kappa_{\min})} \\
&\quad= \frac{c_{11}c_9^2}{2\sqrt{\ln 2}}Y^2\frac{\sqrt{\ln(1+2d)}\sqrt{\ln\Bigl[1+2d\sqrt{\ln 2}/\sqrt{\ln(1+2d)}\Bigr]}}{\ln\bigl(2+16d^2\bigr)} \\
&\quad\geqslant \frac{c_{11}c_9^2}{2\sqrt{\ln 2}}Y^2c_{12}\ ,
\end{aligned}$$

where c_{12} denotes the infimum of the last fraction over all $d \geqslant 1$; in particular, $c_{13} > 0$. It is now easy to see that by choosing the absolute constant c_9 small enough (say $c_9 \leqslant c_{13}$), the right-hand side of (40) is smaller than half the right-hand side of the last inequality. Therefore, choosing c_9 and $c_8 \triangleq \max\bigl\{c_9/(2\sqrt{\ln 2}), c_9/(c_7c_{10})\bigr\}$ such that $c_9 < 1/\sqrt{2}$ (condition (35)), $c_9 < 1/4$, and $c_9 \leqslant c_{13}$, then the condition (29) also holds and we get

$$\mathbb{E}\Bigl[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\Bigr] \leqslant \frac{1}{2}\frac{c_{11}c_9^2}{\ln\bigl(2+16d^2\bigr)}dY^2\kappa\sqrt{\ln(1+1/\kappa)}\ .$$

Substituing the last inequality in (32), we get that

$$\mathbb{E}\biggl[\mathbb{I}_{\mathcal{A}}\left(\widehat{L}_T - \inf_{\|\boldsymbol{u}\|_1\leqslant 1} L_T(\boldsymbol{u})\right)\biggr] \geqslant \frac{1}{2}\frac{c_{11}c_9^2}{\ln\bigl(2+16d^2\bigr)}dY^2\kappa\sqrt{\ln(1+1/\kappa)}\ .$$

We can finally upper bound the last expectation by the supremum over all individual sequences. This tells us that there is an individual sequence $(\boldsymbol{x}_t, y_t)_{1\leqslant t\leqslant T}$ in $\mathbb{R}^d \times \mathbb{R}$ which lies in $\mathcal{A} \triangleq \bigcap_{t=1}^T \bigl\{|y_t| \leqslant Y\bigr\}$ and whose regret is at least as large as the above lower bound. Noting in addition that for all $t = 1, \dots, T$, we have $\|\boldsymbol{x}_t\|_\infty \leqslant \gamma\sqrt{2} \leqslant X$ (since $\gamma \triangleq c_9X$ and $c_9 \leqslant 1/\sqrt{2}$), and setting $c_1 \triangleq c_{11}c_9^2/2$, we have proved, for all online forecasters $(\widetilde{f}_t)_{t\geqslant 1}$ whose predictions lie in $[-Y, Y]$, that

$$\begin{aligned}
&\sup_{\substack{\|\boldsymbol{x}_t\|_\infty\leqslant X\\ |y_t|\leqslant Y}}\left\{\sum_{t=1}^T\bigl(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\bigr)^2 - \inf_{\|\boldsymbol{u}\|_1\leqslant U}\sum_{t=1}^T(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2\right\} \\
&\qquad\geqslant \frac{c_1}{\ln\bigl(2+16d^2\bigr)}dY^2\sqrt{\ln\left(1+1/\kappa\right)}\ ,
\end{aligned}$$

which concludes the proof when $\kappa_{\min} \leqslant \kappa \leqslant 1$ by (22).

Assume now that $\kappa > 1$.
The stated lower bound follows from the case when $\kappa = 1$ and by monotonicity of the minimax regret in κ (when d and Y are kept constant).

More formally, by the first part of this proof (when $\kappa = 1$), we can fix $T \geqslant 1$, $U_1 > 0$, and $X > 0$ such that $\sqrt{T}U_1X/(2dY) = 1$ and

$$\inf_{(\widetilde{f}_t)_t} \sup_{\substack{\|\boldsymbol{x}_t\|_\infty \leqslant X \\ |y_t| \leqslant Y}} \left\{ \sum_{t=1}^T \bigl(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\bigr)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U_1} \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \right\}$$
$$\geqslant \frac{c_1}{\ln\bigl(2 + 16d^2\bigr)} dY^2 \sqrt{\ln 2} \; ,$$

where the infimum is taken over all online forecasters $(\widetilde{f}_t)_{t \geqslant 1}$, and where the supremum is taken over all individual sequences bounded by X and Y.

Now take $\kappa > 1$, and set $U \triangleq \kappa U_1 > U$, so that $\sqrt{T}UX/(2dY) = \kappa$ (since $\sqrt{T}U_1X/(2dY) = 1$). Moreover, for all individual sequences bounded by X and Y, the regret on $B_1(U)$ is at least as large as the regret on $B_1(U_1)$ (since $U > U_1$). Combining the latter remark with the lower bound above and setting $c_2 \triangleq c_1\sqrt{\ln 2}$ concludes the proof. □

B Lemmas

The following lemma is a straightforward[16] consequence of [CBMS07, Corollary 1]. We use the same notations as in Section 3.

Lemma 2 (Corollary 1 of [CBMS07]). *The weighted majority forecaster used in Section 3.2 satisfies*

$$\sum_{t=1}^T \sum_{\substack{1 \leqslant j \leqslant d \\ \varepsilon \in \{0,1\}}} w_{2j-1+\varepsilon,t} (-1)^\varepsilon \nabla \widetilde{\ell}_t\bigl(\widehat{\boldsymbol{u}}_t\bigr) \cdot U \boldsymbol{e}_j - \min_{\substack{1 \leqslant j \leqslant d \\ \varepsilon \in \{0,1\}}} \sum_{t=1}^T (-1)^\varepsilon \nabla \widetilde{\ell}_t\bigl(\widehat{\boldsymbol{u}}_t\bigr) \cdot U \boldsymbol{e}_j$$
$$\leqslant U \max_{1 \leqslant t \leqslant T} \left\| \nabla \widetilde{\ell}_t \right\|_\infty \left(2\sqrt{T \ln(2d)} + 4 \ln(2d) + 6 \right) \; .$$

Next we recall a regret bound satisfied by the standard exponentially weighted average forecaster applied to clipped base forecasts. Assume that at each time $t = 1, \ldots, T$, the forecaster has access to $K \geqslant 1$ base forecasts $\widehat{y}_t^{(k)} \in \mathbb{R}$, $k = 1, \ldots, K$, and that for some known bound $Y > 0$ on the observations, the forecaster predicts at time $t = 1, \ldots, T$ as

$$\widehat{y}_t \triangleq \sum_{k=1}^K p_{k,t} \bigl[\widehat{y}_t^{(k)}\bigr]_Y \; ,$$

where $[x]_Y \triangleq \min\{Y, \max\{-Y, x\}\}$ for all $x \in \mathbb{R}$, and where the weight vectors $\boldsymbol{p}_t \in \mathbb{R}^K$ are given by $\boldsymbol{p}_1 = (1/K, \ldots, 1/K) \in \mathbb{R}^K$ and, for all

[16] The weight vectors $\boldsymbol{w}_t \in \mathbb{R}^{2d}$ used in Section 3 are exactly the weight vectors of the exponentially weighted forecaster of [CBMS07, Corollary 1] when applied to the loss vectors $\left(\nabla \widetilde{\ell}_t\bigl(\widehat{\boldsymbol{u}}_t\bigr) \cdot U \boldsymbol{e}_j, -\nabla \widetilde{\ell}_t\bigl(\widehat{\boldsymbol{u}}_t\bigr) \cdot U \boldsymbol{e}_j \right)_{1 \leqslant j \leqslant d} \in \mathbb{R}^{2d}$, $t = 1, \ldots, T$.

$t = 2, \dots, T$, by

$$p_{k,t} \triangleq \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(y_s - \left[\widehat{y}_s^{(k)}\right]_Y\right)^2\right)}{\sum_{j=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} \left(y_s - \left[\widehat{y}_s^{(j)}\right]_Y\right)^2\right)} \;, \quad 1 \leqslant k \leqslant K \;,$$

for some inverse temperature parameter $\eta > 0$ to be chosen below. The next lemma is a straigthforward consequence of Theorem 3.2 and Proposition 3.1 of [CBL06].

Lemma 3 (Exponential weighting with clipping). *Assume that the forecaster knows beforehand a bound $Y > 0$ on the observations $|y_t|$, $t = 1, \dots, T$. Then, the exponentially weighted average forecaster tuned with $\eta \leqslant 1/(8Y^2)$ and with clipping $[\cdot]_Y$ satisfies*

$$\sum_{t=1}^{T} \left(y_t - \widehat{y}_t\right)^2 \leqslant \min_{1 \leqslant k \leqslant K} \sum_{t=1}^{T} \left(y_t - \widehat{y}_t^{(k)}\right)^2 + \frac{\ln K}{\eta} \;.$$

Proof (of Lemma 3). The proof follows straightforwardly from Theorem 3.2 and Proposition 3.1 of [CBL06]. To apply the latter result, recall from [Vov01, Remark 3] that the square loss is $1/(8Y^2)$-exp-concave on $[-Y, Y]$ and thus η-exp-concave[17] (since $\eta \leqslant 1/(8Y^2)$ by assumption). Therefore, by definition of our forecaster above, Proposition 3.1 of [CBL06] yields

$$\sum_{t=1}^{T} \left(y_t - \widehat{y}_t\right)^2 \leqslant \min_{1 \leqslant k \leqslant K} \sum_{t=1}^{T} \left(y_t - \left[\widehat{y}_t^{(k)}\right]_Y\right)^2 + \frac{\ln K}{\eta} \;.$$

To conclude the proof, note for all $t = 1, \dots, T$ and $k = 1, \dots, K$ that $|y_t| \leqslant Y$ by assumption, so that clipping the base forecasts to $[-Y, Y]$ can only improve prediction, i.e., $\left(y_t - \left[\widehat{y}_t^{(k)}\right]_Y\right)^2 \leqslant \left(y_t - \widehat{y}_t^{(k)}\right)^2$. □

C Additional tools

The next approximation argument is originally due to Maurey, and was used under various forms, e.g., in [Nem00,Tsy03,BN08,SSSZ10].

Lemma 4 (Approximation argument). *Let $U > 0$ and $m \in \mathbb{N}^*$. Define the following finite subset of $B_1(U)$,*

$$\widetilde{B}_{U,m} \triangleq \left\{ \left(\frac{k_1 U}{m}, \dots, \frac{k_d U}{m} \right) : (k_1, \dots, k_d) \in \mathbb{Z}^d, \sum_{j=1}^{d} |k_j| \leqslant m \right\} \subset B_1(U) \;.$$

Then, for all $(\boldsymbol{x}_t, y_t)_{1 \leqslant t \leqslant T} \in \left(\mathbb{R}^d \times \mathbb{R}\right)^T$ such that $\max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty \leqslant X$,

$$\inf_{\boldsymbol{u} \in \widetilde{B}_{U,m}} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \leqslant \inf_{\boldsymbol{u} \in B_1(U)} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 + \frac{T U^2 X^2}{m} \;.$$

[17] This means that for all $y \in [-Y, Y]$, the function $x \mapsto \exp\left(-\eta(y - x)^2\right)$ is concave on $[-Y, Y]$.

Proof. The proof is quite standard and follows the same lines as [Nem00, Proposition 5.2.2] or [BN08, Theorem 2] who addressed the aggregation task in the stochastic setting. We rewrite this argument below in our online deterministic setting.

Fix $\boldsymbol{u}^* \in \operatorname{argmin}_{\boldsymbol{u}\in B_1(U)} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2$. Define the probability distribution $\pi = (\pi_{-d}, \dots, \pi_d) \in \mathbb{R}_+^{2d+1}$ by

$$\pi_j \triangleq \begin{cases} \dfrac{(u_j^*)_+}{U} & \text{if } j \geqslant 1; \\ \dfrac{(u_j^*)_-}{U} & \text{if } j \leqslant -1; \\ 1 - \displaystyle\sum_{j=1}^d \frac{|u_j^*|}{U} & \text{if } j = 0\ . \end{cases}$$

Let $J_1, \dots, J_m \in \{-d, \dots, d\}$ be i.i.d. random integers drawn from π, and set

$$\widetilde{\boldsymbol{u}} \triangleq \frac{U}{m} \sum_{k=1}^m \boldsymbol{e}_{J_k}\ ,$$

where $(\mathbf{e}_j)_{1\leqslant j\leqslant d}$ is the canonical basis of $\mathbb{R}^d$, where $\boldsymbol{e}_0 \triangleq \mathbf{0}$, and where $\boldsymbol{e}_{-j} \triangleq -\boldsymbol{e}_j$ for all $1 \leqslant j \leqslant d$.

Note that $\widetilde{\boldsymbol{u}} \in \widetilde{B}_{U,m}$ by construction. Therefore,

$$\inf_{\boldsymbol{u}\in\widetilde{B}_{U,m}} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \leqslant \mathbb{E}\left[\sum_{t=1}^T (y_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t)^2\right]\ . \tag{41}$$

The rest of the proof is dedicated to upper bounding the last expectation. Expanding all the squares $(y_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t)^2 = (y_t - \boldsymbol{u}^*\cdot\boldsymbol{x}_t + \boldsymbol{u}^*\cdot\boldsymbol{x}_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t)^2$, first note that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T (y_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t)^2\right] &= \sum_{t=1}^T (y_t - \boldsymbol{u}^*\cdot\boldsymbol{x}_t)^2 + \sum_{t=1}^T \mathbb{E}\big[(\boldsymbol{u}^*\cdot\boldsymbol{x}_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t)^2\big] \\ &\quad + 2\sum_{t=1}^T (y_t - \boldsymbol{u}^*\cdot\boldsymbol{x}_t)\,\mathbb{E}\big[\boldsymbol{u}^*\cdot\boldsymbol{x}_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t\big]\ . \end{aligned}\tag{42}$$

But by definition of $\widetilde{\boldsymbol{u}}$ and π,

$$\begin{aligned}\mathbb{E}\big[\widetilde{\boldsymbol{u}}\big] &= U\,\mathbb{E}\big[\boldsymbol{e}_{J_1}\big] = U\sum_{j=-d}^d \pi_j \boldsymbol{e}_j \\ &= U\sum_{j=1}^d \left(\frac{(u_j^*)_+}{U}\boldsymbol{e}_j + \frac{(u_j^*)_-}{U}(-\boldsymbol{e}_j)\right) = U\sum_{j=1}^d \frac{u_j^*}{U}\boldsymbol{e}_j = \boldsymbol{u}^*\ ,\end{aligned}$$

so that $\mathbb{E}\big[\widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t\big] = \boldsymbol{u}^*\cdot\boldsymbol{x}_t$ for all $1 \leqslant t \leqslant T$. Therefore, the last sum in (42) above equals zero, and

$$\mathbb{E}\Big[\big(\boldsymbol{u}^*\cdot\boldsymbol{x}_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t\big)^2\Big] = \operatorname{Var}\big(\widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t\big) = \frac{U^2}{m^2}\sum_{k=1}^m \operatorname{Var}\big(\boldsymbol{e}_{J_k}\cdot\boldsymbol{x}_t\big) \leqslant \frac{U^2X^2}{m}\ ,$$

where the second equality follows from $\widetilde{\boldsymbol{u}} \cdot \boldsymbol{x}_t = (U/m) \sum_{k=1}^m \boldsymbol{e}_{J_k} \cdot \boldsymbol{x}_t$ and from the independence of the J_k, $1 \leqslant k \leqslant m$, and where the last inequality follows from $|\boldsymbol{e}_{J_k} \cdot \boldsymbol{x}_t| \leqslant \|\boldsymbol{e}_{J_k}\|_1 \|\boldsymbol{x}_t\|_\infty \leqslant X$ for all $1 \leqslant k \leqslant m$.

Combining (42) with the remarks above, we get

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T (y_t - \widetilde{\boldsymbol{u}} \cdot \boldsymbol{x}_t)^2\right] &\leqslant \sum_{t=1}^T (y_t - \boldsymbol{u}^* \cdot \boldsymbol{x}_t)^2 + \frac{TU^2X^2}{m} \\ &= \inf_{\boldsymbol{u} \in B_1(U)} \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 + \frac{TU^2X^2}{m} \ ,\end{aligned}$$

where the last line follows by definition of $\boldsymbol{u}^*$. Substituting the last inequality in (41) concludes the proof. □

The combinatorial result below (or variants of it) is well-known; see, e.g., [Tsy03,BN08]. We reproduce its proof for the convenience of the reader.

Lemma 5 (An elementary combinatorial upper bound). *Let $m, d \in \mathbb{N}^*$. Denoting by $|E|$ the cardinality of a set E, we have*

$$\left|\left\{(k_1, \dots, k_d) \in \mathbb{Z}^d : \sum_{j=1}^d |k_j| \leqslant m\right\}\right| \leqslant \left(\frac{\mathrm{e}(2d+m)}{m}\right)^m \ .$$

Proof (of Lemma 5). Setting $(k'_{-j}, k'_j) \triangleq \big((k_j)_-, (k_j)_+\big)$ for all $1 \leqslant j \leqslant d$, and $k'_0 \triangleq 1 - \sum_{j=1}^d |k_j|$, we have

$$\begin{aligned}&\left|\left\{(k_1, \dots, k_d) \in \mathbb{Z}^d : \sum_{j=1}^d |k_j| \leqslant m\right\}\right| \\ &\quad\leqslant \left|\left\{(k'_{-d}, \dots, k'_d) \in \mathbb{N}^{2d+1} : \sum_{j=-d}^d k'_j = m\right\}\right| \\ &\quad= \binom{2d+m}{m} \qquad (43) \\ &\quad\leqslant \left(\frac{\mathrm{e}(2d+m)}{m}\right)^m \ . \qquad (44)\end{aligned}$$

To get inequality (43), we used the (elementary) fact that the number of $2d+1$ integer-valued tuples summing up to m is equal to the number of lattice paths from $(1,0)$ to $(2d+1, m)$ in $\mathbb{N}^2$, which is equal to $\binom{2d+1+m-1}{m}$. As for inequality (44), it follows straightforwardly from a classical combinatorial result stated, e.g., in [Mas07, Proposition 2.5]. □