Adaptive and optimal online linear regression on ℓ^1-balls

Sébastien Gerchinovitz, Jia Yuan Yu

Adaptive and optimal online linear regression on ℓ^1-balls

Sébastien Gerchinovitz[a,1,*], Jia Yuan Yu[b]

[a] *École Normale Supérieure, 45 rue d'Ulm, 75005 Paris, France*
[b] *IBM Research, Damastown Technology Campus, Dublin 15, Ireland*

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 d, T, and 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., that do not require the knowledge of U, X, Y, and T, but still achieve nearly optimal regret bounds.

Keywords: Online learning, Linear regression, Adaptive algorithms, Minimax regret

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 larger than the number of data points. Examples include analysis of DNA sequences, collaborative filtering, astronomical data analysis, and cross-country growth regression. In such high-dimensional problems, performing linear regression on an ℓ^1-ball of small diameter may be helpful if the best linear 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 $(y_t)_{t\geqslant 1}$ 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.

In the sequel, $\boldsymbol{u} \cdot \boldsymbol{v}$ denotes the standard inner product between $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, and we set $\|\boldsymbol{u}\|_\infty \triangleq \max_{1\leqslant j\leqslant d} |u_j|$ and $\|\boldsymbol{u}\|_1 \triangleq \sum_{j=1}^d |u_j|$. The ℓ^1-ball of radius $U > 0$ is the following bounded subset of $\mathbb{R}^d$:

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

*Corresponding author
Email addresses: sebastien.gerchinovitz@ens.fr (Sébastien Gerchinovitz), jiayuanyu@ie.ibm.com (Jia Yuan Yu)

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

Given a fixed radius $U > 0$ and a time horizon $T \geqslant 1$, the goal of the forecaster is to predict almost as well as the best linear forecaster in the reference set $\left\{\boldsymbol{x} \in \mathbb{R}^d \mapsto \boldsymbol{u} \cdot \boldsymbol{x} \in \mathbb{R} : \boldsymbol{u} \in B_1(U)\right\}$, 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[2] $(\boldsymbol{x}_t, y_t)_{1 \leqslant t \leqslant T}$ 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 depend on four important quantities: U, X, Y, and T, which may be known or unknown to the forecaster.

1.2. Contributions and related works

In the next paragraphs we detail the main contributions of this paper in view of related works in online linear regression.

Our first contribution (Section 2) consists of a minimax analysis of online linear regression on ℓ^1-balls in the arbitrary sequence setting. We first provide a refined regret bound 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[3] at $\kappa = 1$ or $d = \sqrt{T}UX/(2Y)$. Namely, for $\kappa < 1$, the regret is of the order of $dY^2\kappa$ (proportional to $\sqrt{T}$), whereas it is of the order of $dY^2 \ln \kappa$ (proportional to $\ln T$) for $\kappa > 1$.

The derivation of this regret bound partially relies on a Maurey-type argument used under various forms with i.i.d. data, e.g., in [1, 2, 3, 4] (see also [5]). We adapt it in a straightforward way to the deterministic setting. Therefore, this is yet another technique that can be applied to both the stochastic and individual sequence settings.

Unsurprisingly, the refined regret bound mentioned above matches the optimal risk bounds for stochastic settings[4] [6, 2] (see also [7]). Hence, linear regression is just as hard in the stochastic setting as in the arbitrary sequence setting. Using the standard online to batch conversion, we make the latter statement more precise by establishing a lower bound for all κ at least of the order of $\sqrt{\ln d}/d$. This lower bound extends those of [8, 9], which only hold for small κ of the order of $1/d$.

The algorithm achieving our minimax regret bound is both computationally inefficient and non-adaptive (i.e., it requires prior knowledge of the quantities U, X, Y, and T that may be unknown in practice). Those two issues were first overcome by [10] via an automatic tuning termed *self-confident* (since the forecaster somehow trusts himself in tuning its parameters). They indeed proved that the self-confident p-norm algorithm with $p = 2 \ln d$ and tuned with U has a cumulative loss $\widehat{L}_T = \sum_{t=1}^{T}(y_t - \widehat{y}_t)^2$ bounded by

$$\begin{aligned}\widehat{L}_T &\leqslant L_T^* + 8UX\sqrt{(e \ln d)\, L_T^*} + (32e \ln d)\, U^2X^2 \\ &\leqslant 8UXY\sqrt{eT \ln d} + (32e \ln d)\, U^2X^2 \ ,\end{aligned}$$

where $L_T^* \triangleq \min_{\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U\}} \sum_{t=1}^{T}(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \leqslant TY^2$. This algorithm is efficient, and our lower bound in terms of κ shows that it is optimal up to logarithmic factors in the regime $\kappa \leqslant 1$ without prior knowledge of X, Y, and T.

Our second contribution (Section 3) is to show that similar adaptivity and efficiency properties can be obtained via exponential weighting. We consider a variant of the EG$^{\pm}$ algorithm [9]. The latter has a manageable computational complexity and our lower bound shows that it is nearly optimal in the regime

[2] 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.

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

[4] 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.

$\kappa \leqslant 1$. However, the EG$^{\pm}$ algorithm requires prior knowledge of U, X, Y, and T. To overcome this adaptivity issue, we study a modification of the EG$^{\pm}$ algorithm that relies on the variance-based automatic tuning of [11]. The resulting algorithm – called *adaptive EG$^{\pm}$ algorithm* – can be applied to general convex and differentiable loss functions. When applied to the square loss, it yields an algorithm of the same computational complexity as the EG$^{\pm}$ algorithm that also achieves a nearly optimal regret but without needing to know X, Y, and T beforehand.

Our third contribution (Section 3.3) is a generic technique called *loss Lipschitzification*. It transforms the loss functions $\boldsymbol{u} \mapsto (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$ (or $\boldsymbol{u} \mapsto \left|y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t\right|^{\alpha}$ if the predictions are scored with the α-loss for a real number $\alpha \geqslant 2$) into Lipschitz continuous functions. We illustrate this technique by applying the generic adaptive EG$^{\pm}$ algorithm to the modified loss functions. When the predictions are scored with the square loss, this yields an algorithm (the LEG algorithm) whose main regret term slightly improves on that derived for the adaptive EG$^{\pm}$ algorithm without Lipschtizification. The benefits of this technique are clearer for loss functions with higher curvature: if $\alpha > 2$, then the resulting regret bound roughly grows as U instead of a naive $U^{\alpha/2}$.

Finally, in Section 4, we provide a simple way to achieve minimax regret uniformly over all ℓ^1-balls $B_1(U)$ for $U > 0$. This method aggregates instances of an algorithm that requires prior knowledge of U. For the sake of simplicity, we assume that X, Y, and T are known, but explain in the discussions how to extend the method to a fully adaptive algorithm that requires the knowledge neither of U, X, Y, nor T.

This 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 — the adaptive EG$^{\pm}$ algorithm with or without loss Lipschitzification — that achieves the optimal regret on $B_1(U)$ when U is known. In Section 4, we use an aggregating strategy to achieve an optimal regret uniformly over all ℓ^1-balls $B_1(U)$, for $U>0$, when X, Y, and T are known. Finally, in Section 5, we discuss as an extension a fully automatic algorithm that requires no prior knowledge of U, X, Y, or T. Some proofs and additional tools are postponed to the appendix.

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.

We consider the following definition to avoid any ambiguity. We call *online forecaster* any sequence $F = (\widetilde{f}_t)_{t \geqslant 1}$ of functions such that $\widetilde{f}_t : \mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R})^{t-1} \to \mathbb{R}$ 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 $\widetilde{f}_t\big(\boldsymbol{x}_t; (\boldsymbol{x}_s, y_s)_{1 \leqslant s \leqslant t-1}\big)$. Depending on the context, the latter prediction may be simply denoted by $\widetilde{f}_t(\boldsymbol{x}_t)$ or by $\widehat{y}_t$.

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 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 $|y_1|, \dots, |y_T| \leqslant Y$ and $\|\boldsymbol{x}_1\|_\infty, \dots, \|\boldsymbol{x}_T\|_\infty \leqslant X$.

Theorem 1 improves the bound of [9, 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 provides a partial answer to a question[5] raised in [9] about the gap of $\sqrt{\ln(2d)}$ between the upper and lower bounds.

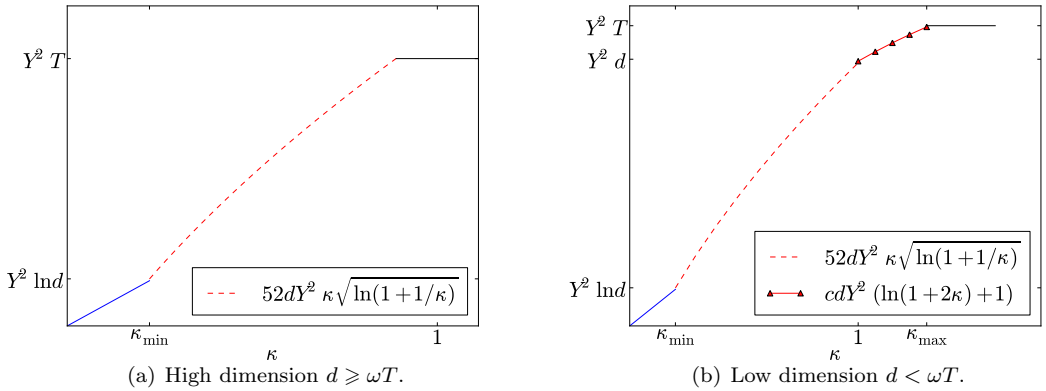
Before proving the theorem (see below), we state the following immediate corollary. It expresses the upper bound of Theorem 1 in terms of an intrinsic quantity $\kappa \triangleq \sqrt{T}UX/(2dY)$ that relates $\sqrt{T}UX/(2Y)$ to the ambient dimension d.

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 parametrization by (d, Y, κ) helps to unify the different upper bounds of Theorem 1: on both regimes $\kappa \leqslant 1$ and $\kappa > 1$, the regret bound scales as dY^2, the only difference lies in the dependence in κ (linear versus logarithmic).

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 $\kappa \geqslant 1$.



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

Figure 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 $\kappa_{\max} = (e^{(T/d-1)/c} - 1)/2$.

[5] The authors of [9] asked: "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."

We now prove Theorem 1. The main part of the proof relies on a Maurey-type argument. Although this argument was used in the stochastic setting [1, 2, 3, 4], we adapt it to the deterministic setting. This is yet another technique that can be applied to both the stochastic and individual sequence settings.

Proof (of Theorem 1): First note from Lemma 5 in Appendix B that the minimax regret on $B_1(U)$ is upper bounded[6] by

$$\min\left\{3UXY\sqrt{2T\ln(2d)},\, 32\,dY^2\ln\left(1+\frac{\sqrt{T}UX}{dY}\right)+dY^2\right\}\ . \tag{2}$$

Therefore, the first case $U < \frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}}$ and the third case $U > \frac{dY}{X\sqrt{T}}$ are straightforward.

Therefore, we assume in the sequel that $\frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}} \leqslant U \leqslant \frac{2dY}{\sqrt{T}X}$.
We use a Maurey-type argument to refine the regret bound (2). This technique was used under various forms in the stochastic setting, e.g., in [1, 2, 3, 4]. It consists of discretizing $B_1(U)$ and looking at a random point in this discretization to study its approximation properties. We also use clipping 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

$$\widetilde{B}_{U,m} \triangleq \left\{\left(\frac{k_1U}{m},\dots,\frac{k_dU}{m}\right) : (k_1,\dots,k_d)\in\mathbb{Z}^d, \sum_{j=1}^d |k_j| \leqslant m\right\} \subset B_1(U)\ ,$$

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 7 in Appendix C, and since $m > 0$ (see below), we indeed have

$$\begin{aligned}\inf_{\boldsymbol{u}\in\widetilde{B}_{U,m}}\sum_{t=1}^T(y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t)^2 &\leqslant \inf_{\boldsymbol{u}\in B_1(U)}\sum_{t=1}^T(y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t)^2+\frac{TU^2X^2}{m}\\ &\leqslant \inf_{\boldsymbol{u}\in B_1(U)}\sum_{t=1}^T(y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t)^2+\frac{2}{\sqrt{\ln 2}}\,UXY\sqrt{T\ln\left(1+\frac{2dY}{\sqrt{T}UX}\right)}\ ,\end{aligned} \tag{3}$$

where (3) follows from $m \triangleq \lfloor\alpha\rfloor \geqslant \alpha/2$ since $\alpha \geqslant 1$ (in particular, $m > 0$ as stated above).

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 (3). This is the case for the standard exponentially weighted average forecaster applied to the clipped predictions

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

[6] As proved in Lemma 5, the regret bound (2) is achieved either by the EG$^{\pm}$ algorithm, the algorithm SeqSEW$^{B,\eta}_{\tau}$ of [12] (we could also get a slightly worse bound with the sequential ridge regression forecaster [13, 14]), or the trivial null forecaster.

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

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

By Lemma 6 in Appendix B, the above forecaster tuned with $\eta = 1/(8Y^2)$ satisfies

$$\begin{aligned}
\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 - \inf_{\boldsymbol{u} \in \widetilde{B}_{U,m}} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 &\leqslant 8Y^2 \ln\big|\widetilde{B}_{U,m}\big| \\
&\leqslant 8Y^2 \ln\left(\frac{\mathrm{e}(2d+m)}{m}\right)^m && (4) \\
&= 8Y^2 m\big(1 + \ln(1 + 2d/m)\big) \leqslant 8Y^2 \alpha\big(1 + \ln(1 + 2d/\alpha)\big) && (5) \\
&= 8Y^2\alpha + 8Y^2\alpha \ln\left(1 + \frac{2dY}{\sqrt{T}UX} \sqrt{\frac{\ln\big(1 + 2dY/(\sqrt{T}UX)\big)}{\ln 2}}\right) \\
&\leqslant 8Y^2\alpha + 16Y^2\alpha \ln\left(1 + \frac{2dY}{\sqrt{T}UX}\right) && (6) \\
&\leqslant \left(\frac{8}{\sqrt{\ln 2}} + 16\sqrt{\ln 2}\right) UXY \sqrt{T \ln\left(1 + \frac{2dY}{\sqrt{T}UX}\right)} \ . && (7)
\end{aligned}$$

To get (4) we used Lemma 8 in Appendix C. Inequality (5) follows by definition of $m \leqslant \alpha$ and the fact that $x \mapsto x\big(1 + \ln(1 + A/x)\big)$ is nondecreasing on $\mathbb{R}_+^*$ for all $A > 0$. Inequality (6) follows from the assumption $U \leqslant 2dY/(\sqrt{T}X)$ and the elementary inequality $\ln\big(1 + x\sqrt{\ln(1+x)/\ln 2}\big) \leqslant 2\ln(1+x)$ which holds for all $x \geqslant 1$ and was used, e.g., at the end of [3, Theorem 2-a)]. Finally, elementary manipulations combined with the assumption that $2dY/(\sqrt{T}UX) \geqslant 1$ lead to (7).

Putting Eqs. (3) and (7) 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 proof since $10/\sqrt{\ln 2} + 16\sqrt{\ln 2} \leqslant 26$. □

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[7] up to logarithmic factors.

[7] 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).

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 is taken 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 [8, 9], which hold for small κ of the order of $1/d$. The proof is postponed to Appendix A.1. We perform a reduction to the stochastic batch setting—via the standard online to batch conversion—and employ a version of a lower bound of [2].

Note that in the proof of Theorem 2, we are free to choose the values of two parameters among T, U, and X, provided that $\sqrt{T}UX/(2dY) = \kappa$. This liberty is possible since the problem is now parametrized by d, Y, and κ only (as shown in Corollary 1, these three parameters are sufficient to express the regret bound of Theorem 1, and they actually help to unify the upper bounds of the two regimes). A more ambitious lower bound would consist in proving that the upper bound of Theorem 1 cannot be substantially improved for any fixed value of (d, Y, T, U, X). This question is left for future work.

3. Adaptation to unknown X, Y and T via exponential weights

Although the proof of Theorem 1 already gives an algorithm that achieves the minimax regret, the latter takes as inputs U, X, Y, and T, 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, Y, and T provided that U is known. Adaptation to an unknown U is considered in Section 4. Our method consists of modifying an underlying efficient linear regression algorithm such as the EG$^\pm$ algorithm [9] or the sequential ridge regression forecaster [14, 13]. Next, we show that automatically tuned variants of the EG$^\pm$ algorithm nearly achieve the minimax regret for the regime $d \geqslant \sqrt{T}UX/(2Y)$. A similar modification could be applied to the ridge regression forecaster — with a total computational efficiency of the same order as that of the standard ridge algorithm — 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 hence is omitted.

3.1. An adaptive EG$^\pm$ algorithm for general convex and differentiable loss functions

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 at each time t. Next we introduce a version of the EG$^\pm$ algorithm — called the *adaptive EG$^\pm$ algorithm* — that does not require prior knowledge of X, Y and T (as opposed to the original EG$^\pm$ algorithm of [9]). This version relies on the automatic tuning of [11]. We first present a generic version suited for general convex and differentiable loss functions. The application to the square loss and to other α-losses will be dealt with in Sections 3.2 and 3.3.

The generic setting with arbitrary convex and differentiable loss functions corresponds to the online convex optimization setting [15, 16] and unfolds as follows: at each time $t \geqslant 1$, the forecaster chooses a linear combination $\widehat{\boldsymbol{u}}_t \in \mathbb{R}^d$, then the environment chooses and reveals a convex and differentiable loss function $\ell_t : \mathbb{R}^d \to \mathbb{R}$, and the forecaster incurs the loss $\ell_t(\widehat{\boldsymbol{u}}_t)$. In online linear regression under the square loss, the loss functions are given by $\ell_t(\boldsymbol{u}) = (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$.

Parameter: radius $U > 0$.

Initialization: $\boldsymbol{p}_1 = (p^+_{1,1}, p^-_{1,1}, \dots, p^+_{d,1}, p^-_{d,1}) \triangleq \big(1/(2d), \dots, 1/(2d)\big) \in \mathbb{R}^{2d}$.

At each time round $t \geqslant 1$,

1. Output the linear combination $\widehat{\boldsymbol{u}}_t \triangleq U \sum_{j=1}^{d} \big(p^+_{j,t} - p^-_{j,t}\big)\, \boldsymbol{e}_j \in B_1(U)$;
2. Receive the loss function $\ell_t : \mathbb{R}^d \to \mathbb{R}$ and update the parameter η_{t+1} according to (8);
3. Update the weight vector $\boldsymbol{p}_{t+1} = (p^+_{1,t+1}, p^-_{1,t+1}, \dots, p^+_{d,t+1}, p^-_{d,t+1}) \in \mathcal{X}_{2d}$ defined for all $j = 1, \dots, d$ and $\gamma \in \{+, -\}$ by[a]

$$p^\gamma_{j,t+1} \triangleq \frac{\exp\left(-\eta_{t+1} \sum_{s=1}^{t} \gamma U \nabla_j \ell_s(\widehat{\boldsymbol{u}}_s)\right)}{\sum_{\substack{1 \leqslant k \leqslant d \\ \mu \in \{+,-\}}} \exp\left(-\eta_{t+1} \sum_{s=1}^{t} \mu U \nabla_k \ell_s(\widehat{\boldsymbol{u}}_s)\right)} \,.$$

[a] For all $\gamma \in \{+, -\}$, by a slight abuse of notation, γU denotes U or $-U$ if $\gamma = +$ or $\gamma = -$ respectively.

Figure 2: The adaptive EG$^\pm$ algorithm for general convex and differentiable loss functions (see Proposition 1).

The adaptive EG$^\pm$ algorithm for general convex and differentiable loss functions is defined in Figure 2. We denote by $(\mathbf{e}_j)_{1 \leqslant j \leqslant d}$ the canonical basis of $\mathbb{R}^d$, by $\nabla \ell_t(\boldsymbol{u})$ the gradient of ℓ_t at $\boldsymbol{u} \in \mathbb{R}^d$, and by $\nabla_j \ell_t(\boldsymbol{u})$ the j-th component of this gradient. The adaptive EG$^\pm$ algorithm uses as a blackbox the exponentially weighted majority forecaster of [11] on $2d$ experts — namely, the vertices $\pm U \boldsymbol{e}_j$ of $B_1(U)$ — as in [9]. It adapts to the unknown gradient amplitudes $\|\nabla \ell_t\|_\infty$ by the particular choice of η_t due to [11] and defined for all $t \geqslant 2$ by

$$\eta_t = \min\left\{ \frac{1}{\widehat{E}_{t-1}}, C\sqrt{\frac{\ln(2d)}{V_{t-1}}} \right\} \,, \tag{8}$$

where $C \triangleq \sqrt{2(\sqrt{2}-1)/(\mathrm{e}-2)}$ and where we set, for all $t = 1, \dots, T$,

$$z^+_{j,s} \triangleq U \nabla_j \ell_s(\widehat{\boldsymbol{u}}_s) \quad \text{and} \quad z^-_{j,s} \triangleq -U \nabla_j \ell_s(\widehat{\boldsymbol{u}}_s) \,, \quad j = 1, \dots, d, \quad s = 1, \dots, t \,,$$

$$\widehat{E}_t \triangleq \inf_{k \in \mathbb{Z}} \left\{ 2^k : 2^k \geqslant \max_{1 \leqslant s \leqslant t} \max_{\substack{1 \leqslant j,k \leqslant d \\ \gamma,\mu \in \{+,-\}}} \left| z^\gamma_{j,s} - z^\mu_{k,s} \right| \right\} \,,$$

$$V_t \triangleq \sum_{s=1}^{t} \sum_{\substack{1 \leqslant j \leqslant d \\ \gamma \in \{+,-\}}} p^\gamma_{j,s} \left(z^\gamma_{j,s} - \sum_{\substack{1 \leqslant k \leqslant d \\ \mu \in \{+,-\}}} p^\mu_{k,s} z^\mu_{k,s} \right)^2 \,.$$

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

Proposition 1 (The adaptive EG$^{\pm}$ algorithm for general convex and differentiable loss functions)**.** *Let $U > 0$. Then, the adaptive EG$^{\pm}$ algorithm on $B_1(U)$ defined in Figure 2 satisfies, for all $T \geqslant 1$ and all sequences of convex and differentiable*[8] *loss functions $\ell_1, \dots, \ell_T : \mathbb{R}^d \to \mathbb{R}$,*

$$\sum_{t=1}^{T} \ell_t(\widehat{\boldsymbol{u}}_t) - \min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \leqslant 4U \sqrt{\left(\sum_{t=1}^{T} \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty^2 \right) \ln(2d)} + U \left(8 \ln(2d) + 12\right) \max_{1 \leqslant t \leqslant T} \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty \ .$$

In particular, the regret is bounded by $4U \big(\max_{1 \leqslant t \leqslant T} \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty\big) \big(\sqrt{T \ln(2d)} + 2 \ln(2d) + 3\big)$.

Proof: The proof follows straightforwardly from a linearization argument and from a regret bound of [11] applied to appropriately chosen loss vectors. Indeed, first note that by convexity and differentiability of $\ell_t : \mathbb{R}^d \to \mathbb{R}$ for all $t = 1, \dots, T$, we get that

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(\widehat{\boldsymbol{u}}_t) - \min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) &= \max_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \big(\ell_t(\widehat{\boldsymbol{u}}_t) - \ell_t(\boldsymbol{u})\big) \leqslant \max_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \nabla \ell_t(\widehat{\boldsymbol{u}}_t) \cdot (\widehat{\boldsymbol{u}}_t - \boldsymbol{u}) \\
&= \max_{\substack{1 \leqslant j \leqslant d \\ \gamma \in \{+,-\}}} \sum_{t=1}^{T} \nabla \ell_t(\widehat{\boldsymbol{u}}_t) \cdot (\widehat{\boldsymbol{u}}_t - \gamma U \boldsymbol{e}_j) \qquad (9) \\
&= \sum_{t=1}^{T} \sum_{\substack{1 \leqslant j \leqslant d \\ \gamma \in \{+,-\}}} p_{j,t}^{\gamma} \, \gamma U \nabla_j \ell_t(\widehat{\boldsymbol{u}}_t) - \min_{\substack{1 \leqslant j \leqslant d \\ \gamma \in \{+,-\}}} \sum_{t=1}^{T} \gamma U \nabla_j \ell_t(\widehat{\boldsymbol{u}}_t) \ , \qquad (10)
\end{aligned}$$

where (9) follows by linearity of $\boldsymbol{u} \mapsto \sum_{t=1}^{T} \nabla \ell_t(\widehat{\boldsymbol{u}}_t) \cdot (\widehat{\boldsymbol{u}}_t - \boldsymbol{u})$ on the polytope $B_1(U)$, and where (10) follows from the particular choice of $\widehat{\boldsymbol{u}}_t$ in Figure 2.

To conclude the proof, note that our choices of the weight vectors $\boldsymbol{p}_t \in \mathcal{X}_{2d}$ in Figure 2 and of the time-varying parameter η_t in (8) correspond to the exponentially weighted average forecaster of [11, Section 4.2] when it is applied to the loss vectors $\big(U \nabla_j \ell_t(\widehat{\boldsymbol{u}}_t), -U \nabla_j \ell_t(\widehat{\boldsymbol{u}}_t)\big)_{1 \leqslant j \leqslant d} \in \mathbb{R}^{2d}$, $t = 1, \dots, T$. Since at time t the coordinates of the last loss vector lie in an interval of length $E_t \leqslant 2U \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty$, we get from [11, Corollary 1] that

$$\sum_{t=1}^{T} \sum_{\substack{1 \leqslant j \leqslant d \\ \gamma \in \{\pm 1\}}} p_{j,t}^{\gamma} \, \gamma U \nabla_j \ell_t(\widehat{\boldsymbol{u}}_t) - \min_{\substack{1 \leqslant j \leqslant d \\ \gamma \in \{\pm 1\}}} \sum_{t=1}^{T} \gamma U \nabla_j \ell_t(\widehat{\boldsymbol{u}}_t) \leqslant 4U \sqrt{\left(\sum_{t=1}^{T} \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty^2 \right) \ln(2d)} + U \left(8 \ln(2d) + 12\right) \max_{1 \leqslant t \leqslant T} \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty \ .$$

Substituting the last upper bound in (10) concludes the proof. □

3.2. Application to the square loss

In the particular case of the square loss $\ell_t(\boldsymbol{u}) = (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$, the gradients are given by $\nabla \ell_t(\boldsymbol{u}) = -2(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t) \, \boldsymbol{x}_t$ for all $\boldsymbol{u} \in \mathbb{R}^d$. Applying Proposition 1, we get the following regret bound for the adaptive EG$^{\pm}$ algorithm.

[8] Gradients can be replaced with subgradients if the loss functions $\ell_t : \mathbb{R}^d \to \mathbb{R}$ are convex but not differentiable.

Corollary 2 (The adaptive EG$^{\pm}$ algorithm under the square loss)**.**
Let $U > 0$. Consider the online linear regression setting defined in the introduction. Then, the adaptive EG$^{\pm}$ algorithm (see Figure 2) tuned with U and applied to the loss functions $\ell_t : \boldsymbol{u} \mapsto (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$ satisfies, for all individual sequences $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_T, y_T) \in \mathbb{R}^d \times \mathbb{R}$,

$$\begin{aligned}
&\sum_{t=1}^{T} (y_t - \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t)^2 - \min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \\
&\qquad \leqslant 8UX \sqrt{\left(\min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \right) \ln(2d)} + \big(137 \ln(2d) + 24\big) \big(UXY + U^2 X^2\big) \\
&\qquad \leqslant 8UXY \sqrt{T \ln(2d)} + \big(137 \ln(2d) + 24\big) \big(UXY + U^2 X^2\big) \ ,
\end{aligned}$$

where the quantities $X \triangleq \max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty$ and $Y \triangleq \max_{1 \leqslant t \leqslant T} |y_t|$ are unknown to the forecaster.

Using the terminology of [17, 11], the first bound of Corollary 2 is an *improvement for small losses*: it yields a small regret when the optimal cumulative loss $\min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$ is small. As for the second regret bound, it indicates that the adaptive EG$^{\pm}$ algorithm achieves approximately the regret bound of Theorem 1 in the regime $\kappa \leqslant 1$, i.e., $d \geqslant \sqrt{T} UX/(2Y)$. In this regime, our algorithm thus has a manageable computational complexity (linear in d at each time t) and it is adaptive in X, Y, and T.

In particular, the above regret bound is similar[9] to that of the original EG$^{\pm}$ algorithm [9, Theorem 5.11], but it is obtained without prior knowledge of X, Y, and T. Note also that this bound is similar to that of the self-confident p-norm algorithm of [10] with $p = 2 \ln d$ (see Section 1.2). The fact that we were able to get similar adaptivity and efficiency properties via exponential weighting corroborates the similarity that was already observed in a non-adaptive context between the original EG$^{\pm}$ algorithm and the p-norm algorithm (in the limit $p \to +\infty$ with an appropriate initial weight vector, or for p of the order of $\ln d$ with a zero initial weight vector, cf. [18]).

Proof (of Corollary 2): We apply Proposition 1 with the square loss $\ell_t(\boldsymbol{u}) = (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$. It yields

$$\begin{aligned}
&\sum_{t=1}^{T} \ell_t(\widehat{\boldsymbol{u}}_t) - \min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \\
&\qquad \leqslant 4U \sqrt{\left(\sum_{t=1}^{T} \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty^2 \right) \ln(2d)} + U \big(8 \ln(2d) + 12\big) \max_{1 \leqslant t \leqslant T} \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty \ . \qquad (11)
\end{aligned}$$

Using the equality $\nabla \ell_t(\boldsymbol{u}) = -2(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)\, \boldsymbol{x}_t$ for all $\boldsymbol{u} \in \mathbb{R}^d$, we get that, on the one hand, by the upper bound $\|\boldsymbol{x}_t\|_\infty \leqslant X$,

$$\|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty^2 \leqslant 4X^2 \ell_t(\widehat{\boldsymbol{u}}_t) \ , \qquad (12)$$

and, on the other hand, $\max_{1 \leqslant t \leqslant T} \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty \leqslant 2(Y + UX)X$ (indeed, by Hölder's inequality, $\big|\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t\big| \leqslant \|\widehat{\boldsymbol{u}}_t\|_1 \, \|\boldsymbol{x}_t\|_\infty \leqslant UX$). Substituting the last two inequalities in (11), setting $\widehat{L}_T \triangleq \sum_{t=1}^{T} \ell_t(\widehat{\boldsymbol{u}}_t)$ as well as $L_T^* \triangleq \min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \ell_t(\boldsymbol{u})$, we get that

$$\widehat{L}_T \leqslant L_T^* + 8UX \sqrt{\widehat{L}_T \ln(2d)} + \underbrace{\big(16 \ln(2d) + 24\big) \big(UXY + U^2 X^2\big)}_{\triangleq C} \ .$$

[9] By Theorem 5.11 of [9], the original EG$^{\pm}$ algorithm satisfies the regret bound $2UX\sqrt{2B \ln(2d)} + 2U^2 X^2 \ln(2d)$, where B is an upper bound on $\min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$ (in particular, $B \leqslant TY^2$). Note that our main regret term is larger by a multiplicative factor of $2\sqrt{2}$. However, contrary to [9], our algorithm does not require the prior knowledge of X and B — or, alternatively, X, Y, and T.

Solving for $\widehat{L}_T$ via Lemma 4 in Appendix B, we get that

$$\begin{aligned}\widehat{L}_T &\leqslant L_T^* + C + \left(8UX\sqrt{\ln(2d)}\right)\sqrt{L_T^* + C} + \left(8UX\sqrt{\ln(2d)}\right)^2 \\ &\leqslant L_T^* + 8UX\sqrt{L_T^*\ln(2d)} + 8UX\sqrt{C\ln(2d)} + 64U^2X^2\ln(2d) + C\ .\end{aligned}$$

Using that

$$\begin{aligned}UX\sqrt{C\ln(2d)} &= UX\ln(2d)\sqrt{\big(16 + 24/\ln(2d)\big)\big(UXY + U^2X^2\big)} \\ &\leqslant \sqrt{U^2X^2 + UXY}\ \ln(2d)\,\sqrt{\big(16 + 24/\ln(2)\big)\big(UXY + U^2X^2\big)} \\ &= \sqrt{16 + 24/\ln(2)}\,\big(UXY + U^2X^2\big)\ln(2d)\end{aligned}$$

and performing some simple upper bounds concludes the proof of the first regret bound. The second one follows immediately by noting that $\min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \leqslant \sum_{t=1}^T y_t^2 \leqslant TY^2$ (since $\mathbf{0} \in B_1(U)$). □

3.3. A refinement via Lipschitzification of the loss function

In Corollary 2 we used the adaptive EG$^\pm$ algorithm in conjunction with the square loss functions $\ell_t : \boldsymbol{u} \mapsto (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2$. In this section we use yet another instance of the adaptive EG$^\pm$ algorithm applied to a modification $\widetilde{\ell}_t : \mathbb{R}^d \to \mathbb{R}$ of the square loss (or the α-loss, see below) which is Lipschitz continuous with respect to $\|\cdot\|_1$. This leads to slightly refined regret bounds; see Theorem 3 below and Corollaries 3 and 4 thereafter.

We first present the Lipschtizification technique; its use with the adaptive EG$^\pm$ algorithm is to be addressed in a few paragraphs. Since our analysis is generic enough to handle both the square loss and other loss functions with higher curvature, we consider below a slightly more general setting than online linear regression *stricto sensu*. Namely, we fix a real number $\alpha \geqslant 2$ and assume that the predictions $\widehat{y}_t$ of the forecaster and the base linear predictions $\boldsymbol{u}\cdot\boldsymbol{x}_t$ are scored with the α-loss, i.e., with the loss functions $x \mapsto |y_t - x|^\alpha$ for all $t \geqslant 1$. The particular case of the square loss ($\alpha = 2$) is considered in Corollary 3 below, while loss functions with higher curvature ($\alpha > 2$) are addressed in Corollary 4.

The Lipschitzification proceeds as follows. At each time $t \geqslant 1$, we set

$$B_t \triangleq \left(2^{\lceil \log_2(\max_{1\leqslant s\leqslant t-1} |y_s|^\alpha)\rceil}\right)^{1/\alpha}\ ,$$

where $\lceil x\rceil \triangleq \min\{k \in \mathbb{Z} : k \geqslant x\}$ for all $x \in \mathbb{R}$. Note that $\max_{1\leqslant s\leqslant t-1}|y_s| \leqslant B_t \leqslant 2^{1/\alpha}\max_{1\leqslant s\leqslant t-1}|y_s|$. The modified (or *Lipschitzified*) loss function $\widetilde{\ell}_t : \mathbb{R}^d \to \mathbb{R}$ is constructed as follows:

- if $|y_t| > B_t$, then
$$\widetilde{\ell}_t(\boldsymbol{u}) \triangleq 0 \quad \text{for all } \boldsymbol{u} \in \mathbb{R}^d\ ;$$
- if $|y_t| \leqslant B_t$, then $\widetilde{\ell}_t$ is the convex function that coincides with the loss function $\boldsymbol{u} \mapsto |y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t|^\alpha$ when $\big|\boldsymbol{u}\cdot\boldsymbol{x}_t\big| \leqslant B_t$ and is linear elsewhere. An example of such function is shown in Figure 3 in the case where $\alpha = 2$. It can be formally defined as

$$\widetilde{\ell}_t(\boldsymbol{u}) \triangleq \begin{cases} \big|y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big|^\alpha & \text{if} \quad \big|\boldsymbol{u}\cdot\boldsymbol{x}_t\big| \leqslant B_t, \\ \big|y_t - B_t\big|^\alpha + \alpha\big|y_t - B_t\big|^{\alpha-1}(\boldsymbol{u}\cdot\boldsymbol{x}_t - B_t) & \text{if} \quad \boldsymbol{u}\cdot\boldsymbol{x}_t > B_t, \\ \big|y_t + B_t\big|^\alpha - \alpha\big|y_t + B_t\big|^{\alpha-1}(\boldsymbol{u}\cdot\boldsymbol{x}_t + B_t) & \text{if} \quad \boldsymbol{u}\cdot\boldsymbol{x}_t < -B_t. \end{cases}$$

Observe that in both cases $|y_t| > B_t$ and $|y_t| \leqslant B_t$, the function $\widetilde{\ell}_t$ is continuously differentiable. By construction it is also Lipschitz continuous with respect to $\|\cdot\|_1$ with an easy-to-control Lipschitz constant (see Appendix A.2). Another key property that we can glean from Figure 3 is that, when $|y_t| \leqslant B_t$, the

modified loss function $\widetilde{\ell}_t : \mathbb{R}^d \to \mathbb{R}$ lies in between the α-loss function $\boldsymbol{u} \mapsto |y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t|^\alpha$ and its clipped version:

$$\forall \boldsymbol{u} \in \mathbb{R}^d, \quad \left|y_t - [\boldsymbol{u} \cdot \boldsymbol{x}_t]_{B_t}\right|^\alpha \leqslant \widetilde{\ell}_t(\boldsymbol{u}) \leqslant \left|y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t\right|^\alpha , \tag{13}$$

where the clipping operator $[\cdot]_B$ is defined by $[x]_B \triangleq \min\{B, \max\{-B, x\}\}$ for all $x \in \mathbb{R}$ and all $B > 0$.

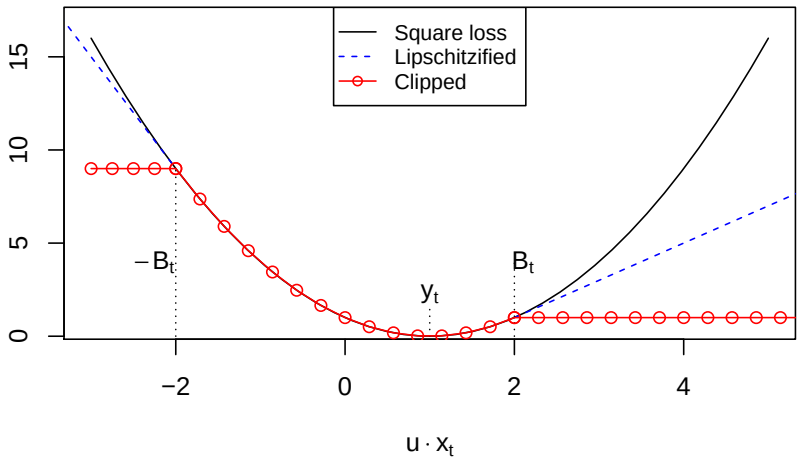


Figure 3: Example with the square loss ($\alpha = 2$) when $|y_t| \leqslant B_t$. The square loss $(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$, its clipped version $\left(y_t - [\boldsymbol{u} \cdot \boldsymbol{x}_t]_{B_t}\right)^2$ and its Lipschitzified version $\widetilde{\ell}_t(\boldsymbol{u})$ are plotted as a function of $\boldsymbol{u} \cdot \boldsymbol{x}_t$.

Next we illustrate the Lipschitzification technique introduced above: we apply the adaptive EG$^\pm$ algorithm to the Lipschitzified loss functions $\widetilde{\ell}_t$. The resulting algorithm is called the *Lipschitzifying Exponentiated Gradient* (LEG) algorithm and is formally defined in Figure 4. Recall that $(\mathbf{e}_j)_{1 \leqslant j \leqslant d}$ denotes the canonical basis of $\mathbb{R}^d$ and that ∇_j denotes the j-th component of the gradient.

We point out that this technique is not specific to the pair of dual norms $(\|\cdot\|_1, \|\cdot\|_\infty)$ and to the EG$^\pm$ algorithm; it could be used with other pairs $(\|\cdot\|_q, \|\cdot\|_p)$ (with $1/p + 1/q = 1$) and other gradient-based algorithms, such as the p-norm algorithm [18, 10] and its regularized variants (SMIDAS and COMID) [19, 20].

The next theorem bounds the cumulative α-loss of the LEG algorithm. The proof is postponed to Appendix A.2. It follows from the bound on the adaptive EG$^\pm$ algorithm for general convex and differentiable loss functions that we derived in Proposition 1 (Section 3.1). See Corollaries 3 and 4 below for regret bounds in the particular cases of the square loss ($\alpha = 2$) or of losses with higher curvature ($\alpha > 2$).

Theorem 3. *Assume that the predictions are scored with the α-loss $x \mapsto |y_t - x|^\alpha$, where $\alpha \geqslant 2$ is a real number. Let $U > 0$. Then, the LEG algorithm defined in Figure 4 and tuned with U satisfies, for all $T \geqslant 1$ and all individual sequences $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T) \in \mathbb{R}^d \times \mathbb{R}$,*

$$\sum_{t=1}^T |y_t - \widehat{y}_t|^\alpha \leqslant \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T \widetilde{\ell}_t(\boldsymbol{u}) + a_\alpha U X Y^{\alpha/2-1} \sqrt{\left(\inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T \widetilde{\ell}_t(\boldsymbol{u})\right) \ln(2d)}$$
$$+ \left(a'_\alpha \ln(2d) + 12 b_\alpha\right) U X Y^{\alpha-1} + a''_\alpha \ln(2d)\, U^2 X^2 Y^{\alpha-2} + a'''_\alpha\, Y^\alpha ,$$

where the Lipschitzified loss functions $\widetilde{\ell}_t$ are defined above, where the quantities $X \triangleq \max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty$ and $Y \triangleq \max_{1 \leqslant t \leqslant T} |y_t|$ are unknown to the forecaster, and where, setting $a_\alpha \triangleq 4\alpha \left(1 + 2^{1/\alpha}\right)^{\alpha/2-1}$ and

Parameter: radius $U > 0$.

Initialization: $B_1 \triangleq 0$, $\boldsymbol{p}_1 = (p^+_{1,1}, p^-_{1,1}, \dots, p^+_{d,1}, p^-_{d,1}) \triangleq \big(1/(2d), \dots, 1/(2d)\big) \in \mathbb{R}^{2d}$.

At each time round $t \geqslant 1$,

1. Compute the linear combination $\widehat{\boldsymbol{u}}_t \triangleq U \sum_{j=1}^{d} \big(p^+_{j,t} - p^-_{j,t}\big)\, \boldsymbol{e}_j \in B_1(U)$;
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 function $\widetilde{\ell}_t : \mathbb{R}^d \to \mathbb{R}$ as above;
4. Update the parameter η_{t+1} according to (8);
5. Update the weight vector $\boldsymbol{p}_{t+1} = (p^+_{1,t+1}, p^-_{1,t+1}, \dots, p^+_{d,t+1}, p^-_{d,t+1}) \in \mathcal{X}_{2d}$ defined for all $j = 1, \dots, d$ and $\gamma \in \{+, -\}$ by[a]
$$p^\gamma_{j,t+1} \triangleq \frac{\exp\left(-\eta_{t+1} \sum_{s=1}^{t} \gamma U \nabla_j \widetilde{\ell}_s(\widehat{\boldsymbol{u}}_s)\right)}{\sum_{\substack{1 \leqslant k \leqslant d \\ \mu \in \{+,-\}}} \exp\left(-\eta_{t+1} \sum_{s=1}^{t} \mu U \nabla_k \widetilde{\ell}_s(\widehat{\boldsymbol{u}}_s)\right)} \,.$$
6. Update the threshold $\quad B_{t+1} \triangleq \big(2^{\lceil \log_2(\max_{1 \leqslant s \leqslant t} |y_s|^\alpha) \rceil}\big)^{1/\alpha}$.

[a]For all $\gamma \in \{+, -\}$, by a slight abuse of notation, γU denotes U or $-U$ if $\gamma = +$ or $\gamma = -$ respectively.

Figure 4: The Lipschitzifying Exponentiated Gradient (LEG) algorithm.

$b_\alpha \triangleq \alpha \big(1 + 2^{1/\alpha}\big)^{\alpha - 1}$, the constants $a'_\alpha, a''_\alpha, a'''_\alpha > 0$ are defined by

$$\begin{cases} a'_\alpha \triangleq a_\alpha \left(\sqrt{b_\alpha \big(4 + 6/\ln 2\big)} + 2\big(1 + 2^{-1/\alpha}\big)^{\alpha/2} / \sqrt{\ln 2}\right) + 8 b_\alpha \\ a''_\alpha \triangleq a_\alpha \left(\sqrt{b_\alpha \big(4 + 6/\ln 2\big)} + a_\alpha\right) \\ a'''_\alpha \triangleq 4\big(1 + 2^{-1/\alpha}\big)^\alpha \,. \end{cases}$$

Corollary 3 (Application to the square loss)**.** *Consider the online linear regression setting under the square loss (i.e., $\alpha = 2$). Let $U > 0$. Then, the LEG algorithm defined in Figure 4 and tuned with U satisfies, for all $T \geqslant 1$ and all individual sequences $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_T, y_T) \in \mathbb{R}^d \times \mathbb{R}$,*

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \widetilde{\ell}_t(\boldsymbol{u}) + 8UX \sqrt{\left(\inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \widetilde{\ell}_t(\boldsymbol{u})\right) \ln(2d)} + \big(134 \ln(2d) + 58\big)\big(UXY + U^2X^2\big) + 12Y^2 \,,$$

where the Lipschitzified loss functions $\widetilde{\ell}_t$ are defined above and where the quantities $X \triangleq \max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty$ and $Y \triangleq \max_{1 \leqslant t \leqslant T} |y_t|$ are unknown to the forecaster.

Note that, in the case of the square loss, the first two terms of the bound of Corollary 3 slightly improve on those obtained without Lipschitzification (cf. Corollary 2) since we always have

$$\inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \widetilde{\ell}_t(\boldsymbol{u}) \leqslant \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \,,$$

where we used the key property $\widetilde{\ell}_t(\boldsymbol{u}) \leqslant (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2$ that holds for all $\boldsymbol{u}\in\mathbb{R}^d$ and all $t=1,\dots,T$ (by (13) if $|y_t|\leqslant B_t$, obvious otherwise). In particular, the LEG algorithm is adaptive in X, Y, and T; it achieves approximately — and efficiently — the regret bound of Theorem 1 in the regime $\kappa\leqslant 1$, i.e., $d\geqslant \sqrt{T}UX/(2Y)$.

In the case of α-losses with a higher curvature than that of the square loss ($\alpha>2$), the improvement is more substantial as indicated after the following corollary.

Corollary 4 (Application to α-losses with $\alpha>2$)**.** *Assume that the predictions are scored with the α-loss $x\mapsto |y_t-x|^\alpha$, where $\alpha>2$. Then, the regret of the LEG algorithm on $B_1(U)$ is at most of the order of*

$$UXY^{\alpha-1}\sqrt{T\ln(2d)} + \Big(UXY^{\alpha-1} + U^2X^2Y^{\alpha-2}\Big)\ln(2d) + Y^\alpha\ ,$$

where $X\triangleq \max_{1\leqslant t\leqslant T}\|\boldsymbol{x}_t\|_\infty$ and $Y\triangleq\max_{1\leqslant t\leqslant T}|y_t|$ are unknown to the forecaster. The above regret bound improves on the bound we would have obtained via a similar analysis for the adaptive $EG^\pm$ algorithm applied to the original losses $\ell_t(\boldsymbol{u})=|y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t|^\alpha$ (without Lipschitzification), namely, a bound of the order of

$$UX(Y+UX)^{\alpha/2-1}Y^{\alpha/2}\sqrt{T\ln(2d)} + \Big(UX(Y+UX)^{\alpha-1} + U^2X^2(Y+UX)^{\alpha-2}\Big)\ln(2d)\ .$$

The main difference between the two regret bounds above lies in the dependence in U: our main regret term scales as $UXY^{\alpha-1}$ while the one obtained without Lipschitzification scales as $UX(Y+UX)^{\alpha/2-1}Y^{\alpha/2}$. The first term grows linearly in U while the second one grows as $U^{\alpha/2}$, hence a clear improvement for $\alpha>2$. The last property stems from the fact that, thanks to Lipschitzification, the gradients $\Big\|\nabla\widetilde{\ell}_t\Big\|_\infty$ are bounded as $U\to+\infty$ (cf. (A.29) in Appendix A.2).

Remark 1 (Another benefit of Lipschitzification)**.**
Another benefit of Lipschitzification is that all online convex optimization regret bounds expressed in terms of the maximal dual norm of the gradients — i.e., $\max_{1\leqslant t\leqslant T}\|\nabla\widetilde{\ell}_t\|_\infty$ in our case — can be used fruitfully with the Lipschitzified loss functions $\widetilde{\ell}_t$. For instance, in the case of the square loss, using the very last bound of Proposition 1, we get that

$$\sum_{t=1}^T (y_t-\widehat{y}_t)^2 - \inf_{\|\boldsymbol{u}\|_1\leqslant U}\sum_{t=1}^T (y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \leqslant c_1UXY\left(\sqrt{T\ln(2d)}+8\ln(2d)\right)+c_2Y^2\ ,$$

where $c_1\triangleq 8\big(\sqrt{2}+1\big)$ and $c_2\triangleq 4\big(1+1/\sqrt{2}\big)^2$. The bound is no longer an improvement for small losses (as compared to Corollary 2), but it does not require to solve any quadratic inequality. The corresponding simple proof is postponed to the end of Appendix A.2.

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, Y, and T are known: we explain in Section 5 how to simultaneously adapt to all parameters. Note that from now on, we consider again the main framework of this paper, i.e., online linear regression under the square loss (cf. Section 1.1).

We define

$$R\triangleq \lceil \log_2(2T/c)\rceil_+ \quad\text{and}\quad U_r\triangleq \frac{Y}{X}\frac{2^r}{\sqrt{T\ln(2d)}},\quad \text{for } r=0,\dots,R\ , \tag{14}$$

Parameters: $X, Y, \eta > 0$, $T \geqslant 1$, and $c > 0$ (a constant).
Initialization: $R = \lceil \log_2(2T/c) \rceil_+$, $\boldsymbol{w}_1 = \left(\frac{1}{R+1}, \cdots, \frac{1}{R+1}\right) \in \mathbb{R}^{R+1}$.
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 $\widehat{y}_t = \sum_{r=0}^{R} \frac{w_t^{(r)}}{\sum_{r'=0}^{R} w_t^{(r')}} \left[\widehat{y}_t^{(r)}\right]_Y$.
3. Update $w_{t+1}^{(r)} = w_t^{(r)} \exp\left(-\eta\left(y_t - \left[\widehat{y}_t^{(r)}\right]_Y\right)^2\right)$ for $r = 0, \ldots, R$.

Figure 5: The Scaling algorithm.

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

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

The Scaling algorithm of Figure 5 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*

$$cUXY\sqrt{T\ln(2d)} + c'Y^2 \quad \text{for } T \geqslant T_0 \ , \tag{15}$$

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

$$\begin{aligned}\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 + 2c\left\|\boldsymbol{u}\right\|_1 XY\sqrt{T\ln(2d)}\right\} \\ + 8Y^2\ln\left(\lceil\log_2(2T/c)\rceil_+ + 1\right) + (c+c')Y^2. \end{aligned} \tag{16}$$

In particular, for every $U > 0$,

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

Remark 2. *By Remark 1 the LEG algorithm satisfies assumption (15) with $T_0 = \ln(2d)$, $c \triangleq 9c_1 = 72\left(\sqrt{2}+1\right)$, 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 $\mathcal{A}(U_r) = \big(\widehat{y}_t^{(r)}\big)_{t\geqslant 1}$, $r = 0, \dots, R$, we have, by Lemma 6 in Appendix B,

$$\begin{aligned}\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 &\leqslant \min_{r=0,\dots,R}\sum_{t=1}^{T}\left(\widehat{y}_t^{(r)} - \widehat{y}_t\right)^2 + 8Y^2\ln(R+1)\\ &\leqslant \min_{r=0,\dots,R}\left\{\inf_{\boldsymbol{u}\in B_1(U_r)}\left\{\sum_{t=1}^{T}(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2\right\} + cU_rXY\sqrt{T\ln(2d)}\right\} + z\ , \end{aligned} \tag{17}$$

where the last inequality follows by assumption (15), 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 $r^* \triangleq \min\big\{r = 0,\dots,R : U_r \geqslant \left\|\boldsymbol{u}_T^*\right\|_1\big\}$. Note that $r^* \geqslant 1$ since $\left\|\boldsymbol{u}_T^*\right\|_1 > U_0$. By (17) 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 (16) 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 (17) we have

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \left\{\sum_{t=1}^{T}(y_t - \boldsymbol{u}_T^*\cdot\boldsymbol{x}_t)^2 + cU_0XY\sqrt{T\ln(2d)}\right\} + z\ , \tag{18}$$

which yields (16) by the equality $cU_0XY\sqrt{T\ln(2d)} = cY^2$ (by definition of U_0), by adding the nonnegative quantity $2c\left\|\boldsymbol{u}_T^*\right\|_1 XY\sqrt{T\ln(2d)}$, 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 (16) by definition of $\boldsymbol{u}_T^*$.

This concludes the proof of the first claim (16). The second claim follows by bounding $\left\|\boldsymbol{u}\right\|_1 \leqslant U$. □

5. Extension to a fully adaptive algorithm

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 already used for 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 $\{U_0, \ldots, U_R\}$ by another exponential sequence $\{U'_0, \ldots, U'_{R'}\}$:

$$U'_r \triangleq \frac{1}{T^k} \frac{2^r}{\sqrt{T \ln(2d)}} \ , \quad r = 0, \ldots, R' \ , \tag{19}$$

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 $T \geqslant T_0 \triangleq \max\{(X/Y)^{1/k}, (Y/X)^{1/k}\}$, we have (cf. (14) and (19)),

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

Therefore, the analysis of Theorem 4 applied to the grid $\{U'_0, \ldots, U_{R'}\}$ yields[10] a regret bound of the order of $UXY\sqrt{T \ln d} + Y^2 \ln(R'+1)$. On the other hand, clipping the predictions to $[-Y, Y]$ ensures the crude regret bound $4Y^2 T_0$ for small $T < T_0$. 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 $\{U'_{-R'(t)}(t), \ldots, U'_{R'(t)}(t)\}$ where $R'(t)$ grows at the rate of $k \ln(t)$. This gives[11] 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 $k > 1$).

Acknowledgments

The authors would like to thank Gilles Stoltz for his valuable comments and suggestions, as well as two anonymous reviewers for their insightful feedback. 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.

An extended abstract of the present paper appeared in the *Proceedings of the 22nd International Conference on Algorithmic Learning Theory (ALT'11)*.

Appendix A. Proofs

Appendix A.1. 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 [2] for convex aggregation.

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

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

We first need the following notations. Let $T \in \mathbb{N}^*$. Let (S, μ) be a probability space for which we can find an orthonormal family[12] $(\varphi_j)_{1 \leqslant j \leqslant d}$ with d elements 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

$$Y_t = \gamma \varphi_{\boldsymbol{u}}(X_t) + \sigma \varepsilon_t \in \mathbb{R} \;, \tag{A.1}$$

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 [2, 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 [2])**.**
Let $d, T \in \mathbb{N}^*$ 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 (A.1)*. 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[13] $\widehat{f}_T : S \to \mathbb{R}$, where the supremum is taken over all nonnegative vectors with total mass at most 1*, and where $\|f\|_\mu^2 \triangleq \int_S f(x)^2 \mu(\mathrm{d}x)$ 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 $\kappa \triangleq \sqrt{T} U X / (2dY)$). 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;
- step 2: application of Lemma 1;
- step 3: concentration argument.

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

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

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

[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.

[13] As usual, an estimator is a measurable function of the sample $(X_t, Y_t)_{1 \leqslant t \leqslant T}$, but the dependency on the sample is omitted.

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

$$\inf_{(\widetilde{f}_t)_t} \sup_{\substack{\|\boldsymbol{x}_t\|_\infty \leqslant X \\ |y_t| \leqslant Y}} \left\{ \sum_{t=1}^T \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \right\}$$
$$= \inf_{\substack{(\widetilde{f}_t)_t \\ |\widetilde{f}_t| \leqslant Y}} \sup_{\substack{\|\boldsymbol{x}_t\|_\infty \leqslant X \\ |y_t| \leqslant Y}} \left\{ \sum_{t=1}^T \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \right\}, \tag{A.3}$$

where the first infimum is taken over all online forecasters[14] $(\widetilde{f}_t)_t$, 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 conversion to bound from below the right-hand side of (A.3) by T times the lower bound of Lemma 1, which we apply to the particular case where $S = [-\pi,\pi]$, where $\mu(\mathrm{d}x) = \mathrm{d}x/(2\pi)$, and where $\varphi_j(x) = \sqrt{2}\sin(jx)$ for all $1 \leqslant j \leqslant d$ and $x \in [-\pi,\pi]$. Let

$$\gamma \triangleq c_8 X \quad \text{and} \quad \sigma \triangleq \frac{c_9 Y}{\sqrt{\ln T}}\,, \tag{A.4}$$

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

$$\widehat{f}_T\Big(X'; (X_t,Y_t)_{1\leqslant t\leqslant T}\Big) \triangleq \frac{1}{T}\sum_{t=1}^T \widetilde{f}_t\Big(\gamma\boldsymbol{\varphi}(X'); (\gamma\boldsymbol{\varphi}(X_s), Y_s)_{1\leqslant s\leqslant t-1}\Big)\,, \tag{A.5}$$

where $\boldsymbol{\varphi} \triangleq (\varphi_1,\dots,\varphi_d)$, and where we explicitely wrote all the dependencies[14] of the $\widetilde{f}_t$, $t = 1,\dots,T$.
Take $\boldsymbol{u}^* \in \mathbb{R}^d_+$ achieving the supremum[15] in Lemma 1 for the estimator $\widehat{f}_T$. Note that $\|\boldsymbol{u}^*\|_1 \leqslant 1$. Besides, consider the i.i.d. 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

$$\boldsymbol{x}_t \triangleq \big(\gamma\varphi_1(X_t), \dots, \gamma\varphi_d(X_t)\big) \quad \text{and} \quad y_t \triangleq \gamma\varphi_{\boldsymbol{u}^*}(X_t) + \sigma\varepsilon_t\,, \tag{A.6}$$

where $\varphi_{\boldsymbol{u}^*} \triangleq \sum_{j=1}^d u_j^*\varphi_j$ (so that $y_t = \boldsymbol{u}^*\cdot\boldsymbol{x}_t + \sigma\varepsilon_t$ for all t), 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 are thus taken with respect to the probability distribution $\mathbb{P}_{\boldsymbol{u}^*}^{\gamma,\sigma}$.

By standard manipulations (e.g., using the tower rule and Jensen's inequality), we get the following lower bound. A detailed proof can be found after the proof of the present theorem (page 24).

Lemma 2 (Reduction to the batch setting)**.**
With $(\widetilde{f}_t)_{1\leqslant t\leqslant T}$, $\widehat{f}_T$, and $\boldsymbol{u}^*$ defined above, we have

$$\mathbb{E}\left[\sum_{t=1}^T \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1\leqslant 1}\sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2\right] \geqslant T\,\mathbb{E}\left\|\widehat{f}_T - \gamma\varphi_{\boldsymbol{u}^*}\right\|_\mu^2\,.$$

[14]Recall that 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 dependency in $(\boldsymbol{x}_s,y_s)_{1\leqslant s\leqslant t-1}$, and only write $\widetilde{f}_t(\boldsymbol{x}_t)$.

[15]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.

Step 2: application of Lemma 1.
Next we use Lemma 1 to prove that, for some absolute constants $c_9, c_{11} > 0$,

$$T\,\mathbb{E}\left\|\widehat{f}_T - \gamma\varphi_{\boldsymbol{u}^*}\right\|_\mu^2 \geqslant \frac{c_{11}c_9^2}{\ln\bigl(2+16d^2\bigr)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\ . \tag{A.7}$$

By Lemma 1 and by definition of $\boldsymbol{u}^*$, we have

$$\begin{aligned}\mathbb{E}\left\|\widehat{f}_T - \gamma\varphi_{\boldsymbol{u}^*}\right\|_\mu^2 &\geqslant \begin{cases} c_4\frac{d\sigma^2}{T} & \text{if} \quad \frac{d}{\sqrt{T}} \leqslant c_5\frac{\gamma}{\sigma}\ , \\ c_6\gamma\sigma\sqrt{\frac{1}{T}\ln\left(1+\frac{d\sigma}{\sqrt{T}\gamma}\right)} & \text{if} \quad c_5\frac{\gamma}{\sigma} < \frac{d}{\sqrt{T}} \leqslant \frac{c_7\gamma d}{\sigma\sqrt{\ln(1+d)}}\ . \end{cases} \\ &\geqslant \begin{cases} \frac{c_4c_9^2}{T(\ln T)}dY^2 & \text{if} \quad \frac{d}{\sqrt{T}} \leqslant c_5\frac{\gamma}{\sigma}\ , \\ \frac{c_6c_8c_9}{\sqrt{\ln T}}UXY\sqrt{\frac{1}{T}\ln\left(1+\frac{c_9dY}{c_8\sqrt{T(\ln T)}UX}\right)} & \text{if} \quad c_5\frac{\gamma}{\sigma} < \frac{d}{\sqrt{T}} \leqslant \frac{c_7\gamma d}{\sigma\sqrt{\ln(1+d)}}\ , \end{cases}\end{aligned} \tag{A.8}$$

where the last inequality follows from (A.4) and from $U = 1$.

The above lower bound is only meaningful if the following condition holds true:

$$\frac{d}{\sqrt{T}} \leqslant \frac{c_7\gamma d}{\sigma\sqrt{\ln(1+d)}}\ . \tag{A.9}$$

But, by definition of $T \triangleq 1+\bigl\lceil(4d\kappa)^2\bigr\rceil$ and by the assumption $\sqrt{\ln(1+2d)}/\bigl(2d\sqrt{\ln 2}\bigr) \leqslant \kappa$, elementary manipulations show that (A.9) actually holds true whenever[16] $c_9 \leqslant c_7c_8c_{10}$, 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 $c_{10} > 0$).

Therefore, if $c_9 \leqslant c_7c_8c_{10}$, then (A.8) entails that

$$\mathbb{E}\left\|\widehat{f}_T - \gamma\varphi_{\boldsymbol{u}^*}\right\|_\mu^2 \geqslant \min\left\{\frac{c_4c_9^2}{T(\ln T)}dY^2,\ \frac{c_6c_8c_9}{\sqrt{\ln T}}UXY\sqrt{\frac{1}{T}\ln\left(1+\frac{c_9dY}{c_8\sqrt{T(\ln T)}UX}\right)}\right\}\ . \tag{A.10}$$

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

$$\begin{aligned}&\frac{c_6c_8c_9}{\sqrt{\ln T}}UXY\sqrt{\frac{1}{T}\ln\left(1+\frac{c_9dY}{c_8\sqrt{T(\ln T)}UX}\right)} \\ &\quad\geqslant \frac{c_6c_9^2}{2\ln T}UXY\sqrt{\frac{1}{T}\ln\left(1+\frac{2dY}{\sqrt{T}UX}\right)} = \frac{c_6c_9^2}{T(\ln T)}dY^2\kappa\sqrt{\ln(1+1/\kappa)}\ ,\end{aligned}$$

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

$$c_9 \leqslant c_7c_8c_{10} \quad\text{and}\quad c_9 \leqslant c_8 2\sqrt{\ln 2}\ . \tag{A.11}$$

[16] By definition of γ and σ, (A.9) is equivalent to $T\ln T \geqslant c_9^2/(c_7^2c_8^2)(Y/X)^2\ln(1+d)$. But by definition of X and by the assumption $\kappa \geqslant \sqrt{\ln(1+2d)}/(2d\sqrt{\ln 2})$, we have $Y/X \leqslant 1/c_{10}$. Therefore, (A.9) is implied by $T\ln T \geqslant c_9^2/(c_7^2c_8^2c_{10}^2)\ln(1+d)$, which in turn is implied by the condition $c_9 \leqslant c_7c_8c_{10}$ (by definition of T).

Therefore, by the above remarks, by the fact that $\ln T \triangleq \ln\bigl(1 + \lceil (4d\kappa)^2 \rceil\bigr) \leqslant \ln\bigl(2 + 16d^2\bigr)$ (since $\kappa \leqslant 1$ by assumption), and multiplying both sides of (A.10) by T, we get

$$T\,\mathbb{E}\left\|\widehat{f}_T - \gamma\varphi_{\boldsymbol{u}^*}\right\|_\mu^2 \geqslant \min\left\{\frac{c_4 c_9^2}{\ln\bigl(2+16d^2\bigr)} dY^2,\ \frac{c_6 c_9^2}{\ln\bigl(2+16d^2\bigr)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\right\}$$
$$\geqslant \frac{c_{11} c_9^2}{\ln\bigl(2+16d^2\bigr)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\ ,$$

where we set $c_{11} \triangleq \min\bigl\{c_4/\sqrt{\ln 2}, c_6\bigr\}$, and where we used the fact that $x \mapsto x\sqrt{\ln(1+1/x)}$ is nondecreasing on $\mathbb{R}_+^*$, so that its value at $x = \kappa \leqslant 1$ is smaller than $\sqrt{\ln 2}$. This concludes the proof of (A.7).

Combining Lemma 2 and (A.7), we get

$$\mathbb{E}\left[\sum_{t=1}^T \bigl(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\bigr)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant 1} \sum_{t=1}^T \bigl(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\bigr)^2\right] \geqslant \frac{c_{11} c_9^2}{\ln\bigl(2+16d^2\bigr)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\ . \tag{A.12}$$

Step 3: concentration argument.
At this stage it would be tempting to conclude by using (A.12) that since the expectation is lower bounded, then there is at least one individual sequence with the same lower bound. However, we have no boundedness guarantee about such individual sequence since the random observations y_t lie outside of $[-Y, Y]$ with positive probability. Next we prove that the probability of the event

$$\mathcal{A} \triangleq \bigcap_{t=1}^T \bigl\{|y_t| \leqslant Y\bigr\}$$

is actually close to 1, and that

$$\mathbb{E}\left[\mathbb{I}_{\mathcal{A}}\left(\sum_{t=1}^T \bigl(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\bigr)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant 1} \sum_{t=1}^T \bigl(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\bigr)^2\right)\right] \geqslant \frac{1}{2}\frac{c_{11} c_9^2}{\ln\bigl(2+16d^2\bigr)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\ . \tag{A.13}$$

(Note a missing factor of 2 between (A.12) and (A.13).) The last lower bound will then enable us to conclude the proof of this theorem.

Set $\widehat{L}_T \triangleq \sum_{t=1}^T \bigl(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\bigr)^2$ and $L_T(\boldsymbol{u}) \triangleq \sum_{t=1}^T \bigl(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\bigr)^2$ 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. By the equality $\mathbb{I}_{\mathcal{A}} = 1 - \mathbb{I}_{\mathcal{A}^c}$, we have

$$\mathbb{E}\left[\mathbb{I}_{\mathcal{A}}\left(\widehat{L}_T - \inf_{\|\boldsymbol{u}\|_1 \leqslant 1} L_T(\boldsymbol{u})\right)\right] = \mathbb{E}\left[\widehat{L}_T - \inf_{\|\boldsymbol{u}\|_1 \leqslant 1} L_T(\boldsymbol{u})\right] - \mathbb{E}\left[\mathbb{I}_{\mathcal{A}^c}\left(\widehat{L}_T - \inf_{\|\boldsymbol{u}\|_1 \leqslant 1} L_T(\boldsymbol{u})\right)\right]$$
$$\geqslant \frac{c_{11} c_9^2}{\ln\bigl(2+16d^2\bigr)} dY^2\kappa\sqrt{\ln(1+1/\kappa)} - \mathbb{E}\Bigl[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\Bigr]\ , \tag{A.14}$$

where the last inequality follows by (A.12) 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 bounding the above quantity $\mathbb{E}\bigl[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\bigr]$ by half the term on its left. This way, we will have proved (A.13).

First note that

$$\begin{aligned}\mathbb{E}\Big[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\Big] &\triangleq \mathbb{E}\left[\mathbb{I}_{\mathcal{A}^c}\sum_{t=1}^{T}\big(y_t-\widetilde{f}_t(\boldsymbol{x}_t)\big)^2\right]\\ &\leqslant \mathbb{E}\left[\mathbb{I}_{\mathcal{A}^c}\sum_{t=1}^{T}\Big(4Y^2\mathbb{I}_{\{|y_t|\leqslant Y\}}+\big(y_t-\widetilde{f}_t(\boldsymbol{x}_t)\big)^2\mathbb{I}_{\{|y_t|>Y\}}\Big)\right] \qquad \text{(A.15)}\\ &\leqslant 4TY^2\mathbb{P}\big(\mathcal{A}^c\big)+\sum_{t=1}^{T}\mathbb{E}\Big[\big(y_t-\widetilde{f}_t(\boldsymbol{x}_t)\big)^2\mathbb{I}_{\big\{|\varepsilon_t|>\frac{Y}{2\sigma}\big\}}\Big]\ , \qquad \text{(A.16)}\end{aligned}$$

where (A.15) follows from the fact that the online forecaster $(\widetilde{f}_t)_t$ outputs its predictions in $[-Y,Y]$. As for (A.16), 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 $\gamma\triangleq c_8X$, and since $X\leqslant Y/2$ (by definition of X), we get $|y_t|\leqslant c_8\sqrt{2}\,Y/2+\sigma|\varepsilon_t|\leqslant Y/2+\sigma|\varepsilon_t|$ provided that

$$c_8\leqslant\frac{1}{\sqrt{2}}\ , \qquad \text{(A.17)}$$

which we assume thereafter. The above remarks show that $\{|y_t|>Y\}\subset\{|\varepsilon_t|>Y/(2\sigma)\}$, which entails (A.16). By the same comments and since $|\widetilde{f}_t|\leqslant Y$, we have, for all $t=1,\dots,T$,

$$\begin{aligned}\mathbb{E}\Big[\big(y_t-\widetilde{f}_t(\boldsymbol{x}_t)\big)^2\mathbb{I}_{\big\{|\varepsilon_t|>\frac{Y}{2\sigma}\big\}}\Big] &\leqslant \mathbb{E}\Big[\big(Y/2+\sigma|\varepsilon_t|+Y\big)^2\mathbb{I}_{\big\{|\varepsilon_t|>\frac{Y}{2\sigma}\big\}}\Big]\\ &\leqslant 2\left(\frac{3Y}{2}\right)^2\mathbb{P}\left(|\varepsilon_t|>\frac{Y}{2\sigma}\right)+2\sigma^2\mathbb{E}\Big[\varepsilon_t^2\mathbb{I}_{\big\{|\varepsilon_t|>\frac{Y}{2\sigma}\big\}}\Big] \qquad \text{(A.18)}\\ &\leqslant \frac{9Y^2}{2}\mathbb{P}\left(|\varepsilon_t|>\frac{Y}{2\sigma}\right)+2\sigma^2\sqrt{3}\,\mathbb{P}^{1/2}\left(|\varepsilon_t|>\frac{Y}{2\sigma}\right) \qquad \text{(A.19)}\\ &\leqslant 9Y^2T^{-1/(8c_9^2)}+2\frac{c_9^2Y^2}{\ln 2}\sqrt{6}\,T^{-1/(16c_9^2)}\ , \qquad \text{(A.20)}\end{aligned}$$

where we used the following arguments. Inequality (A.18) follows by the elementary inequality $(a+b)^2\leqslant 2(a^2+b^2)$ for all $a,b\in\mathbb{R}$. To get (A.19) 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, (A.20) follows by definition of $\sigma\triangleq c_9Y/\sqrt{\ln T}\leqslant c_9Y/\sqrt{\ln 2}$ and from the fact that, since ε_t is a standard Gaussian random variable[17],

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

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 and substituting (A.20) in (A.16), we get

$$\begin{aligned}\mathbb{E}\Big[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\Big] &\leqslant 8Y^2T^{2-1/(8c_9^2)}+9Y^2T^{1-1/(8c_9^2)}+\frac{2c_9^2\sqrt{6}}{\ln 2}Y^2T^{1-1/(16c_9^2)}\\ &\leqslant 8Y^2 2^{2-1/(8c_9^2)}+9Y^2 2^{1-1/(8c_9^2)}+\frac{2c_9^2\sqrt{6}}{\ln 2}Y^2 2^{1-1/(16c_9^2)}\ , \qquad \text{(A.21)}\end{aligned}$$

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).

[17] We use a standard deviation inequality for subgaussian random variables; see, e.g., [21, Equation (2.5)] with $\sigma^2=1$.

In order to further upper bound $\mathbb{E}\big[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\big]$, we use the following technical lemma, which is proved after the proof of the present theorem (see page 24). It relies on the following elementary argument: since $d\,\kappa$ is large enough and since the left-hand side of the next inequality (Lemma 3) decreases exponentially fast as $c_9 \to 0$, then this inequality holds true for all $c_9 > 0$ small enough.

Lemma 3. *There exists an absolute constant $c_{13} > 0$ such that, for all $c_9 \in (0, c_{13})$,*

$$8Y^2 2^{2-1/(8c_9^2)} + 9Y^2 2^{1-1/(8c_9^2)} + \frac{2c_9^2\sqrt{6}}{\ln 2} Y^2 2^{1-1/(16c_9^2)} \leqslant \frac{1}{2}\frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\;.$$

We can now fix the values of the constants c_8 and c_9 and conclude the proof. Choosing c_9 and $c_8 \triangleq \max\big\{c_9/(2\sqrt{\ln 2}), c_9/(c_7 c_{10})\big\}$ such that $c_8 < 1/\sqrt{2}$ (condition (A.17)), $c_9 < 1/4$, and $c_9 < c_{13}$, then the condition (A.11) also holds, and (A.21) combined with Lemma 3 entails that

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

Substituting the last inequality in (A.14), we get that

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

By the above lower bound and the fact that, $\mathbb{P}_{\boldsymbol{u}^*}^{\gamma;\sigma}$-almost surely, $\|\boldsymbol{x}_t\|_\infty \leqslant \gamma\sqrt{2} \leqslant X$ for all $t = 1,\dots,T$ (since $\gamma \triangleq c_8 X$ and $c_8 \leqslant 1/\sqrt{2}$), we get that

$$\sup_{\substack{\|\boldsymbol{x}_1\|_\infty,\dots,\|\boldsymbol{x}_T\|_\infty\leqslant X\\ y_1,\dots,y_T\in\mathbb{R}}} \left\{\mathbb{I}_{\mathcal{A}}\left(\widehat{L}_T - \inf_{\|\boldsymbol{u}\|_1\leqslant 1} L_T(\boldsymbol{u})\right)\right\} \geqslant \frac{1}{2}\frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\;.$$

Therefore, by definition of $\mathcal{A} \triangleq \bigcap_{t=1}^T\big\{|y_t| \leqslant Y\big\}$, of $\widehat{L}_T \triangleq \sum_{t=1}^T\big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2$, and of $L_T(\boldsymbol{u}) \triangleq \sum_{t=1}^T(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2$, we get that, for all online forecasters $(\widetilde{f}_t)_{t\geqslant 1}$ whose predictions lie in $[-Y, Y]$,

$$\sup_{\substack{\|\boldsymbol{x}_1\|_\infty,\dots,\|\boldsymbol{x}_T\|_\infty\leqslant X\\ |y_1|,\dots,|y_T|\leqslant Y}} \left\{\sum_{t=1}^T\big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1\leqslant U}\sum_{t=1}^T(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2\right\} \geqslant \frac{1}{2}\frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\;.$$

Combining the last lower bound with (A.3) and setting $c_1 \triangleq c_{11}c_9^2/2$ concludes the proof under the assumption $\sqrt{\ln(1+2d)}/\big(2d\sqrt{\ln 2}\big) \leqslant \kappa \leqslant 1$.

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 $\sqrt{T}U_1 X/(2dY) = 1$ and

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

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 $U \triangleq \kappa U_1 > U_1$, 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. □

Proof (of Lemma 2): We use the same notations as in Step 1 of the proof of Theorem 2. Let (X', y') be a random copy of (X_1, y_1) independent of the sample $(X_t, y_t)_{1\leqslant t\leqslant T}$, and define the random vector $\boldsymbol{x}' \triangleq \big(\gamma\varphi_1(X'), \dots, \gamma\varphi_d(X')\big)$. By the tower rule, we have

$$\mathbb{E}\big[(y_t - \widetilde{f}_t(\boldsymbol{x}_t)^2\big] = \mathbb{E}\Big[\mathbb{E}\big[(y_t - \widetilde{f}_t(\boldsymbol{x}_t))^2\big|(\boldsymbol{x}_s, y_s)_{s\leqslant t-1}\big]\Big] = \mathbb{E}\big[(y' - \widetilde{f}_t(\boldsymbol{x}')^2\big] \ ,$$

where we used the fact that $\widetilde{f}_t$ is built on the past data $(\boldsymbol{x}_s, y_s)_{s\leqslant t-1}$ and that $(\boldsymbol{x}', y')$ and $(\boldsymbol{x}_t, y_t)$ are both independent of $(\boldsymbol{x}_s, y_s)_{s\leqslant t-1}$ and are identically distributed. Similarly $\mathbb{E}\big[(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2\big] = \mathbb{E}\big[(y' - \boldsymbol{u}\cdot\boldsymbol{x}')^2\big]$. Using the last equalities and the fact that $\mathbb{E}\big[\inf\{\dots\}\big] \leqslant \inf\mathbb{E}\big[\{\dots\}\big]$, we get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1\leqslant 1}\sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2\right] \\
&\quad\geqslant T\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\Big[\big(y' - \widetilde{f}_t(\boldsymbol{x}')\big)^2\Big] - \inf_{\|\boldsymbol{u}\|_1\leqslant 1}\mathbb{E}\Big[\big(y' - \boldsymbol{u}\cdot\boldsymbol{x}'\big)^2\Big]\right) \\
&\quad\geqslant T\left(\mathbb{E}\Big[\big(y' - \widehat{f}_T(X')\big)^2\Big] - \inf_{\|\boldsymbol{u}\|_1\leqslant 1}\mathbb{E}\Big[\big(y' - \boldsymbol{u}\cdot\boldsymbol{x}'\big)^2\Big]\right) &\text{(A.22)} \\
&\quad= T\,\mathbb{E}\Big[\big(\gamma\varphi_{\boldsymbol{u}^*}(X') - \widehat{f}_T(X')\big)^2\Big] &\text{(A.23)} \\
&\quad= T\,\mathbb{E}\left\|\widehat{f}_T - \gamma\varphi_{\boldsymbol{u}^*}\right\|_\mu^2 \ .
\end{aligned}$$

Inequality (A.22) follows by definition of $\widehat{f}_T \triangleq T^{-1}\sum_{t=1}^T \widetilde{f}_t$ (see (A.5)) and by Jensen's inequality. As for Inequality (A.23), it follows by expanding the square

$$\big(y' - \widehat{f}_T(X')\big)^2 = \big(\gamma\varphi_{\boldsymbol{u}^*}(X') - \widehat{f}_T(X') + y' - \gamma\varphi_{\boldsymbol{u}^*}(X')\big)^2 \ ,$$

by noting that $\mathbb{E}\big[y' - \gamma\varphi_{\boldsymbol{u}^*}(X')\big|X'\big] = 0$ (via (A.6)) 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 $\boldsymbol{u}\cdot\boldsymbol{x}' = \gamma\varphi_{\boldsymbol{u}}(X')$. This concludes the proof. □

Proof (of Lemma 3): We use the same notations and assumptions as in the proof of Theorem 2. Since the function $x \mapsto x\sqrt{\ln(1+1/x)}$ is nondecreasing on $\mathbb{R}_+^*$ and since $\kappa \geqslant \kappa_{\min} \triangleq \sqrt{\ln(1+2d)}/(2d\sqrt{\ln 2})$ by assumption, we have

$$\begin{aligned}
&\frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)}dY^2\kappa\sqrt{\ln(1+1/\kappa)} \\
&\quad\geqslant \frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)}dY^2\kappa_{\min}\sqrt{\ln(1+1/\kappa_{\min})} \\
&\quad= \frac{c_{11}c_9^2}{2\sqrt{\ln 2}}Y^2\frac{\sqrt{\ln(1+2d)}\sqrt{\ln\Big[1+2d\sqrt{\ln 2}/\sqrt{\ln(1+2d)}\Big]}}{\ln\big(2+16d^2\big)} &\text{(A.24)} \\
&\quad\geqslant \frac{c_{11}c_9^2}{2\sqrt{\ln 2}}Y^2c_{12} \ , &\text{(A.25)}
\end{aligned}$$

where c_{12} denotes the infimum of the last fraction of (A.24) over all $d \geqslant 1$; in particular, $c_{12} > 0$. It is now easy to see that by choosing the absolute constant $c_{13} > 0$ small enough (where c_{13} can be expressed in terms of c_{11} and c_{12}), we have, for all $c_9 \in (0, c_{13})$,

$$8 \cdot 2^{2-1/(8c_9^2)} + 9 \cdot 2^{1-1/(8c_9^2)} + \frac{2c_9^2\sqrt{6}}{\ln 2} 2^{1-1/(16c_9^2)} \leqslant \frac{c_{11}c_9^2}{2\sqrt{\ln 2}} c_{12} \ .$$

Multiplying both sides of the last inequality by Y^2 and combining it with (A.25) concludes the proof. □

Appendix A.2. Proofs of Theorem 3 and Remark 1

Proof (of Theorem 3): The proof follows directly from Proposition 1 and from the fact that the Lipschitzified losses are larger than their clipped versions. Indeed, first note that, by definition of $\widehat{y}_t$ and $B_{t+1} \geqslant |y_t|$, we have

$$\begin{aligned}
\sum_{t=1}^T |y_t - \widehat{y}_t|^\alpha &\leqslant \sum_{\substack{t=1 \\ t:|y_t| \leqslant B_t}}^T \left| y_t - \left[\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t\right]_{B_t} \right|^\alpha + \sum_{\substack{t=1 \\ t:|y_t| > B_t}}^T (B_{t+1} + B_t)^\alpha \\
&\leqslant \sum_{\substack{t=1 \\ t:|y_t| \leqslant B_t}}^T \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) + \left(1 + 2^{-1/\alpha}\right)^\alpha \sum_{\substack{t=1 \\ t:B_{t+1} > B_t}}^T B_{t+1}^\alpha \\
&\leqslant \sum_{t=1}^T \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) + 4\left(1 + 2^{-1/\alpha}\right)^\alpha Y^\alpha \ , \qquad \text{(A.26)}
\end{aligned}$$

where the second inequality follows from the fact that:

- if $|y_t| \leqslant B_t$ then $|y_t - [\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t]_{B_t}|^\alpha \leqslant \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)$ by Eq. (13);
- if $|y_t| > B_t$, which is equivalent to $B_{t+1} > B_t$ by definition of B_{t+1}, then $B_t \leqslant B_{t+1}/2^{1/\alpha}$, so that $B_{t+1} + B_t \leqslant \left(1 + 2^{-1/\alpha}\right) B_{t+1}$.

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 $\sum_{t:B_{t+1}>B_t}^T B_{t+1}^\alpha$ in the same way as in [11, Theorem 6], i.e., setting $K \triangleq \left\lceil \log_2 \max_{1 \leqslant t \leqslant T} |y_t|^\alpha \right\rceil$,

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

To bound (A.26) further from above, we now use the fact that, by construction, the LEG algorithm is the adaptive EG$^\pm$ algorithm applied to the modified loss functions $\widetilde{\ell}_t$. Therefore, we get from Proposition 1 that

$$\begin{aligned}
\sum_{t=1}^T \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) \leqslant \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T \widetilde{\ell}_t(\boldsymbol{u}) \\
+ 4U\sqrt{\left(\sum_{t=1}^T \left\| \nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) \right\|_\infty^2\right) \ln(2d)} + U\left(8\ln(2d) + 12\right) \max_{1 \leqslant t \leqslant T} \left\| \nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) \right\|_\infty \ . \qquad \text{(A.27)}
\end{aligned}$$

We can now follow the same lines as in Corollary 2, except that we use the particular shape of the Lipschitzified losses. We first derive some properties of the gradients $\nabla\widetilde{\ell}_t$. Observe from the definition of $\widetilde{\ell}_t$ in Section 3.3 that in both cases $|y_t| > B_t$ and $|y_t| \leqslant B_t$, the function $\widetilde{\ell}_t$ is continuously differentiable. Moreover, if $|y_t| \leqslant B_t$, then

$$\forall \boldsymbol{u} \in \mathbb{R}^d \ , \quad \nabla\widetilde{\ell}_t(\boldsymbol{u}) = -\alpha \operatorname{sgn}\left(y_t - [\boldsymbol{u} \cdot \boldsymbol{x}_t]_{B_t}\right) |y_t - [\boldsymbol{u} \cdot \boldsymbol{x}_t]_{B_t}|^{\alpha-1} \boldsymbol{x}_t \ ,$$

where for all $x \in \mathbb{R}$, the quantity $\mathrm{sgn}(x)$ equals 1 (resp. -1, 0) if $x > 0$ (resp. $x < 0$, $x = 0$).

Therefore, in both cases $|y_t| > B_t$ and $|y_t| \leqslant B_t$, the function $\widetilde{\ell}_t$ is Lipschitz continuous with respect to $\|\cdot\|_1$ with Lipschitz constant $\sup_{\boldsymbol{u}\in\mathbb{R}^d} \left\|\nabla\widetilde{\ell}_t\right\|_\infty$ bounded as follows: for all $\boldsymbol{u} \in \mathbb{R}^d$,

$$\left\|\nabla\widetilde{\ell}_t(\boldsymbol{u})\right\|_\infty \leqslant \alpha \, |y_t - [\boldsymbol{u}\cdot\boldsymbol{x}_t]_{B_t}|^{\alpha-1} \, \|\boldsymbol{x}_t\|_\infty \tag{A.28}$$

$$\leqslant \alpha \left(|y_t| + B_t\right)^{\alpha-1} \|\boldsymbol{x}_t\|_\infty \leqslant \alpha \left(1 + 2^{1/\alpha}\right)^{\alpha-1} \left(\max_{1\leqslant s\leqslant t} |y_s|\right)^{\alpha-1} \|\boldsymbol{x}_t\|_\infty \ , \tag{A.29}$$

where we used the fact that $B_t \leqslant 2^{1/\alpha} \max_{1\leqslant s\leqslant t-1} |y_s|$.

We can draw several consequences from the inequalities above. First note that, by (A.29),

$$\max_{1\leqslant t\leqslant T} \|\nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)\|_\infty \leqslant \alpha\big(1 + 2^{1/\alpha}\big)^{\alpha-1} X Y^{\alpha-1} \ . \tag{A.30}$$

Moreover, using (A.28) and the definition of $\widehat{y}_t$ in Figure 4, we can see that the gradients $\nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)$ satisfy $\left\|\nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)\right\|_\infty \leqslant \alpha \, |y_t - \widehat{y}_t|^{\alpha-1} \, \|\boldsymbol{x}_t\|_\infty \leqslant \alpha X \, |y_t - \widehat{y}_t|^{\alpha-1}$. This entails that

$$\left\|\nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)\right\|_\infty^2 \leqslant \alpha^2 X^2 \big|y_t - \widehat{y}_t\big|^{2\alpha-2} = \alpha^2 X^2 \big|y_t - \widehat{y}_t\big|^{\alpha-2} \big|y_t - \widehat{y}_t\big|^{\alpha}$$

$$\leqslant \alpha^2 X^2 \big((1 + 2^{1/\alpha})Y\big)^{\alpha-2} \big|y_t - \widehat{y}_t\big|^{\alpha} \ , \tag{A.31}$$

where we used the upper bounds $|y_t| \leqslant Y$ and $|\widehat{y}_t| \triangleq \left|\left[\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t\right]_{B_t}\right| \leqslant B_t \leqslant 2^{1/\alpha} Y$. Substituting (A.30) and (A.31) in (A.27) and combining the resulting bound with (A.26), we get

$$\sum_{t=1}^T |y_t - \widehat{y}_t|^\alpha \leqslant \inf_{\|\boldsymbol{u}\|_1\leqslant U} \sum_{t=1}^T \widetilde{\ell}_t(\boldsymbol{u}) + a_\alpha U X Y^{\alpha/2-1} \sqrt{\left(\sum_{t=1}^T |y_t - \widehat{y}_t|^\alpha\right) \ln(2d)}$$

$$+ \underbrace{\big(8\ln(2d) + 12\big) \, b_\alpha \, U X Y^{\alpha-1}}_{\triangleq C_1} + \underbrace{4\big(1 + 2^{-1/\alpha}\big)^\alpha Y^\alpha}_{\triangleq C_2} \ ,$$

where we set $a_\alpha \triangleq 4\alpha \left(1 + 2^{1/\alpha}\right)^{\alpha/2-1}$ and $b_\alpha \triangleq \alpha \left(1 + 2^{1/\alpha}\right)^{\alpha-1}$.

To simplify the notations we also set $\widehat{L}_T \triangleq \sum_{t=1}^T |y_t - \widehat{y}_t|^\alpha$ and $\widetilde{L}_T^* \triangleq \min_{\|\boldsymbol{u}\|_1\leqslant U} \sum_{t=1}^T \widetilde{\ell}_t(\boldsymbol{u})$, so that the previous inequality can be rewritten as

$$\widehat{L}_T \leqslant \widetilde{L}_T^* + C_1 + C_2 + a_\alpha U X Y^{\alpha/2-1} \sqrt{\widehat{L}_T \ln(2d)} \ .$$

Solving for $\widehat{L}_T$ via Lemma 4 in Appendix B (used with $a = \widetilde{L}_T^* + C_1 + C_2$ and $b = a_\alpha U X Y^{\alpha/2-1}\sqrt{\ln(2d)}$), we get that

$$\widehat{L}_T \leqslant \widetilde{L}_T^* + C_1 + C_2 + \left(a_\alpha U X Y^{\alpha/2-1} \sqrt{\ln(2d)}\right) \sqrt{\widetilde{L}_T^* + C_1 + C_2} + \left(a_\alpha U X Y^{\alpha/2-1} \sqrt{\ln(2d)}\right)^2$$

$$\leqslant \widetilde{L}_T^* + a_\alpha U X Y^{\alpha/2-1} \sqrt{\widetilde{L}_T^* \ln(2d)}$$

$$+ a_\alpha U X Y^{\alpha/2-1} \sqrt{(C_1 + C_2)\ln(2d)} + a_\alpha^2 U^2 X^2 Y^{\alpha-2} \ln(2d) + C_1 + C_2 \ . \tag{A.32}$$

To conclude the proof, it just suffices to bound the term $a_\alpha U X Y^{\alpha/2-1} \sqrt{(C_1 + C_2)\ln(2d)}$ from above. First note that

$$\sqrt{(C_1 + C_2)\ln(2d)} \leqslant \sqrt{C_1 \ln(2d)} + \sqrt{C_2 \ln(2d)}$$

$$\leqslant \sqrt{C_1 \ln(2d)} + 2\big(1 + 2^{-1/\alpha}\big)^{\alpha/2} Y^{\alpha/2} \sqrt{\ln(2d)} \ , \tag{A.33}$$

where the last inequality follows by definition of C_2 above. Now, to upper bound $\sqrt{C_1 \ln(2d)}$, we note that, by definition of C_1,

$$\begin{aligned}\sqrt{C_1 \ln(2d)} &= \ln(2d) \sqrt{\big(8 + 12/\ln(2d)\big)\, b_\alpha\, UXY^{\alpha-1}} \\ &\leqslant \ln(2d) \sqrt{\big(8 + 12/\ln 2\big)\, b_\alpha}\; \frac{UXY^{\alpha/2-1} + Y^{\alpha/2}}{\sqrt{2}} \;,\end{aligned}$$

where we used the elementary upper bound $\sqrt{ab} \leqslant (a+b)/2$ with $a = UXY^{\alpha/2-1}$ and $b = Y^{\alpha/2}$. Substituting the last inequality in (A.33) and using $\sqrt{\ln(2d)} \leqslant \ln(2d)/\sqrt{\ln 2}$, we finally get that

$$\begin{aligned}&a_\alpha UXY^{\alpha/2-1} \sqrt{(C_1 + C_2)\ln(2d)} \\ &\quad\leqslant a_\alpha \ln(2d) \left(\sqrt{b_\alpha\big(4 + 6/\ln 2\big)} + 2\big(1 + 2^{-1/\alpha}\big)^{\alpha/2}/\sqrt{\ln 2} \right) UXY^{\alpha-1} \\ &\qquad + a_\alpha \ln(2d) \sqrt{b_\alpha\big(4 + 6/\ln 2\big)}\; U^2X^2Y^{\alpha-2} \;.\end{aligned}$$

Substituting the last inequality into (A.32) and rearranging terms concludes the proof. □

Proof (of Remark 1): Recall that in this remark, we focus on the square loss (i.e., $\alpha = 2$) and that we set $c_1 \triangleq 8\big(\sqrt{2}+1\big)$ and $c_2 \triangleq 4\left(1 + 1/\sqrt{2}\right)^2$. By the key property (13) that holds for all rounds t such that $|y_t| \leqslant B_t$ (the other rounds accounting only for an additional total loss at most of $c_2 Y^2$, see (A.26)), we get

$$\begin{aligned}\sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 &\leqslant \sum_{t=1}^T \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T \widetilde{\ell}_t(\boldsymbol{u}) + c_2 Y^2 \\ &\leqslant 4U \max_{1\leqslant t\leqslant T} \left\|\nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)\right\|_\infty \left(\sqrt{T\ln(2d)} + 2\ln(2d) + 3\right) + c_2 Y^2 \qquad \text{(A.34)} \\ &\leqslant c_1 UXY \left(\sqrt{T\ln(2d)} + 8\ln(2d)\right) + c_2 Y^2 \;, \qquad \text{(A.35)}\end{aligned}$$

where (A.34) follows from the remark in Proposition 1 involving the uniform bound $\max_{1\leqslant t\leqslant T} \|\nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)\|_\infty$, and where (A.35) follows from $\max_{1\leqslant t\leqslant T} \|\nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)\|_\infty \leqslant 2\big(1+\sqrt{2}\big)XY$ (by (A.29)) and from the elementary inequality $3 \leqslant 6\ln(2d)$. □

Appendix B. Lemmas

The next elementary lemma is due to [22, Appendix III]. It is useful to compute an upper bound on the cumulative loss $\widehat{L}_T$ of a forecaster when $\widehat{L}_T$ satisfies an inequality of the form (B.1).

Lemma 4. *Let $a, b \geqslant 0$. Assume that $x \geqslant 0$ satisfies the inequality*

$$x \leqslant a + b\sqrt{x} \;. \qquad \text{(B.1)}$$

Then,

$$x \leqslant a + b\sqrt{a} + b^2 \;.$$

The next lemma is useful to prove Theorem 1. At the end of this section, we also provide an elementary lemma about the exponentially weighted average forecaster combined with clipping.

Lemma 5. *Let $d, T \in \mathbb{N}^*$, and $U, X, Y > 0$. The minimax regret on $B_1(U)$ for bounded base predictions and observations satisfies*

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

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 $\|\boldsymbol{x}_1\|_\infty, \ldots, \|\boldsymbol{x}_T\|_\infty \leqslant X$.

Proof: We treat each of the two terms in the above minimum separately.

Step 1: We prove that their exists a forecaster F whose worst-case regret on $B_1(U)$ is upper bounded by $3UXY\sqrt{2T\ln(2d)}$.

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 forecaster F 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 [9, 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

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

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

Step 2: We prove that their exists a forecaster F whose worst-case regret on $B_1(U)$ is upper bounded by $32\, dY^2 \ln\!\left(1 + \frac{\sqrt{T}UX}{dY}\right) + dY^2$.

Such a forecaster is given by the sparsity-oriented algorithm SeqSEW$^{B,\eta}_\tau$ of [12] (we could also get a slightly worse bound with the sequential ridge regression forecaster of [13, 14]). Indeed, by [12, Proposition 1], the cumulative square loss of the algorithm SeqSEW$^{B,\eta}_\tau$ tuned with $B = Y$, $\eta = 1/(8Y^2)$ and $\tau = Y/(\sqrt{T}X)$ is upper bounded by

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

where the last inequality follows by monotonicity[18] in $\|\boldsymbol{u}\|_0$ and $\|\boldsymbol{u}\|_1$ of the second term of the left-hand side. This concludes the proof. □

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

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

In the equation above, $[x]_Y \triangleq \min\{Y, \max\{-Y, x\}\}$ for all $x \in \mathbb{R}$, and the weight vectors $\boldsymbol{p}_t \in \mathbb{R}^K$ are given by $\boldsymbol{p}_1 = (1/K, \dots, 1/K)$ and, for all $t = 2, \dots, T$, by

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

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 [17].

Lemma 6 (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*

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

Proof (of Lemma 6): The proof follows straightforwardly from Theorem 3.2 and Proposition 3.1 of [17]. To apply the latter result, recall from [14, Remark 3] that the square loss is $1/(8Y^2)$-exp-concave on $[-Y, Y]$ and thus η-exp-concave[19] (since $\eta \leqslant 1/(8Y^2)$ by assumption). Therefore, by definition of our forecaster above, Theorem 3.2 and Proposition 3.1 of [17] yield

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

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., $\left(y_t - \big[\widehat{y}_t^{(k)}\big]_Y\right)^2 \leqslant \big(y_t - \widehat{y}_t^{(k)}\big)^2$. □

Appendix C. Additional tools

The next approximation argument is originally due to Maurey, and was used under various forms, e.g., in [1, 2, 3, 4] (see also [5]).

[18] 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 $\mathbb{R}_+$.

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

Lemma 7 (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 $\max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty \leqslant X$,

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

Proof: The proof is quite standard and follows the same lines as [1, Proposition 5.2.2] or [3, 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,

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

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

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

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 (C.2) above equals zero, and

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

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 (C.2) 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 (C.1) concludes the proof. □

The combinatorial result below (or variants of it) is well-known; see, e.g., [2, 3]. We reproduce its proof for the convenience of the reader. We use the notation $\mathrm{e} \triangleq \exp(1)$.

Lemma 8 (An elementary combinatorial upper bound)**.**
Let $m, d \in \mathbb{N}^*$. Denoting by $|E|$ the cardinality of a set E, we have

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

Proof (of Lemma 8): Setting $(k'_{-j}, k'_j) \triangleq \big((k_j)_-, (k_j)_+\big)$ for all $1 \leqslant j \leqslant d$, and $k'_0 \triangleq m - \sum_{j=1}^{d} |k_j|$, we have

$$\begin{aligned}
\left| \left\{ (k_1, \ldots, k_d) \in \mathbb{Z}^d : \sum_{j=1}^{d} |k_j| \leqslant m \right\} \right| &\leqslant \left| \left\{ (k'_{-d}, \ldots, k'_d) \in \mathbb{N}^{2d+1} : \sum_{j=-d}^{d} k'_j = m \right\} \right| \\
&= \binom{2d+m}{m} \qquad \text{(C.3)} \\
&\leqslant \left(\frac{\mathrm{e}(2d+m)}{m} \right)^m \ . \qquad \text{(C.4)}
\end{aligned}$$

To get inequality (C.3), 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 (C.4), it follows straightforwardly from a classical combinatorial result stated, e.g., in [21, Proposition 2.5]. □

References

[1] A. Nemirovski, Topics in Non-Parametric Statistics, Springer, Berlin/Heidelberg/New York, 2000.

[2] A. B. Tsybakov, Optimal rates of aggregation, in: Proceedings of the 16th Annual Conference on Learning Theory (COLT'03), 2003, pp. 303–313.

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

[4] S. Shalev-Shwartz, N. Srebro, T. Zhang, Trading accuracy for sparsity in optimization problems with sparsity constraints, SIAM J. Optim. 20 (6) (2010) 2807–2832.
[5] Y. Yang, Aggregating regression procedures to improve performance, Bernoulli 10 (1) (2004) 25–47.
[6] L. Birgé, P. Massart, Gaussian model selection, J. Eur. Math. Soc. 3 (2001) 203–268.
[7] G. Raskutti, M. J. Wainwright, B. Yu, Minimax rates of estimation for high-dimensional linear regression over ℓ^q-balls, IEEE Trans. Inform. Theory 57 (10) (2011) 6976–6994.
[8] N. Cesa-Bianchi, Analysis of two gradient-based algorithms for on-line regression, J. Comput. System Sci. 59 (3) (1999) 392–411.
[9] J. Kivinen, M. K. Warmuth, Exponentiated gradient versus gradient descent for linear predictors, Inform. and Comput. 132 (1) (1997) 1–63.
[10] P. Auer, N. Cesa-Bianchi, C. Gentile, Adaptive and self-confident on-line learning algorithms, J. Comp. Sys. Sci. 64 (2002) 48–75.
[11] N. Cesa-Bianchi, Y. Mansour, G. Stoltz, Improved second-order bounds for prediction with expert advice, Mach. Learn. 66 (2/3) (2007) 321–352.
[12] S. Gerchinovitz, Sparsity regret bounds for individual sequences in online linear regression, JMLR Workshop and Conference Proceedings 19 (COLT 2011 Proceedings) (2011) 377–396.
[13] K. S. Azoury, M. K. Warmuth, Relative loss bounds for on-line density estimation with the exponential family of distributions, Mach. Learn. 43 (3) (2001) 211–246.
[14] V. Vovk, Competitive on-line statistics, Internat. Statist. Rev. 69 (2001) 213–248.
[15] M. Zinkevich, Online convex programming and generalized infinitesimal gradient ascent, in: Proceedings of the 20th International Conference on Machine Learning (ICML'03), 2003, pp. 928–936.
[16] S. Shalev-Shwartz, O. Shamir, N. Srebro, K. Sridharan, Stochastic convex optimization, in: Proceedings of the 22nd Annual Conference on Learning Theory (COLT'09), 2009, pp. 177–186.
[17] N. Cesa-Bianchi, G. Lugosi, Prediction, Learning, and Games, Cambridge University Press, 2006.
[18] C. Gentile, The robustness of the p-norm algorithms, Mach. Learn. 53 (3) (2003) 265–299.
[19] S. Shalev-Shwartz, A. Tewari, Stochastic methods for ℓ^1-regularized loss minimization, in: Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09), 2009, pp. 929–936.
[20] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, A. Tewari, Composite objective mirror descent, in: Proceedings of the 23rd Annual Conference on Learning Theory (COLT'10), 2010, pp. 14–26.
[21] P. Massart, Concentration Inequalities and Model Selection, Vol. 1896 of Lecture Notes in Mathematics, Springer, Berlin, 2007.
[22] N. Cesa-Bianchi, G. Lugosi, G. Stoltz, Minimizing regret with label efficient prediction, IEEE Trans. Inform. Theory 51 (6) (2005) 2152–2162.