Efficient online algorithms for fast-rate regret bounds under sparsity

Pierre Gaillard, Olivier Wintenberger

▶ **To cite this version:**

Pierre Gaillard, Olivier Wintenberger. Efficient online algorithms for fast-rate regret bounds under sparsity. NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, Dec 2018, Montreal, France. hal-01798201

HAL Id: hal-01798201
https://hal.science/hal-01798201

Submitted on 23 May 2018

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

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

EFFICIENT ONLINE ALGORITHMS FOR FAST-RATE REGRET BOUNDS UNDER SPARSITY

PIERRE GAILLARD[1] AND OLIVIER WINTENBERGER[2]

Abstract. We consider the online convex optimization problem. In the setting of arbitrary sequences and finite set of parameters, we establish a new fast-rate quantile regret bound. Then we investigate the optimization into the ℓ_1-ball by discretizing the parameter space. Our algorithm is projection free and we propose an efficient solution by restarting the algorithm on adaptive discretization grids. In the adversarial setting, we develop an algorithm that achieves several rates of convergence with different dependences on the sparsity of the objective. In the i.i.d. setting, we establish new risk bounds that are adaptive to the sparsity of the problem and to the regularity of the risk (ranging from a rate $1/\sqrt{T}$ for general convex risk to $1/T$ for strongly convex risk). These results generalize previous works on sparse online learning. They are obtained under a weak assumption on the risk (Łojasiewicz's assumption) that allows multiple optima which is crucial when dealing with degenerate situations.

1. Introduction

We consider the following setting of online convex prediction. Let $(\ell_t : \mathbb{R}^d \to \mathbb{R})_{t\geqslant 1}$ be a collection of random convex sub-differentiable loss functions sequentially observed. At each time step $t \geqslant 1$, a learner forms a prediction $\widehat{\theta}_{t-1} \in \mathbb{R}^d$ based on past observations $\mathcal{F}_{t-1} = \{\ell_1, \widehat{\theta}_1, \dots, \ell_{t-1}, \widehat{\theta}_{t-1}\}$. The learner aims at minimizing its average risk

$$R_T(\theta) := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{t-1}\big[\ell_t(\widehat{\theta}_{t-1})\big] - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{t-1}\big[\ell_t(\theta)\big] \qquad \text{where} \quad \mathbb{E}_{t-1} = \mathbb{E}[\,\cdot\,|\mathcal{F}_{t-1}]\,, \tag{1}$$

with respect to all θ in some reference set $\Theta \subseteq \mathcal{B}_1 := \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leqslant 1\}$. By considering the Dirac masses, one obtains $\ell_t = \mathbb{E}_{t-1}[\ell_t]$ and the average risk matches the definition $(1/T)\sum_{t=1}^{T}\big(\ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta)\big)$ of the average regret more commonly used in the online learning literature. We will first consider finite set Θ. Then we will show how to extend the results to the unit ℓ_1-ball $\mathcal{B}_1$ providing sparsity guarantees for sparse $\theta \in \mathcal{B}_1$.

Related work. The case of finite reference set Θ corresponds to the setting of prediction with expert advice (see Section 2.2 or [Cesa-Bianchi and Lugosi, 2006, Freund and Schapire, 1997, Vovk, 1998]), where a learner makes sequential predictions over a series of rounds with the help of K experts. Littlestone and Warmuth [1994] and Vovk [1990] introduced the exponentially weighted average algorithm (Hedge) which achieves the optimal rate of convergence $\mathcal{O}(1/\sqrt{T})$ for the average regret for general convex functions. Several works focused on improving the rate of convergence under nice properties of the loss or the data. For instance, Hedge ensures a rate $\mathcal{O}(1/T)$ for exp-concave loss functions. We refer to Van Erven et al. [2015] for a thorough review of fast-rate type assumptions on the losses.

The extension from finite reference sets to convex sets is natural. The seminal paper Kivinen and Warmuth [1997] introduced the Exponentiated Gradient algorithm (EG), a version of Hedge using gradient version of the losses. The latter guarantees a $\mathcal{O}(1/\sqrt{T})$ average regret uniformly over the unit ℓ_1-ball $\mathcal{B}_1$. Another approach consists in projecting gradient descent steps (see Zinkevich [2003] for general convex set, Duchi et al. [2008] for the ℓ_1-ball, or Agarwal et al. [2012] for fast rates under sparsity).

[1] INRIA - Département d'informatique de l'ENS, Ecole normale supérieure, CNRS, INRIA, PSL Research University, 75005 Paris, France

[2] Sorbonne Université, LPSM, Paris, France

E-mail addresses: pierre.gaillard@inria.fr, olivier.wintenberger@upmc.fr.

First works in i.i.d. online convex optimization under sparsity was done by Agarwal et al. [2012], Gaillard and Wintenberger [2017], Steinhardt et al. [2014] that obtained sparse rates of order $\tilde{\mathcal{O}}(\|\theta^*\|_0 \ln d/T)$[1]. Their settings are very close to the one of Bunea et al. [2007] used for studying the convergence properties of the LASSO batch procedure. Their methods differ; the one of Steinhardt et al. [2014] uses a ℓ_1-penalized gradient descent whereas the one of Agarwal et al. [2012] and Gaillard and Wintenberger [2017] are based on restarting a subroutine centered around the current estimate, on sessions of exponentially growing length. These works compete with the optima over $\mathbb{R}^d$ assumed to be (approximately in Agarwal et al. [2012]) sparse with a known ℓ_1-bound. In contrast, we only compete here with optima over $\mathcal{B}_1$ which are more likely to be sparse.

Little work was done on sparsity under adversarial data. The papers Langford et al. [2009], Xiao [2010], Duchi et al. [2010] focus on providing sparse estimators with rates of order $\mathcal{O}(1/\sqrt{T})$ or a linear dependency on the dimension d. Recent work (see Foster et al. [2016], Kale et al. [2017] and references therein) considers the problem where the learner only observes a sparse subset of coordinates at each round. Though they also compare themselves with sparse parameters, they also suffer a bound larger than $\mathcal{O}(1/\sqrt{T})$. Fast rate sparse regret bounds involving $\|\theta\|_0$ were, to our knowledge, only obtained through non-efficient procedures (see Gerchinovitz [2011] or Rakhlin and Sridharan [2015]).

Contributions and outline of the paper. In this paper we focus on providing fast rate regret bounds involving the sparsity of the objective $\|\theta\|_0$.

In Section 2 we start with the finite case $\Theta = \{\theta_1, \dots, \theta_K\}$. We extend the results of Wintenberger [2014] and Koolen and Van Erven [2015] under a weak version of exp-concavity, see Assumption (A2). We show in Theorem 2.1 that the Bernstein Online Aggregation (BOA) and Squint algorithms achieve a fast rate with high probability: i.e. $R_T(\theta) \leqslant \mathcal{O}((\ln K)/T)$ for arbitrary data. The theorem also provides a quantile bound on the risk which improves the dependency on K if many experts are performing well. This is the first quantile-like bound on the average risk that provides fast-rate with high probability. Mehta [2016] developed high-probability quantile bounds but it was degrading with an additional gap term.

In Section 3, we consider the case $\Theta = \mathcal{B}_1$. The standard reduction using the "gradient trick" of Kivinen and Warmuth [1997], looses the fast-rate guaranty obtained under Assumption (A2). Considering BOA on a discretization grid Θ_0 of Θ and applying Theorem 2.1 yields optimal convergence rate under (A2). Yet, the complexity of the discretization is prohibitive. We thus investigate how an a-priori discretization grid Θ_0 may be used to improve the regret bound. We provide in Theorem 3.2 a bound of the form $R_T(\theta) \leqslant \mathcal{O}(D(\theta, \Theta_0)/\sqrt{T})$ which we call *accelerable*, i.e. the rate may decrease if $D(\theta, \Theta_0)$ decreases with T. Here D is a pseudo-metric that we call *averaging accelerability* and $D(\theta, \Theta_0)$ is the distance of θ with Θ_0 in this pseudo-metric. Our bound yields an oracle bound of the form $R_T(\theta) \leqslant \mathcal{O}(\|\theta\|_1/\sqrt{T})$ which was recently studied by Foster et al. [2017]. The following sections 3.3 and 3.4 build the grid Θ_0 adaptively in order to ensure a small regret under a sparsity scenario: Section 3.3 in the adversarial setting and Section 3.4 for i.i.d. losses.

In Section 3.3, we work under the strong convexity assumption on the losses in the adversarial setting. Using a doubling trick, we show that including sparse versions of the leader of the last session in Θ_0 is enough to ensure that $R_T(\theta) \leqslant \tilde{\mathcal{O}}\big((\sqrt{d\|\theta\|_0}/T) \wedge (\sqrt{\|\theta\|_0}/T^{3/4})\big)$ for all $\theta \in \mathcal{B}_1$. The rate is faster than the usual rate of convergence $\mathcal{O}(d/T)$ obtained by online gradient descent or online newton step Hazan et al. [2007]. The gain $\sqrt{\|\theta\|_0/d} \wedge \sqrt{\|\theta\|_0/T}$ is significant for sparse parameters θ. The numerical and space complexities of the algorithm, called BOA+, are $\tilde{\mathcal{O}}(dT)$. Notice that the rate can be decreased to $\tilde{\mathcal{O}}(d_0/T)$ whenever the leaders and the parameter θ are d_0-sparse. This favorable case is not likely to happen in the adversarial setting but do happen in the i.i.d. setting treated in Section 3.4.

A new difficulty raises in the i.i.d. setting: we accept only assumptions on the risk $\mathbb{E}[\ell_t]$ and not on the losses ℓ_t. To do so, we need to enrich the grid Θ_0 with good approximations of the optima of the risk $\mathbb{E}[\ell_t]$. However, the risk is not observed and the minimizer of the empirical risk (the leader)

[1]Throughout the paper $\lesssim$ denotes an approximative inequality which holds up to universal constants and $\tilde{\mathcal{O}}$ denotes an asymptotic inequality up to logarithmic terms in T and dependence on parameters not clarified.

suffer a rate of convergence linear in d. Thus, we develop another algorithm, called SABOA, that sequentially enriches Θ_0 by averaging the estimations of the algorithms on the last session. We extend the setting of strong convexity on $\mathbb{R}^d$ of the preceding results of Steinhardt et al. [2014], Gaillard and Wintenberger [2017], Agarwal et al. [2012] to the weaker Łojasiewicz's assumption (A3) on the ℓ_1-ball only. The latter was introduced by Łojasiewicz [1963, 1993] and states that there exist $\beta > 0$ and $\mu > 0$ such that for all $\theta \in \mathcal{B}_1$, it exists a minimizer θ^* of the risk over $\mathcal{B}_1$ satisfying

$$\mu \big\| \theta - \theta^* \big\|_2^2 \leqslant \mathbb{E}[\ell_t(\theta) - \ell_t(\theta^*)]^\beta \,.$$

The Łojasiewicz's assumption depends on a parameter $\beta \in [0,1]$ that ranges from general convex functions ($\beta = 0$) to strongly convex functions ($\beta = 1$). Under this condition our algorithm achieves a fast rate upper-bound on the average risk of order $\tilde{\mathcal{O}}((\|\theta^*\|_0 \ln(d)/T)^{1/(2-\beta)})$ when the optimal parameters have ℓ_1-norm bounded by $c < 1$. When some optimal parameters θ^* lie on the border of the ball, the bound suffers an additional factor $\|\theta^*\|_0$. Łojasiewicz's Assumption (A3) also allows multiple optima which is crucial when we are dealing with degenerated collinear design (allowing zero eigenvalues in the Gram matrix). The complexity of the algorithm, called SABOA, is $\tilde{\mathcal{O}}(dT)$ and it is fully adaptive to all parameters except for the Lipschitz constant.

To summarize our contributions, we provide
- the first hight-probability quantile bound achieving a fast rate (Theorem 2.1);
- a new bound on $R_T(\theta)$ that is small whenever θ is close to a grid provided in hindsight (Thm. 3.2);
- two efficient algorithms with improved average risks when θ is sparse in the adversarial setting with strongly convex losses (BOA+, Thm. 3.3) and in the i.i.d. setting with Łojasiewicz's assumption (SABOA, Thm. 3.4).

2. Finite reference set

In this section, we focus on finite reference set $\Theta := \{\theta_1, \dots, \theta_K\} \subset \mathcal{B}_1$. This is the case of the setting of prediction with expert advice presented in Section 2.2. We will consider the following two assumptions on the loss:

(A1) *Lipschitz loss*[2]: $\nabla \ell_t$ are sub-differential and for all $t \geqslant 1$, $\max_{\theta \in \mathcal{B}_1} \big\| \nabla \ell_t \big\|_\infty \leqslant G$.

(A2) *Weak exp-concavity:* There exist $\alpha > 0$ and $\beta \in [0,1]$ such that for all $t \geqslant 1$, for all $\theta_1, \theta_2 \in \mathcal{B}_1$, almost surely

$$\mathbb{E}_{t-1}\big[\ell_t(\theta_1) - \ell_t(\theta_2)\big] \leqslant \mathbb{E}_{t-1}\big[\nabla \ell_t(\theta_1)^\top (\theta_1 - \theta_2)\big] - \mathbb{E}_{t-1}\Big[\Big(\alpha\big(\nabla \ell_t(\theta_1)^\top (\theta_1 - \theta_2)\big)^2\Big)^{1/\beta}\Big].$$

For convex losses (ℓ_t), Assumption (A2) is satisfied with $\beta = 0$ and $\alpha < G^{-2}$. Fast rates are obtained for $\beta > 0$. It is worth pointing out that Assumption (A2) is weak even in the strongest case $\beta = 1$. It is implied by several common assumptions such as:

– *Strong convexity of the risk*: under the boundedness of the gradients, assumption (A2) with $\alpha = \mu/(2G^2)$ is implied by the μ-strong convexity of the risks $(\mathbb{E}_{t-1}[\ell_t])$.

– *Exp-concavity of the loss*: Lemma 4.2, Hazan [2016] states that (A2) with $\alpha \leqslant \frac{1}{4}\min\{\frac{1}{8G}, \kappa\}$ is implied by κ-exp-concavity of the loss functions (ℓ_t). Our assumption is slightly weaker since its needs to hold in conditional expectation only.

2.1. Fast-rate quantile bound with high probability. For prediction with $K \geqslant 1$ expert advice, Wintenberger [2014] showed that a fast rate $\mathcal{O}\big((\ln K)/T\big)$ can be obtained by the BOA algorithm under the LIST condition (i.e., Lipschitz and strongly convex losses) and i.i.d. estimators. Here, we show that Assumption (A2) is enough. By using the Squint algorithm of Koolen and Van Erven [2015] (see Algorithm 1), we also replace the dependency on the total number of experts with a quantile bound. The latter is smaller when many experts perform well. Note that Algorithm 1 uses Squint with a discrete prior over a finite set of learning rates. It corresponds to BOA of Wintenberger [2014], where each expert is replicated multiple times with different constant learning rates. The proof (with the exact constants) is deferred to Appendix C.1.

[2]Throughout the paper, we assume that the Lipschitz constant G in (A1) is known. It can be calibrated online with standard tricks such as the doubling trick (see Cesa-Bianchi et al. [2007] for instance) under sub-Gaussian conditions.

Algorithm 1: Squint – BOA with multiple constant learning rates assigned to each parameter

Inputs: $\Theta_0 = \{\theta_1, \dots, \theta_K\} \subset \mathcal{B}_1$, $E > 0$ and $\widehat{\pi}_0 \in \Delta_K$[3].
Initialization: For $1 \leqslant i \leqslant \ln(ET)$, define $\eta_i := (e^i E)^{-1}$
For $t = 1, \dots, T$

– predict $\widehat{\theta}_{t-1} = \sum_{k=1}^K \widehat{\pi}_{k,t-1}\theta_k$ and observe $\nabla \ell_t(\widehat{\theta}_{t-1})$,

– update component-wise for all $1 \leqslant k \leqslant K$

$$\widehat{\pi}_{k,t} = \frac{\sum_{i=1}^{\ln(ET)} \eta_i e^{\eta_i \sum_{s=1}^t (r_{k,s} - \eta_i r_{k,s}^2)} \pi_{k,0}}{\sum_{i'=1}^{\ln(ET)} \mathbb{E}_{j\sim\widehat{\pi}_0}\left[\eta_{i'} e^{\eta_{i'} \sum_{s=1}^t (r_{j,s} - \eta_{i'} r_{j,s}^2)}\right]}\,, \text{ where } r_{k,s} = \nabla \ell_t(\widehat{\theta}_{s-1})^\top (\widehat{\theta}_{s-1} - \theta_k)\,.$$

.

Theorem 2.1. *Let $\Theta = \{\theta_1, \dots, \theta_K\} \subset \mathcal{B}_1$ and $x > 0$. Assume (A1) and (A2). Apply Algorithm 1 with grid $\Theta_0 = \Theta$, parameter $E = 4G/3$ and initial weight vector $\widehat{\pi}_0 \in \Delta_K$. Then, for all $T \geqslant 1$ and all $\pi \in \Delta_K$, with probability at least $1 - 2e^{-x}$*

$$\mathbb{E}_{k\sim\pi}\left[R_T(\theta_k)\right] \lesssim \left(\frac{\mathcal{K}(\pi, \widehat{\pi}_0) + \ln\ln(GT) + x}{\alpha T}\right)^{\frac{1}{2-\beta}}\,,$$

where $\mathcal{K}(\pi, \widehat{\pi}_0) := \sum_{k=1}^K \pi_k \ln(\pi_k/\widehat{\pi}_{k,0})$ is the Kullback-Leibler divergence.

A fast rate of this type (without quantiles property) can be obtained in expectation by using the exponential weight algorithm (Hedge) for exp-concave loss functions. However, Theorem 2.1 is stronger. First, Assumption (A2) only needs to hold on the risks $\mathbb{E}_{t-1}[\ell_t]$, which is much weaker than exp-concavity of the losses ℓ_t. It can hold for absolute loss or quantile regression under regularity conditions. Second, the algorithm uses the so-called gradient trick. Therefore, simultaneously with upper-bounding the average risk $\mathcal{O}(T^{-1/(2-\beta)})$ with respect to the experts (θ_k), the algorithm achieves the slow rate $\mathcal{O}(1/\sqrt{T})$ with respect to any convex combination (similarly to EG). Finally, we recall that our result holds with high-probability, which is not the case for Hedge (see Audibert [2008]).

If the algorithm is run with a uniform prior $\widehat{\pi}_0 = (1/K, \dots, 1/K)$, Theorem 2.1 implies that for any subset $\Theta' \subseteq \Theta$, with high probability

$$\max_{\theta\in\Theta'} R_T(\theta) \lesssim \left(\frac{\ln(K/\operatorname{Card}(\Theta')) + \ln\ln(GT)}{\alpha T}\right)^{\frac{1}{2-\beta}}\,.$$

One only pays the proportion of good experts $\ln(K/\operatorname{Card}(\Theta'))$ instead of the total number of experts $\ln(K)$. This is the advantage of quantile bounds. We refer to Koolen and Van Erven [2015] for more details, who obtained a similar result for the regret (not the average risk). Such quantile bounds on the risk were studied by Mehta [2016, Section 7] in a batch i.i.d. setting (i.e., ℓ_t are i.i.d.). A standard online to batch conversion of our results shows that in this case, Theorem 2.1 yields with high probability for any $\pi \in \Delta_K$

$$\mathbb{E}_T\Big[\ell_{T+1}(\bar{\theta}_T) - \mathbb{E}_{k\sim\pi}\big[\ell_{T+1}(\theta_k)\big]\Big] \lesssim \left(\frac{\mathcal{K}(\pi,\widehat{\pi}_0) + \ln\ln(GT) + x}{\alpha T}\right)^{\frac{1}{2-\beta}} \quad \text{where} \quad \bar{\theta}_T = (1/T)\sum_{t=1}^T \widehat{\theta}_{t-1}\,.$$

This improves the bound obtained by Mehta [2016] who suffers the additional gap

$$(e-1)\ \mathbb{E}_T\big[\mathbb{E}_{k\sim\pi}[\ell_{T+1}(\theta_k)] - \min_{\pi^*\in\Delta_K} \ell_{T+1}(\mathbb{E}_{j\sim\pi^*}[\theta_j])\big]\,.$$

2.2. Prediction with expert advice. The framework of prediction with expert advice is widely considered in the literature (see Cesa-Bianchi and Lugosi [2006] for an overview). We recall now this setting and how it can be included in our framework. At the beginning of each round t, a finite set of $K \geqslant 1$ experts forms predictions $\boldsymbol{f}_t = (f_{1,t}, \dots, f_{K,t}) \in [0,1]^K$ that are included into the history $\mathcal{F}_{t-1}$. The learner then chooses a weight vector $\widehat{\theta}_{t-1}$ in the simplex $\Delta_K := \{\theta \in \mathbb{R}_+^K : \|\theta\|_1 = 1\}$ and produces a prediction $\widehat{f}_t := \widehat{\theta}_{t-1}^\top \boldsymbol{f}_t \in \mathbb{R}$ as a linear combination of the experts. Its performance at time t is evaluated thanks to a loss function[4] $g_t : \mathbb{R} \to \mathbb{R}$. The goal of the learner is to approach

[3]Throughout the paper, we denote the simplex of dimension $K \geq 1$ as $\Delta_K = \{\theta \in [0,\infty)^k;\ \|\theta\|_1 = 1, \|\theta\|_0 = 1\}$.
[4]For instance, g_t can be the square loss with respect to some observation $y \mapsto (y - y_t)^2$.

the performance of the best expert on a long run. This can be done by minimizing the average risk $R_{k,T} := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{t-1}[g_t(\widehat{f}_t)] - \mathbb{E}_{t-1}[g_t(f_{k,t})]$, with respect to all experts $k \in \{1,\dots,K\}$.

This setting reduces to our framework with dimension $d = K$. Indeed, it suffices to choose the K-dimensional loss function $\ell_t : \theta \mapsto g_t(\theta^\top \boldsymbol{f}_t)$ and the canonical basis $\Theta := \{\theta \in \mathbb{R}_+^K : \|\theta\|_1 = 1, \|\theta\|_0 = 1\}$ in $\mathbb{R}^K$ as the reference set. Denoting by θ_k the k-th element of the canonical basis, we see that $\theta_k^\top \boldsymbol{f}_t = f_{k,t}$, so that $\ell_t(\theta_k) = g_t(f_{k,t})$. Therefore, $R_{k,T}$ matches our definition of $R_T(\theta_k)$ in Equation (1) and we get under the assumptions of Theorem 2.1 a bound of order:

$$\mathbb{E}_{k\sim\pi}\big[R_{k,T}\big] \lesssim \Big(\tfrac{\mathcal{K}(\pi,\widehat{\pi}_0)+\ln\ln(GT)+x}{\alpha T}\Big)^{\frac{1}{2-\beta}}.$$

It is worth to point out that though the parameters θ_k of the reference set are constant, this method can be used to compare the player with arbitrary strategies $f_{k,t}$ that may evolve over time and depend on recent data. This is why we do not want to assume here that there is a single fixed expert $k^* \in \{1,\dots,K\}$ which is always the best, i.e., $\mathbb{E}_{t-1}[g_t(f_{k^*,t})] \leqslant \min_k \mathbb{E}_{t-1}[g_t(f_{k,t})]$. Hence, we cannot replace (A2) with the closely related Bernstein assumption (see Ass. (A2') or [Koolen et al., 2016, Cond. 1]).

In this setting, Assumption (A2) can be reformulated on the one dimensional loss functions g_t as follows: there exist $\alpha > 0$ and $\beta \in [0,1]$ such that for all $t \geqslant 1$, for all $0 \leqslant f_1, f_2 \leqslant 1$,

$$\mathbb{E}_{t-1}[g_t(f_1) - g_t(f_2)] \leqslant \mathbb{E}_{t-1}\big[g_t'(f_1)(f_1-f_2)\big] - \mathbb{E}_{t-1}\bigg[\Big(\alpha\big(g_t'(f_1)(f_1-f_2)\big)^2\Big)^{1/\beta}\bigg], \quad a.s.$$

It holds with $\alpha = \kappa/(2G^2)$ for κ-strongly convex risk $\mathbb{E}_{t-1}[g_t]$. For instance, the square loss $g_t = (\,\cdot - y_t\,)^2$ satisfies it with $\beta = 1$ and $\alpha = 1/8$.

3. Online optimization in the unit ℓ_1-ball

The aim of this section is to extend the preceding results to the reference set $\Theta = \mathcal{B}_1$ instead of finite $\Theta = \{\theta_1,\dots,\theta_K\}$. A classical reduction from the expert advice setting to the ℓ_1-ball is the so-called "gradient-trick". A direct analysis on BOA applied to the 2d corners of the ℓ_1-ball suffers a slow rate $\mathcal{O}(1/\sqrt{T})$ on the average risk. The goal is to exhibit algorithms that go beyond $\mathcal{O}(1/\sqrt{T})$. In view of the fast rate in Theorem 2.1 the program is clear; in order to accelerate BOA, one has to add in the grid of experts some points of the ℓ_1-ball to the 2d corners. In Section 3.1 one investigate the cases of non adaptive grids that are optimal but yields unfeasible (NP) algorithm. In Section 3.2 we introduce a pseudo-metric in order to bound the regret of grids consisting of the 2d corners and some arbitrary fixed points. From this crucial step, we then derive the form of the adaptive points we have to add to the 2d corners, in the adversarial case, Section 3.3, and in the i.i.d. case, Section 3.4.

3.1. Warmup: fast rate by discretizing the space. As a warmup, we show how to use Theorem 2.1 in order to obtain fast rate on $R_T(\theta)$ for any $\theta \in \mathcal{B}_1$. Basically, if the parameter θ could be included into the grid Θ_0, Theorem 2.1 would turn into a bound on the regret $R_T(\theta)$ with respect to θ. However, this is not possible as we do not know θ in advance. A solution consists in approaching $\mathcal{B}_1$ with $\mathcal{B}_1(\varepsilon)$, a fixed finite ε-covering in ℓ_1-norm of minimal cardinal. In dimension d, it is known that $\mathrm{Card}(\mathcal{B}_1(\varepsilon)) \lesssim (1/\varepsilon)^d$. We obtain the following near optimal rate for the regret on $\mathcal{B}_1$.

Proposition 3.1. *Let $x > 0$ and $T \geqslant 1$. Under Assumptions of Theorem 2.1, applying Algorithm 1 with grid $\Theta_0 = \mathcal{B}_1(T^{-2})$ and uniform prior $\widehat{\pi}_0$ over $\Delta_{\mathrm{Card}(\mathcal{B}_1(T^{-2}))}$ satisfies for all $\theta \in \mathcal{B}_1$*

$$R_T(\theta) \lesssim \Big(\frac{d\ln T + \ln\ln(GT) + x}{\alpha T}\Big)^{\frac{1}{2-\beta}} + \frac{G}{T^2}.$$

Proof. Let $\varepsilon = 1/T^2$ and $\theta \in \mathcal{B}_1$ and $\tilde{\theta}$ be its ε-approximation in $\mathcal{B}_1(\varepsilon)$. The proof follows from Lipschitzness of the loss: $R_T(\theta) \leqslant R_T(\tilde{\theta}) + G\varepsilon$; followed by applying Theorem 2.1 on $R_T(\tilde{\theta})$. □

Following this method and inspired by the work of Rigollet and Tsybakov [2011], one can improve d to $\|\theta\|_0 \ln d$ by carefully choosing the prior $\widehat{\pi}_0$; see Appendix A for details. The obtained rate is optimal up to log-factors. However, the complexity of the discretization is prohibitive (of order T^d) and non realistic for practical purpose.

3.2. Regret bound for arbitrary fixed discretization grid. Let $\Theta_0 \subset \mathcal{B}_1$ of finite size. The aim of this Section is to study the regret of Algorithm 1 with respect to any $\theta \in \mathcal{B}_1$ when applied with the grid Θ_0. Similarly to Proposition 3.1, the average risk may be bounded as

$$R_T(\theta) \lesssim \left(\frac{\ln \mathrm{Card}(\Theta_0)+\ln\ln T+x}{\alpha T}\right)^{\frac{1}{2-\beta}} + G\|\theta' - \theta\|_1\,, \tag{2}$$

for any $\theta' \in \Theta_0$. We say that a regret bound is *accelerable* if it provides a fast rate except a term depending on the distance with the grid (i.e., the term in $\|\theta' - \theta\|_1$ in (2)) which vanishes to zero. This property will be crucial in obtaining fast rates by enriching the grid Θ_0. Hence the regret bound (2) is not accelerable due to the second term that is constant. In order to find an accelerable regret bound, we introduce the notion of *averaging accelerability*, a pseudo-metric that replaces the ℓ_1-norm in (2). We define it now formally but we will give its intuition in the sketch of proof of Theorem 3.2.

Definition 3.1 (averaging accelerability)**.** *For any $\theta, \theta' \in \mathcal{B}_1$, we define*

$$D(\theta, \theta') := \min\left\{0 \leqslant \pi \leqslant 1 : \|\theta - (1-\pi)\theta'\|_1 \leqslant \pi\right\}.$$

This averaging accelerability has several nice properties. In Appendix B, we provide a few concrete upper-bounds in terms of classical distances. For instance, Lemma B.1 provides the upper-bound $D(\theta, \theta') \leqslant \|\theta - \theta'\|_1/(1 - \|\theta'\|_1 \wedge \|\theta\|_1)$. We are now ready to state our regret bound, when Algorithm 1 is applied with an arbitrary approximation grid Θ_0.

Theorem 3.2. *Let $x > 0$. Let $\Theta_0 \subset \mathcal{B}_1$ of finite size such that $\{\theta : \|\theta\|_1 = 1, \|\theta\|_0 = 1\} \subseteq \Theta_0$. Let Assumption (A1) and (A2) be satisfied. Then, Algorithm 1 applied with Θ_0, uniform weight vector $\widehat{\pi}_0$ over the elements of Θ_0 and $E = 8G/3$, satisfies with probability $1 - e^{-x}$,*

$$R_T(\theta) \lesssim \left(\frac{a}{\alpha T}\right)^{\frac{1}{2-\beta}} + GD(\theta, \Theta_0)\sqrt{\frac{a}{T}} + \frac{aG}{T}\,,$$

for all $\theta \in \mathcal{B}_1$, where $a = \ln \mathrm{Card}(\Theta_0) + \ln\ln(GT) + x$ and $D(\theta, \Theta_0) := \min_{\theta' \in \Theta_0} D(\theta, \theta')$.

Sketch of proof. The complete proof can be found in Appendix C.2 but we give here the high-level idea of the proof. Let θ be the unknown parameter the algorithm will be compared with. Let $\theta' \in \Theta_0$ a point in the grid Θ_0 minimizing $D(\theta, \theta')$. Then one can decompose $\theta = (1-\varepsilon)\theta' + \varepsilon\theta''$ for a unique point $\|\theta''\|_1 = 1$ and $\varepsilon := D(\theta, \theta')$. See Appendix C.2 for details. In the analysis, the regret bound with respect to θ can be decomposed into two terms:

- The first one quantifies the cost of picking $\theta' \in \Theta_0$, bounded using Theorem 2.1;
- The second one is the cost of learning $\theta'' \in \mathcal{B}_1$ rescaled by ε. Using a classical slow-rate bound in $\mathcal{B}_1$, it is of order $\mathcal{O}(1/\sqrt{T})$.

The average risk $\mathrm{Reg}(\theta)$ is thus of the order

$$(1-\varepsilon)\underbrace{\mathrm{Reg}(\theta')}_{\text{Thm 2.1}} + \varepsilon\underbrace{\mathrm{Reg}(\theta'')}_{G\sqrt{\ln(\mathrm{Card}\,\Theta_0))/T}} \lesssim \left(\frac{\ln \mathrm{Card}(\Theta_0) + \ln\ln(GT) + x}{\alpha T}\right)^{\frac{1}{2-\beta}} + \varepsilon G\sqrt{\frac{\ln \mathrm{Card}(\Theta_0)}{T}}\,. \quad \square$$

Note that the bound of Theorem 3.2 is *accelerable* as it vanishes to zero on the contrary to Inequality (2). Theorem 3.2 provides an upper-bound which may improve the rate $\mathcal{O}(1/\sqrt{T})$ if the distance $D(\theta, \Theta_0)$ is small enough. By using the properties of the averaging accelerability (see Lemma B.1 in Appendix B), Theorem 3.2 provides some interesting properties of the rate in terms of ℓ_1 distance. By including 0 into our approximation grid Θ_0, we get a an oracle-bound of order $\mathcal{O}(\|\theta\|_1/\sqrt{T})$ for any $\theta \in \mathcal{B}_1$. Furthermore, it also yields for any $\|\theta\|_1 \leqslant 1 - \gamma < 1$, a bound of order $R_T(\theta) \leqslant \mathcal{O}\big(\|\theta - \theta_k\|_1/(\gamma\sqrt{T})\big)$ for all $\theta_k \in \Theta_0$.

It is also interesting to notice that the bound on the gradient G can be substituted with the averaged gradient observed by the algorithm. This allows to replace G with the level of the noise in certain situations with vanishing gradients (see for instance Theorem 3 of Gaillard and Wintenberger [2017]).

3.3. Fast-rate sparsity regret bound under adversarial data. In this section, we focus on the adversarial case where $\ell_t = \mathbb{E}_{t-1}[\ell_t]$ are μ-strongly convex deterministic functions. In this case, Assumption (A2) is satisfied with $\beta = 1$ and $\alpha = \mu/(2G^2)$. Our algorithm, called BOA+, is defined as follows. For $i \geqslant 0$, it predicts from time step $t_i = 2^i$ to $t_{i+1} - 1$, by restarting Algorithm 1 with uniform prior, parameter $E = 4G/3$ and updated discretization grid Θ_0 indexed by i:

$$\Theta^{(i)} = \{[\theta_i^*]_k, k = 1, \dots, d\} \cup \{\theta : \|\theta\|_1 = 2, \|\theta\|_0 = 1\},$$

where $\theta_i^* \in \arg\min_{\theta \in \mathcal{B}_1} \sum_{t=1}^{t_i - 1} \ell_t(\theta)$ is the empirical risk minimizer (or the leader) until time $t_i - 1$. The notation $[\cdot]_k$ denotes the hard-truncation with k non-zero values. Remark that θ_i^* for $i = 1, 2, \dots, \ln_2(T)$ can be efficiently computed approximatively as the solution of a strongly convex optimization problem.

Theorem 3.3. *Assume the losses are μ-strongly convex on $\mathcal{B}_2 := \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leqslant 2\}$ with gradients bounded by G in ℓ_∞-norm. The average regret of BOA+ is upper-bounded for all $\theta \in \mathcal{B}_1$ as:*

$$R_T(\theta) \leqslant \tilde{\mathcal{O}}\left(\min\left\{G\sqrt{\frac{\ln d}{T}}, \sqrt{\frac{\|\theta\|_0}{\mu}}\left(G\sqrt{\frac{\ln d}{T}}\right)^{\frac{3}{2}}, \frac{\sqrt{\|\theta\|_0 d} G^2 \ln d}{\mu T}\right\}\right).$$

The proof is deferred to the appendix. It is worth to notice that the bound can be rewritten as follows:

$$R_T(\theta) \leqslant \tilde{\mathcal{O}}\left(\min\left\{G\sqrt{\frac{\ln d}{T}}, \frac{\|\theta\|_0 G^2 \ln d}{\mu T}\right\} \min\left\{G\sqrt{\frac{\ln d}{T}}, \frac{dG^2 \ln d}{\mu T}\right\}\right)^{1/2}.$$

It provides an intermediate rate between known optimal rates without sparsity $\mathcal{O}(\sqrt{\ln d/T})$ and $\tilde{\mathcal{O}}(d/T)$ and known optimal rates with sparsity $\mathcal{O}(\sqrt{\ln d/T})$ and $\tilde{\mathcal{O}}(\|\theta\|_0/T)$ but with non-efficient procedures only. If all θ_i^* are approximatively d_0-sparse it is possible to achieve a rate of order $\tilde{\mathcal{O}}(d_0/T)$, for any $\|\theta\|_0 \leqslant d_0$. This can be achieved in particular in the i.i.d. setting (see next section). However, we leave for future work whether it is possible to achieve it in full generality and efficiently in the adversarial setting.

Remark 3.1. The strongly convex assumption on the losses can be relaxed by only assuming Inequality (30): it exists $\mu > 0$ and $\beta \in [0, 1]$ such that for all $t \geqslant 1$ and $\theta \in \mathcal{B}_1$

$$\mu\|\theta - \theta_t^*\|_2^2 \leqslant \left(\tfrac{1}{t}\sum_{s=1}^t \ell_s(\theta) - \ell_s(\theta_t^*)\right)^\beta, \quad \text{where} \quad \theta_t^* \in \arg\min_{\theta \in \mathcal{B}_1} \sum_{s=1}^t \ell_s(\theta). \tag{3}$$

The rates will then depend on β as it was the case in Theorem 2.1. A specific interesting case is when $\|\theta_t^*\|_1 = 1$. Then θ_t^* is very likely to be sparse. Denote S_t^* its support. Assumption (3) can be weakened in this case. Indeed any $\theta \in \mathcal{B}_1$ satisfies $\|\theta\|_1 \leqslant \|\theta_t^*\|_1$, which from Lemma 6 of Agarwal et al. [2012] yields $\|\theta - \theta_t^*\|_1 \leqslant 2\|[\theta - \theta_t^*]_{S_t^*}\|_1$ where $[\theta]_S = (\theta_i \mathbb{1}_{i \in S})_{1 \leqslant i \leqslant d}$. One can thus restrict Assumption (3) to hold on the support of θ_t^* only. Such restricted conditions for $\beta = 1$ are common in the sparse learning literature and essentially necessary to hold for the existence of efficient and optimal sparse procedures, see Zhang et al. [2014]. In the online setting, the restricted condition (3) with $\beta = 1$ should hold at any time $t \geq 1$, which is unlikely.

3.4. Fast-rate sparsity risk bound under i.i.d. data. In this section, we provide an algorithm with fast-rate sparsity risk-bound on $\mathcal{B}_1$ under i.i.d. data. This is obtained by regularly restarting Algorithm 1 with an updated discretization grid Θ_0 approaching the set of minimizers $\Theta^* := \arg\min_{\theta \in \mathcal{B}_1} \mathbb{E}[\ell_t(\theta)]$.

In this setting, a close inspection of the proof of Theorem 3.4 shows that we can replace Assumption (A2) with the Bernstein condition: it exists $\alpha' > 0$ and $\beta \in [0, 1]$, such that for all $\theta \in \mathcal{B}_1$, all $\theta^* \in \Theta^*$ and all $t \geqslant 1$,

$$\alpha' \mathbb{E}\Big[\big(\nabla \ell_t(\theta)^\top (\theta - \theta^*)\big)^2\Big] \leqslant \mathbb{E}\Big[\nabla \ell_t(\theta)^\top (\theta - \theta^*)\Big]^\beta. \tag{A2'}$$

This fast-rate type stochastic condition is equivalent to the *central condition* (see [Van Erven et al., 2015, Condition 5.2]) and was already considered to obtain faster rates of convergence for the regret (see [Koolen et al., 2016, Condition 1]).

Algorithm 2: SABOA – Sparse Acceleration of BOA

Parameters: $E > 0$

Initialization: $t_i = 2^i$ for $i \geqslant 0$,

For each $i = 0, \dots$

- define $\bar{\theta}^{(i-1)} := 0$ if $i = 0$ and $\bar{\theta}^{(i-1)} := 2^{-i+1} \sum_{t=t_{i-1}}^{t_i - 1} \widehat{\theta}_{t-1}$ otherwise.
- Define $\Theta^{(i)}$ a set of hard-truncated and dilated soft-thresholded versions of $\bar{\theta}^{(i-1)}$ as in (42);
- Denote $K_i := \mathrm{Card}(\Theta^{(i)}) + 2d \leqslant (i+1)(1 + \ln d) + 3d$;
- At time step t_i, restart Algorithm 1 in Δ_{K_i} with parameters $\Theta_0 := \Theta^{(i)} \cup \{\theta : \|\theta\|_1 = 1, \|\theta\|_0 = 1\}$ (denote by $\theta_1, \dots, \theta_{K_i}$ its elements), $E > 0$ and uniform prior $\widehat{\pi}_0$ over Δ_{K_i}. In other words, for time steps $t = t_i, \dots, t_{i+1} - 1$:
  - predict $\widehat{\theta}_{t-1} = \sum_{k=1}^{K_i} \widehat{\pi}_{k,t-1} \theta_k$ and observe $\nabla \ell_t(\widehat{\theta}_{t-1})$
  - define component-wise for all $1 \leqslant k \leqslant K_i$

$$\widehat{\pi}_{k,t} = \frac{\sum_{i=1}^{\ln(ET^2)} \eta_{k,i} e^{\eta_{k,i} \sum_{s=1}^{t} (r_{k,s} - \eta_{k,i} r_{k,s}^2)} \pi_{k,0}}{\sum_{i=1}^{\ln(ET^2)} \mathbb{E}_{\pi_0}\left[\eta_{j,i} e^{\eta_{j,i} \sum_{s=1}^{t} (r_{j,s} - \eta_{j,i} r_{j,s}^2)}\right]} ,$$

  where $r_{k,s} = \nabla \ell_t(\widehat{\theta}_{s-1})^\top (\widehat{\theta}_{s-1} - \theta_k)$.

The Łojasiewicz's assumption. In order to obtain sparse oracle inequalities we also need the Łojasiewicz's Assumption (A3) which is a relaxed version of strong convexity of the risk.

(A3) *Łojasiewicz's inequality:* $(\ell_t)_{t \geqslant 1}$ is i.i.d. and it exists $\beta \in [0,1]$ and $0 < \mu \leqslant 1$ such that, for all $\theta \in \mathbb{R}^d$ with $\|\theta\|_1 \leqslant 1$, it exists $\theta^* \in \Theta^* \subseteq \mathcal{B}_1$ satisfying

$$\mu \left\|\theta - \theta^*\right\|_2^2 \leqslant \mathbb{E}[\ell_t(\theta) - \ell_t(\theta^*)]^\beta .$$

This assumption is fairly mild. It is indeed satisfied with $\beta = 0$ and $\mu = 1$ as soon as the loss is convex. For $\beta = 1$, this assumption is implied by the strong convexity of the risk $\mathbb{E}[\ell_t]$. One should mention that our framework is more general than this classical case because

- multiple optima are allowed, which seems to be new when combined with sparsity bounds;
- on the contrary to Steinhardt et al. [2014] or Gaillard and Wintenberger [2017], our framework does not compete with the minimizer θ^* over $\mathbb{R}^d$ with a known upper-bound on the ℓ_1-norm $\|\theta^*\|_1$. We consider the minimizer over the ℓ_1-ball $\mathcal{B}_1$ only. The latter is more likely to be sparse and Assumption (A3) only needs to hold over $\mathcal{B}_1$.

Assumption (A2) (or (A2')) and (A3) are strongly related. Assumption (A3) is more restrictive because it is heavily design dependent. In linear regression for instance, the constant μ corresponds to the smallest non-zero eigenvalue of the Gram matrix while $\alpha = 1/G^2$. If $\Theta^* = \{\theta^*\}$ is a singleton than Assumption (A3) implies Assumption (A2') with $\alpha' \geqslant \mu/G^2$.

Algorithm and risk bound. Our new procedure is described in Algorithm 2. It is based on the following fact: the bound of Theorem 3.2 is small if one of the estimators in Θ_0 is close to Θ^*. Thus, our algorithm regularly restarts BOA by adding current estimators of Θ^* into an updated grid Θ_0. The estimators are built by averaging past iterates $\widehat{\theta}_{t-1}$ and truncated to be sparse and ensure small ℓ_1-distance. Remark that restart schemes under Łojasiewicz's Assumption is natural and was already used for instance in optimization by Roulet and d'Aspremont [2017]. A stochastic version of the algorithm (sampling randomly a subset of gradient coordinates at each time step) can be implemented as in the experiments of Duchi et al. [2008]. We get the following upper-bound on the average risk. The proof, that computes the exact constants, is postponed to Appendix C.7.

Theorem 3.4. *Let $x > 0$, $\gamma \geqslant 0$. Under Assumptions (A1-3), if $\Theta^* \subseteq \mathcal{B}_{1-\gamma}$, $E = 4/3G \geqslant 1$, Algorithm 2 satisfies with probability at least $1 - e^{-x}$ the bound on the average risk*

$$R_T(\theta^*) \lesssim \left(\frac{\ln d + \ln\ln(GT) + x}{T} \left(\frac{1}{\alpha} + \frac{G^2}{\mu} \left(d_0^2 \wedge \frac{d_0}{\gamma^2} \right) \right) \right)^{\frac{1}{2-\beta}} ,$$

where $d_0 = \max_{\theta^* \in \Theta^*} \|\theta^*\|_0$.

Let us conclude with some important remarks about Theorem 3.4. First, it is worth pointing out that SABOA does not need to know the parameters δ, β, α, μ and d_0 to fulfill the rate of Theorem 3.4.

Approximately sparse optima. Our results can be extended to a unique approximately sparse optimum θ^*. We get $R_T(\theta) \leqslant (1+o(1))\|\theta-\theta^*\|_1 + \tilde{\mathcal{O}}((\|\theta\|_0^2/T)^{1/(2-\beta)})$ for any $\theta \in \mathcal{B}_1$; see Agarwal et al. [2012], Bunea et al. [2007].

On the radius of L1 ball. We only performed the analysis into $\mathcal{B}_1$, the ℓ_1-ball of radius 1. However, one might need to compare with parameters into $\mathcal{B}_1(U)$ the ℓ_1-ball of radius $U > 0$. This can be done by simply rescaling the losses and applying our results to the loss functions $\theta \in \mathcal{B}_1 \mapsto \ell_t(U\theta)$ instead of ℓ_t. If θ^* lies on the border of the ℓ_1-ball, we could not avoid a factor $\|\theta^*\|_0^2$. In that situation, our algorithm needs to recover the support of θ^* without the Irreprensatibility Condition [Wainwright, 2009] (see configuration 3 of Figure 1). In this case, we can actually relax Assumption (A3) to hold in sup-norm.

Conclusion

In this paper, we show that BOA is an optimal online algorithm for aggregating predictors under very weak conditions on the loss. Then we aggregate sparse versions of the leader (BOA+) or of the averaging of BOA's weights (SABOA) in the adversarial or in the i.i.d. setting, respectively. Aggregating both achieves sparse fast-rates of convergence in any case. These rates are deteriorated compared with the optimal one that require restrictive assumption. Our weaker conditions are very sensitive to the radius of the ℓ_1-ball we consider. The optimal choice of the radius, if it is not imposed by the application, is left for future research.

References

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. In *Advances in Neural Information Processing Systems 25*, pages 1538–1546. Curran Associates, Inc., 2012.

[2] J.-Y. Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems*, pages 41–48, 2008.

[3] F. Bunea, A. Tsybakov, M. Wegkamp, et al. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[4] O. Catoni. Universal aggregation rules with exact bias bounds. *preprint*, 510, 1999.

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

[6] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

[7] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM, 2008.

[8] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, pages 14–26, 2010.

[9] D. Foster, S. Kale, and H. Karloff. Online sparse linear regression. In *Conference on Learning Theory*, pages 960–970, 2016.

[10] D. J. Foster, S. Kale, M. Mohri, and K. Sridharan. Parameter-free online learning via model selection. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6020–6030. Curran Associates, Inc., 2017.

[11] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[12] P. Gaillard and O. Wintenberger. Sparse Accelerated Exponential Weights. In *20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Apr. 2017.

[13] S. Gerchinovitz. *Prediction of individual sequences and prediction in the statistical framework: some links around sparse regression and aggregation techniques*. PhD thesis, Université Paris-Sud 11, Orsay, 2011.

[14] E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.
[15] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.
[16] S. Kale, Z. Karnin, T. Liang, and D. Pál. Adaptive feature selection: Computationally efficient online sparse linear regression under rip. *arXiv preprint arXiv:1706.04690*, 2017.
[17] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.
[18] W. M. Koolen and T. Van Erven. Second-order quantile methods for experts and combinatorial games. In *COLT*, volume 40, pages 1155–1175, 2015.
[19] W. M. Koolen, P. Grünwald, and T. van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. In *Advances in Neural Information Processing Systems*, pages 4457–4465, 2016.
[20] J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10(Mar):777–801, 2009.
[21] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.
[22] N. A. Mehta. Fast rates with high probability in exp-concave statistical learning. *arXiv preprint arXiv:1605.01288*, 2016.
[23] A. Rakhlin and K. Sridharan. Online nonparametric regression with general loss functions. *arXiv preprint arXiv:1501.06598*, 2015.
[24] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, pages 731–771, 2011.
[25] V. Roulet and A. d'Aspremont. Sharpness, restart and acceleration. In *Advances in Neural Information Processing Systems*, pages 1119–1129, 2017.
[26] J. Steinhardt, S. Wager, and P. Liang. The statistics of streaming sparse regression. *arXiv preprint arXiv:1412.4182*, 2014.
[27] T. Van Erven, P. D. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.
[28] V. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.
[29] V. G. Vovk. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990.
[30] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55 (5):2183–2202, 2009.
[31] O. Wintenberger. Optimal learning with bernstein online aggregation. Extended version available at arXiv:1404.1356 [stat. ML], 2014.
[32] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.
[33] Y. Yang. Combining forecasting procedures: some theoretical results. *Econometric Theory*, 20 (01):176–222, 2004.
[34] Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948, 2014.
[35] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning, ICML 2003*, 2003.
[36] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, pages 87–89, 1963.
[37] S. Łojasiewicz. Sur la géométrie semi-et sous-analytique. *Annales de l'institut Fourier*, 43(5): 1575–1595, 1993.

Appendix A. Sparse oracle inequality by discretizing the space

Inspired by the work of [24], one can improve d to $\|\theta\|_0 \ln d$ in Proposition 3.1 by carefully choosing the prior $\widehat{\pi}_0$. To do so, we cover $\mathcal{B}_1$ by the subspaces

$$\mathcal{B}_1^\tau := \Big\{\theta \in \mathcal{B}_1 : \forall i \quad \tau_i = 0 \Rightarrow \theta_i = 0\Big\},$$

where $\tau \in \{0,1\}^d$ denotes a sparsity pattern which determines the non-zero components of $\theta \in \mathcal{B}_1^\tau$. For each sparsity pattern $\tau \in \{0,1\}^d$, the subspace $\mathcal{B}_1^\tau$ can be approximated in ℓ_1-norm by an ε-cover $\mathcal{B}_1^\tau(\varepsilon)$ of size $\varepsilon^{-\|\tau\|_0}$. In order to obtain the optimal rate of convergence, we apply Algorithm 1 with $\Theta_0 = \cup_{\tau\in\{0,1\}^d}\mathcal{B}_1^\tau(\varepsilon)$ with a non-uniform prior $\widehat{\pi}_0$. The latter penalizes non-sparse τ to reflect their respective complexities. We assign to any $\theta \in \mathcal{B}_1^\tau(\varepsilon)$ the prior, depending on $\tau \in \{0,1\}^d$,

$$\widehat{\pi}_{\tau,0} = \left(\#\mathcal{B}_1^\tau(\varepsilon)(d+1)\binom{d}{d_0}\right)^{-1} \approx \frac{\varepsilon^{d_0}}{(d+1)\binom{d}{d_0}} \qquad \text{where} \quad d_0 = \|\tau\|_0\,.$$

Note that the sum $\widehat{\pi}_{\tau,0}$ over $\theta \in \mathcal{B}_1^\tau(\varepsilon)$ and $\tau \in \{0,1\}^d$ is one. Therefore, Theorem 2.1 yields

$$R_T(\theta) \lesssim \left(\frac{\|\theta\|_0 \ln(dT/\|\theta\|_0) + x}{\alpha T}\right)^{\frac{1}{2-\beta}} + \frac{\|\theta\|_0 G}{T^2}\,, \tag{4}$$

by noting that $\binom{d}{\|\theta\|_0} \leqslant d^{\|\theta\|_0}$ and choosing $\varepsilon = \|\theta\|_0/T^2$. Similar optimal oracle inequalities for mixing arbitrary regressions functions are obtained by Yang [33] and Catoni [4].

Appendix B. Properties of the averaging accelerability

In this appendix, we give a geometric interpretation of the *averaging accelerability* defined in Definition (3.1). We also provide several properties in terms of classical distances.

Geometric insight. Let $\theta \in \mathcal{B}_1$ be some unknown parameter and $\theta' \in \mathcal{B}_1$ a point approximating θ. Let us define $\theta'' \in \mathcal{B}_1$ the unique point satisfying

$$\|\theta''\|_1 = 1 \qquad \text{and} \qquad \theta'' = \lambda(\theta - \theta') + \theta' \tag{5}$$

for some $\lambda \geq 1$. From this definition, we immediately derive that

$$\left\|\theta - \Big(1 - \frac{1}{\lambda}\Big)\theta'\right\|_1 = \frac{\|\theta''\|_1}{\lambda} = \frac{1}{\lambda}$$

Therefore from Definition 3.1, we have $D(\theta,\theta') \leqslant \frac{1}{\lambda}$. Actually, this is an equality and we can write

$$D(\theta,\theta') = \max\Big\{\lambda \geq 1 : \|\lambda(\theta-\theta') + \theta'\|_1 \leqslant 1\Big\}^{-1}.$$

As the maximum is achieved, the averaging accelerability corresponds to the inverse of λ in the definition (5) of the extrapolation point θ''.

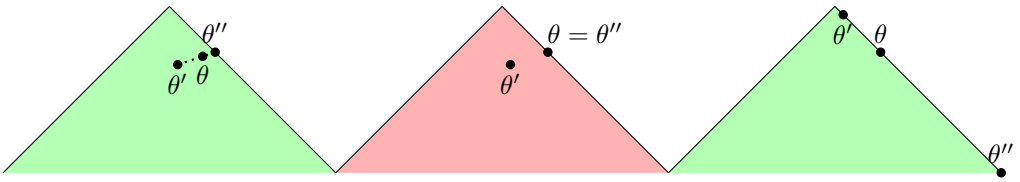

Figure 1. Averaging accelerability for 3 different configurations.

Figure 1 pictures several configurations of θ' and θ that lead to different averaging accelerability. The further θ'' is from θ, the smaller is $D(\theta,\theta')$ and the smaller is the averaging accelerability. When $D(\theta,\theta') = 1/\lambda = 1$, then $\theta = \theta''$ and our regret bound does not improve the classic slow-rate $\mathcal{O}(1/\sqrt{T})$. That typically happens when $\|\theta\|_1 = 1$, as in the second configuration in Figure 1. In this case, a possible solution is to consider a larger ball (for instance of radius 2 instead of 1). This approach was considered in [12], see Figure 1 there. Another solution is to remark that even when $\|\theta\|_1 = 1$, the procedure is still accelerable ($D(\theta,\theta') < 1$) if the approximation θ' satisfies the same

constraints than θ (see the third configuration in Figure 1 where θ'' and θ are on the same edge of the ball). We make this statement more precise in the following subsections.

B.1. The averaging accelerability in terms of classical distances. We provide in the next Lemmas a few concrete upper-bounds in terms of classical distances. The proofs are respectively postponed to Appendices C.3 to C.5. The first Lemma, states that the averaging accelerability can be upper-bounded with the ℓ_1-distance.

Lemma B.1. *We have for any $\theta, \theta' \in \mathcal{B}_1$*

$$D(\theta, \theta') \leqslant \frac{\|\theta - \theta'\|_1}{\|\theta - \theta'\|_1 + 1 - \|\theta\|_1} .$$

The Lemma above has a main drawback. The averaging accelerability does not decrease with the ℓ_1-distance if $\|\theta\|_1 = 1$. In this case, we thus need additional assumptions. The following Corollary upper-bounds the averaging accelerability in sup-norm as soon as a θ' has a support included into the one of θ. This situation is represented in the third configuration of Figure 1.

Lemma B.2. *Let $\theta, \theta' \in \mathcal{B}_1$. Assume that $\|\theta'\|_1 \geqslant \|\theta\|_1$ and $\operatorname{sign}(\theta'_i) \in \{0, \operatorname{sign}(\theta_i)\}$ for all $1 \leqslant i \leqslant d$. Then,*

$$D(\theta, \theta') \leqslant 1 - \min_{1 \leqslant i \leqslant d} \frac{|\theta_i|}{|\theta'_i|} \leqslant \frac{\|\theta - \theta'\|_\infty}{\Delta} ,$$

where $\Delta := \min_{i:\theta'_i \neq 0} |\theta_i|$.

We want to emphasis here the two very different behavior of the averaging accelerability;

- in the case $\|\theta\|_1 < 1$: the averaging accelerability is proportional to $\|\theta - \theta'\|_1$.
- in the case $\|\theta\|_1 = 1$: the averaging accelerability may be smaller than 1 and lead to improved regret guarantees under extra assumptions: $\|\theta'\|_1 = 1$ and the support of θ' is included in the one of θ. The relative gain is then proportional to $\|\theta\|_0 \|\theta - \theta'\|_\infty$.

B.2. The averaging accelerability with an approximation in sup-norm in hand. Let us focus on the second case, where the averaging accelerability is controlled under the knowledge of the support of θ. The second inequality in Lemma B.2 is interesting but yields an undesirable dependence on $\Delta := \min_{i:\theta_i \neq 0} |\theta_i|$, which can be arbitrarily small and which is at best of order $\|\theta\|_1 / \|\theta\|_0$. Moreover, the recovery of the support of θ is a well studied difficult problem, see [30]. Thanks to the following Lemma, we ensure the averaging accelerability from any ℓ_∞-approximation θ' of θ. We use a dilated soft-thresholding version of θ' as an approximation of θ. For any $\varepsilon > 0$, let us introduce S_ε the soft threshold operator so that $S_\varepsilon(x)_i = \operatorname{sign}(x_i)(|x_i| - \varepsilon)_+$ for all $1 \leqslant i \leqslant d$. The soft threshold operator is equivalent to the popular LASSO algorithm in the orthogonal design setting for the square loss. We couple the soft-thresholding with a dilatation that has the benefit of ensuring non thresholded coordinates faraway from zero. This allows to get rid of the unwanted factor $1/\Delta$ of the Lemma B.2. It is replaced with a factor $2\|\theta\|_0 / \|\theta\|_1$ which corresponds (up to the factor 2) to the best possible scenario for the value of Δ.

Lemma B.3. *Let $\theta, \theta' \in \mathcal{B}_1$ such that $\|\theta - \theta'\|_\infty \leqslant \varepsilon$ and $\|\theta\|_0 \leqslant d_0$. Then, define the dilated soft-threshold*

$$\tilde{\theta} := S_\varepsilon(\theta') \left(1 + \frac{2 d_0 \varepsilon}{\|S_\varepsilon(\theta')\|_1}\right) \wedge \frac{1}{\|S_\varepsilon(\theta')\|_1}$$

where by convention $\tilde{\theta} = 0$ when $S_\varepsilon(\theta') = 0$. Then $\tilde{\theta}$ satisfies

(i) $\|\tilde{\theta}\|_1 \geqslant \|\theta\|_1$ *if* $\tilde{\theta} \neq 0$
(ii) $\operatorname{sign}(\tilde{\theta}_i) \in \{0, \operatorname{sign}(\theta_i)\}$ *for all* $1 \leqslant i \leqslant d$
(iii) $D(\theta, \tilde{\theta}) \leqslant 2 d_0 \varepsilon / \|\theta\|_1$.

Performing this transformation requires the knowledge of the values of ε and d_0 that are not observed. However, performing an exponential grid on ε from $1/T$ to U only harms the complexity by a factor $\ln(UT)$.

Appendix C. Proofs

C.1. Proof of Theorem 2.1. Algorithm 1 is a particular case of the Bernstein Online Aggregation algorithm (BOA) with fixed learning rates of [31][5] applied on a particular set of experts $\mathcal{K}$. We make more clear the connexion thereafter. We start our proof with Theorem 3.2 of [31] that states that for any distribution $\tilde{\pi}$ over the set of experts $j \in \mathcal{K}$:

$$\sum_{t=1}^{T} \mathbb{E}_{j\sim\tilde{\pi}}[r_{j,t}] \leqslant \mathbb{E}_{j\sim\tilde{\pi}}\left[\eta_j \sum_{t=1}^{T} r_{j,t}^2 + \frac{\ln(\tilde{\pi}_j/\tilde{\pi}_{j,0})}{\eta_j}\right], \tag{6}$$

where $r_{j,t} = \nabla\ell_t(\widehat{\theta}_{t-1})^\top(\widehat{\theta}_{t-1} - \theta_j)$, where $\tilde{\pi}_{j,0}$ are the initial weights assigned to the experts by the algorithm. In the original version of the BOA algorithm, each expert θ_k is assigned to a single learning rate η_k. In Algorithm 1 each parameter θ_k for $k = 1, \dots, K$ is replicated several times, each replicate being assigned a different learning rate $\eta_i = e^{-i}E^{-1}$ for $1 \leqslant i \leqslant \ln(ET^2)$. Algorithm 1 corresponds to applying BOA on experts indexed by couples $j = (k, i)$ of a parameter θ_k for $k = 1, \dots, K$ and a learning-rate $\eta_i = e^{-i}E^{-1}$. Each couple $j = (k, i)$ is assigned the initial weight $\tilde{\pi}_{j,0} = \widehat{\pi}_{k,0}/\ln(ET^2)$. We refer to these couples of parameter-learning rate $j = (k, i)$ as experts.

For each parameter $\theta_k, k \in \{1, \dots, K\}$, let $1 \leqslant i_k \leqslant \ln(ET^2)$ be the index of a learning rate which will be chosen later by the analysis in order to optimize the final bound. Let π be a distribution over the index set $\{1, \dots, K\}$. We now apply Inequality (6) to a specific distribution $\tilde{\pi}$ on the experts. We choose $\tilde{\pi}$ so that it assigns all the mass π_k on the expert (k, i_k) and no mass on the experts (k, i) for $i \neq i_k$. In other words, $\tilde{\pi}_j = \pi_k \mathbb{1}_{i=i_k}$. Then $\ln(\tilde{\pi}_j/\tilde{\pi}_{j,0}) = \ln(\pi_k/\widehat{\pi}_{k,0}\ln(ET^2))$ and Inequality (6) entails

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_{k\sim\pi}[r_{k,t}] &\leqslant \mathbb{E}_{k\sim\pi}\left[\underbrace{e^{-i_k}E^{-1}}_{:=\lambda_k}\sum_{t=1}^{T} r_{k,t}^2 + e^{i_k}E\big(\ln(\pi_k/\widehat{\pi}_{k,0}) + \ln\ln(ET^2)\big)\right] \\ &= \mathbb{E}_{k\sim\pi}\left[\lambda_k \sum_{t=1}^{T} r_{k,t}^2 + \frac{\ln(\pi_k/\widehat{\pi}_{k,0}) + \ln\ln(ET^2)}{\lambda_k}\right], \end{aligned} \tag{7}$$

where we defined $\lambda_k := e^{-i_k}E^{-1}$. Now, by choosing i_k, this bound may be optimized with respect to any λ_k of the form $e^{-i_k}E^{-1}$, with $1 \leqslant i_k \leqslant \ln(ET^2)$. To get the minimum over any $\lambda_k > 0$, we pay additional additive and multiplicative terms due to edge effects that we compute now. Fix $k > 0$ and define $V_k = \sum_{t=1}^{T} r_{k,t}^2$. The minimum is reached when both terms in (7) are equal. This yields the optimal choice $\lambda_k \approx (V_k/a)^{-1/2}$, where $a_k := \ln(\pi_k/\pi_{k,0}) + \ln\ln(ET^2)$. However, because of edge effects, this is only possible when $1/(ET)^2 \leqslant (V_k/a)^{-1/2} \leqslant 1/(Ee)$. We distinguish three cases:

- if $\sqrt{a_k/V_k} > 1/(eE)$: then, we choose $\lambda_k = 1/(eE)$, which yields:
$$\lambda_k V_k + \frac{a_k}{\lambda_k} \leqslant \frac{2a_k}{\lambda_k} = 2ea_kE \leqslant 6a_kE$$
- if $1/(ET)^2 \leqslant (V_k/a_k)^{-1/2} \leqslant 1/(Ee)$: then, we can choose λ_k such that
$$\frac{\lambda_k}{\sqrt{e}} \leqslant (V_k/a_k)^{-1/2} \leqslant \sqrt{e}\lambda_k\,,$$
which entails $\lambda_k V_k + \frac{a_k}{\lambda_k} \leqslant 2\sqrt{e}\sqrt{a_kV_k} \leqslant 4\sqrt{a_kV_k}$
- if $\sqrt{a_k/V_k} < (ET)^{-2}$: then, the choice $\lambda_k = (ET)^{-2}$ gives
$$\lambda_k V_k + \frac{a_k}{\lambda_k} \leqslant 2\lambda_k V_k = \frac{2V_k}{E^2T^2} \leqslant \frac{2}{T}\,,$$
because $r_{k,t}^2 \leqslant E^2$.

[5]It is also a specific case of Squint of [18] with a discrete distribution over the learning rates

Putting the three cases together and plugging into Inequality (7) yields

$$\sum_{t=1}^{T} \mathbb{E}_{k\sim\pi}[r_{k,t}] \leqslant \mathbb{E}_{k\sim\pi}\left[4\sqrt{a_k V_k} + 6a_k E\right] + \frac{2}{T}. \tag{8}$$

We recall Young's inequality.

Lemma C.1 (Young's inequality)**.** *For all $a, b \geqslant 0$ and $p, q > 0$ such that $1/p + 1/q = 1$, then $ab \leqslant a^p/p + b^q/q$.*

Applying it, with $p = q = 2$, and $a = \sqrt{2\lambda_k V_k}$ and $b = \sqrt{8a_k/\lambda_k}$, we get $4\sqrt{a_k V_k} \leqslant \lambda_k V_k + 4a_k/\lambda_k$ for any $\lambda_k > 0$. Therefore, substituting into Inequality (8), for any distribution π over $\{1, \dots, K\}$, we have

$$\sum_{t=1}^{T} \mathbb{E}_{k\sim\pi}[r_{k,t}] \leqslant \mathbb{E}_{k\sim\pi}\left[\lambda_k V_k + \frac{4a_k}{\lambda_k} + 6a_k E\right] + \frac{2}{T}\,, \tag{9}$$

where we recall that $V_k = \sum_{t=1}^{T} r_{k,t}^2$ and $a_k = \ln(\pi_k/\pi_{k,0}) + \ln\ln(ET^2)$. For simplicity, from now on, we will denote $\mathbb{E}_{k\sim\pi}$ by $\mathbb{E}_\pi$. Using Theorem 4.1 of [31] for $\eta_{j,t} = \lambda_j$ independent of t, we obtain with probability $1 - e^{-x}$ and integrating with respect to π

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_{t-1}[\mathbb{E}_\pi[r_{k,t}]] &\leqslant \sum_{t=1}^{T} \mathbb{E}_\pi[r_{k,t}] + \mathbb{E}_\pi\left[\lambda_k \sum_{t=1}^{T} r_{k,t}^2 + \frac{x}{\lambda_k}\right] \\ &\overset{(9)}{\leqslant} \mathbb{E}_\pi\left[2\lambda_k \sum_{t=1}^{T} r_{k,t}^2 + \frac{x + 4a_k}{\lambda_k} + 6a_k E\right] + \frac{2}{T}. \end{aligned} \tag{10}$$

To apply Assumption (A2), we need to transform the second order term (the sum of $r_{k,s}^2$ in the right-hand side) into a cumulative risk. This can be done using a Poissonian inequality for martingales (see for instance Theorem 9 of [12]): with probability at least $1 - e^{-x}$

$$\sum_{t=1}^{T} r_{k,t}^2 \leqslant 2\sum_{t=1}^{T} \mathbb{E}_{t-1}\big[r_{k,t}^2\big] + \frac{9}{4}E^2 x\,.$$

Substituting into the previous regret inequality, this yields for any $\lambda_k > 0$ and any distribution π over $\{1, \dots, K\}$

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}\Big[\mathbb{E}_\pi[r_{k,t}]\Big] \leqslant \mathbb{E}_\pi\left[4\lambda_k \sum_{t=1}^{T} \mathbb{E}_{t-1}[r_{k,t}^2] + \frac{9}{2}\lambda_k E^2 x + \frac{4a_k + x}{\lambda_k} + 6a_k E\right] + \frac{2}{T}. \tag{11}$$

Now, we are ready to apply Assumption (A2) in order to cancel the sum in the right-hand side. Assumption (A2) ensures that for any time $t \geqslant 1$

$$\mathbb{E}_{t-1}\big[\ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta_k)\big] \leqslant \mathbb{E}_{t-1}[r_{k,t}] - \big(\alpha\mathbb{E}_{t-1}[r_{k,t}^2]\big)^{1/\beta}\,.$$

Therefore, summing over $t = 1, \dots, T$ and using the preceding inequality with probability at least $1 - 2e^{-x}$

$$\begin{aligned}\mathbb{E}_\pi\left[\sum_{t=1}^{T} \mathbb{E}_{t-1}\big[\ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta_k)\big]\right] &\leqslant \mathbb{E}_\pi\left[\sum_{t=1}^{T} \mathbb{E}_{t-1}[r_{k,t}] - \big(\alpha\mathbb{E}_{t-1}[r_{k,t}^2]\big)^{1/\beta}\right] \\ &\leqslant \mathbb{E}_\pi\left[4\lambda_k \sum_{t=1}^{T} \mathbb{E}_{t-1}[r_{k,t}^2] - \sum_{t=1}^{T}\Big(\alpha\mathbb{E}_{t-1}[r_{k,t}^2]\Big)^{1/\beta} + \frac{9}{2}\lambda_k E^2 x + \frac{4a_k + x}{\lambda_k} + 6a_k E\right] + \frac{2}{T}. \end{aligned} \tag{12}$$

Now, we use Young's inequality (see Lemma C.1) again to cancel the two sums in the right-hand side. Let $\gamma > 0$ to be fixed later by the analysis. Using $a = \mathbb{E}_{t-1}[r_{k,t}^2]/\gamma$, $b = \gamma$, $p = 1/\beta$, and $q = 1/(1-\beta)$, it yields

$$\mathbb{E}_{t-1}[r_{k,t}^2] \leqslant \frac{\beta\big(\mathbb{E}_{t-1}[r_{k,t}^2]\big)^{1/\beta}}{\gamma^{1/\beta}} + \big(1-\beta\big)\gamma^{1/(1-\beta)}\,.$$

Thus,

$$\lambda_k \mathbb{E}_{t-1}[r_{k,t}^2] \leqslant \frac{\lambda_k \beta \big(\mathbb{E}_{t-1}[r_{k,t}^2]\big)^{1/\beta}}{\gamma^{1/\beta}} + \lambda_k \big(1-\beta\big)\gamma^{1/(1-\beta)}\,.$$

The choice $\gamma = (4\lambda_k\beta)^\beta/\alpha$ yields $4\lambda_k\beta/\gamma^{1/\beta} = \alpha^{1/\beta}$, which entails

$$\begin{aligned}
4\lambda_k \mathbb{E}_{t-1}[r_{k,t}^2] - \big(\alpha \mathbb{E}_{t-1}[r_{k,t}^2]\big)^{1/\beta} &\leqslant 4\lambda_k\big(1-\beta\big)\gamma^{1/(1-\beta)} \\
&= 4\lambda_k\big(1-\beta\big)\Big(\frac{(4\lambda_k\beta)^\beta}{\alpha}\Big)^{1/(1-\beta)} \\
&= 4\big(1-\beta\big)(4\beta)^{\beta/(1-\beta)}\Big(\frac{\lambda_k}{\alpha}\Big)^{1/(1-\beta)} \\
&\leqslant 4\Big(\frac{4\lambda_k}{\alpha}\Big)^{1/(1-\beta)}\,. \qquad (13)
\end{aligned}$$

Summing over t and substituting into Inequality (12), we get

$$\mathbb{E}_\pi\left[\sum_{t=1}^T \mathbb{E}_{t-1}\big[\ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta_k)\big]\right] \leqslant E_\pi\Bigg[\underbrace{4\Big(\frac{4\lambda_k}{\alpha}\Big)^{1/(1-\beta)}T + \frac{4a_k+x}{\lambda_k}}_{=:R_k} + \frac{9}{2}\lambda_k E^2 x + 6a_k E\Bigg] + \frac{2}{T}\,. \quad (14)$$

We optimize λ_k by equalizing the two main terms of R_k:

$$4\Big(\frac{4\lambda_k}{\alpha}\Big)^{1/(1-\beta)}T = \frac{4a_k+x}{\lambda_k} \Leftrightarrow \lambda_k = \left(\frac{4a_k+x}{4T}\right)^{\frac{1-\beta}{2-\beta}}\left(\frac{\alpha}{4}\right)^{\frac{1}{2-\beta}}\,.$$

We express R_k in termes of λ_k using this identity

$$\frac{R_k}{T} = 2\frac{4a_k+x}{\lambda_k T} = 2\left(\frac{4a_k+x}{\alpha T}\right)^{\frac{1}{2-\beta}} 4^{\frac{1-\beta}{2-\beta}} \leqslant 4\left(\frac{16a_k+4x}{\alpha T}\right)^{\frac{1}{2-\beta}}\,.$$

The choice $\lambda_k = 1/(2E)$ would give

$$\frac{R_T}{T} \leqslant 4\Big(\frac{4\lambda_k}{\alpha}\Big)^{1/(1-\beta)} + \frac{4a_k+x}{T\lambda_k} \leqslant \frac{(4a_k+x)E}{T}\,.$$

So that we can assume $\lambda_k \leqslant 1/(2E)$ and

$$\frac{R_T}{T} \leqslant 4\left(\frac{16a_k+4x}{\alpha T}\right)^{\frac{1}{2-\beta}} + \frac{(4a_k+x)E}{T}$$

Substituting into Inequality (14) and upper-bounding $\lambda_k E^2 \leqslant E/2$, gives

$$\frac{1}{T}\mathbb{E}_\pi\left[\sum_{t=1}^T \mathbb{E}_{t-1}\big[\ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta_k)\big]\right] \leqslant E_\pi\left[4\left(\frac{16a_k+4x}{\alpha T}\right)^{\frac{1}{2-\beta}} + \frac{(10a_k+4x)E}{T}\right] + \frac{2}{T^2}\,.$$

Replacing $a_k = \ln(\pi_k/\pi_{k,0}) + \ln\ln(ET^2)$ concludes the proof.

C.2. Proof of Theorem 3.2. We denote by $\theta_1, \dots, \theta_K$ the elements of Θ_0. We recall that we use a particular case of Algorithm 1. We can thus follow the proof of Theorem 2.1 and start from Inequality (8). We apply it to a Dirac distributions π on $\{1, \dots, K\}$. We get that for any $1 \leqslant k \leqslant K$, for any $\lambda_k > 0$,

$$\sum_{t=1}^T r_{k,t} \leqslant 4\sqrt{a\sum_{t=1}^T r_{k,t}^2} + 6aE + \frac{2}{T}\,. \qquad (15)$$

where $a := \ln(K) + \ln\ln(ET^2)$ and where we remind the notation of the linearized instantaneous regret $r_{k,t} = \nabla\ell_t(\widehat{\theta}_{t-1})^\top(\widehat{\theta}_{t-1} - \theta_k)$ for $1 \leqslant k \leqslant K$.

Let $\theta^* \in \mathbb{R}^d$, let $\varepsilon := D(\theta^*, \Theta_0)$ and $k^* \in \{1 \leqslant k \leqslant K\}$ such that $\|\theta^* - (1-\varepsilon)\theta_{k^*}\|_1 \leqslant \varepsilon$. Then it exists $\tilde{\theta}$ with $\|\tilde{\theta}\|_1 \leqslant 1$ such that

$$\theta^* = (1-\varepsilon)\theta_{k^*} + \varepsilon\tilde{\theta}\,. \qquad (16)$$

Since $\{\theta \in \mathcal{B}_1 : \|\theta\|_1 = 1, \|\theta\|_0 = 1\} \subset \Theta_0$, we can write $\tilde{\theta}$ as a combination of elements of Θ_0. Hence, from (16), it exists a distribution $\pi = (\pi_1, \dots, \pi_K) \in \Delta_K$ such that

$$\theta^* = \sum_{k=1}^K \pi_k \theta_k \quad \text{and} \quad 1 - \pi_{k^*} \leqslant \varepsilon.$$

Denoting $r_t := \nabla \ell_t(\widehat{\theta}_{t-1})^\top (\widehat{\theta}_{t-1} - \theta^*)$, we thus get

$$\begin{aligned} r_t := \nabla \ell_t(\widehat{\theta}_{t-1})^\top (\widehat{\theta}_{t-1} - \theta^*) &= \nabla \ell_t(\widehat{\theta}_{t-1})^\top \Big(\widehat{\theta}_{t-1} - \sum_{k=1}^K \pi_k \theta_k\Big) \\ &= \nabla \ell_t(\widehat{\theta}_{t-1})^\top (\widehat{\theta}_{t-1} - \mathbb{E}_{k\sim\pi}[\theta_k]) = \mathbb{E}_{k\sim\pi}\Big[r_{k,t}\Big], \end{aligned}$$

and integrating Inequality (15) with respect to π, we obtain

$$\begin{aligned} \sum_{t=1}^T r_t &\leqslant \mathbb{E}_{k\sim\pi}\left[4\sqrt{a\sum_{t=1}^T \big(\nabla \ell_t(\widehat{\theta}_{t-1})^\top (\widehat{\theta}_{t-1} - \theta^* + \theta^* - \theta_k)\big)^2}\right] + \frac{2}{T} + 6aE \\ &\leqslant 4\sqrt{a\sum_{t=1}^T r_t^2} + 4\mathbb{E}_{k\sim\pi}\left[\sqrt{a\sum_{t=1}^T \big(\nabla \ell_t(\widehat{\theta}_{t-1})^\top (\theta^* - \theta_k)\big)^2}\right] + \frac{2}{T} + 6aE\,. \end{aligned} \tag{17}$$

Let us upper bound the second term of the right hand side.

$$\begin{aligned} &\mathbb{E}_{k\sim\pi}\left[\sqrt{\sum_{t=1}^T \big(\nabla \ell_t(\widehat{\theta}_{t-1})^\top (\theta^* - \theta_k)\big)^2}\right] \\ &\quad\leqslant \sqrt{\sum_{t=1}^T \|\nabla \ell_t(\widehat{\theta}_{t-1})\|_\infty^2} \sum_{k=1}^K \pi_k \|\theta^* - \theta_k\|_1 \\ &\quad\leqslant \sqrt{\sum_{t=1}^T \|\nabla \ell_t(\widehat{\theta}_{t-1})\|_\infty^2} \Big(\pi_{k^*}\|\theta^* - \theta_{k^*}\|_1 + (1-\pi_{k^*}) \max_{1\leqslant k\leqslant K} \|\theta^* - \theta_k\|_1\Big) \\ &\quad\leqslant \sqrt{\sum_{t=1}^T \|\nabla \ell_t(\widehat{\theta}_{t-1})\|_\infty^2} \Big(\pi_{k^*}\|\theta^* - \theta_{k^*}\|_1 + 2(1-\pi_{k^*})\Big), \end{aligned} \tag{18}$$

where the last inequality is because $\|\theta^* - \theta_k\|_1 \leqslant \|\theta_k\|_1 + \|\theta^*\|_1 \leqslant 2$. We also have from the definition of θ^* (see before (16))

$$\|\theta^* - \theta_{k^*}\|_1 \leqslant \|\theta^* - (1-\varepsilon)\theta_{k^*} + \varepsilon\theta_{k^*}\|_1 \leqslant \|\theta^* - (1-\varepsilon)\theta_{k^*}\|_1 + \varepsilon\|\theta_{k^*}\|_1 \leqslant 2\varepsilon.$$

Therefore, substituting into (18) we get

$$\mathbb{E}_{k\sim\pi}\left[\sqrt{\sum_{t=1}^T \big(\nabla \ell_t(\widehat{\theta}_{t-1})^\top (\theta^* - \theta_k)\big)^2}\right] \leqslant 4\varepsilon\sqrt{\sum_{t=1}^T \|\nabla \ell_t(\widehat{\theta}_{t-1})\|_\infty^2} = 4\varepsilon \bar{G}_T \sqrt{T},$$

where $\bar{G}_T := \sqrt{\frac{1}{T}\sum_{t=1}^T \|\nabla \ell_t(\widehat{\theta}_{t-1})\|_\infty^2} \leqslant G$.

Therefore, substituting into Inequality (17), we have

$$\sum_{t=1}^T r_t \leqslant 4\sqrt{a\sum_{t=1}^T r_t^2} + 16\varepsilon \bar{G}_T\sqrt{aT} + \frac{2}{T} + 6aE\,,$$

which yields by Young's inequality for any $\lambda > 0$

$$\sum_{t=1}^T r_t \leqslant \lambda \sum_{t=1}^T r_t^2 + \frac{4a}{\lambda} + \underbrace{16\varepsilon \bar{G}_T\sqrt{aT} + \frac{2}{T} + 6aE}_{=:z}\,. \tag{19}$$

Now, we recognize an inequality similar to Inequality (9). There only are a few technical differences which do not matter in the analysis: we consider here a Dirac distribution π on the comparison parameter θ^* and we have some additional rest terms that we denote by $z := 16\varepsilon\bar{G}_T\sqrt{aT} + \frac{2}{T} + 6aE$ for simplicity. We can then follow the lines of the proof of Theorem 2.1 after Inequality (9)

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}_{t-1}[r_t] &\overset{\text{Thm 4.1 of [31]}}{\leqslant} \sum_{t=1}^{T} r_t + \lambda\sum_{t=1}^{T} r_t^2 + \frac{x}{\lambda} \\
&\overset{(19)}{\leqslant} 2\lambda\sum_{t=1}^{T} r_t^2 + \frac{4a+x}{\lambda} + z \\
&\overset{\text{Thm 9 of [12]}}{\leqslant} 4\lambda\sum_{t=1}^{T}\mathbb{E}_{t-1}[r_t^2] + \frac{4a+x}{\lambda} + \frac{9}{2}\lambda E^2 x + z. \qquad (20)
\end{aligned}$$

Using Assumption (A2) then yields

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}_{t-1}\big[\ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta^*)\big] &\leqslant \sum_{t=1}^{T}\mathbb{E}_{t-1}[r_t] - \big(\alpha\mathbb{E}_{t-1}[r_t^2]\big)^{1/\beta} \\
&\overset{(20)}{\leqslant} 4\lambda\sum_{t=1}^{T}\mathbb{E}_{t-1}[r_t^2] - \big(\alpha\mathbb{E}_{t-1}[r_t^2]\big)^{1/\beta} + \frac{4a+x}{\lambda} + \frac{9}{2}\lambda E^2 x + z \\
&\overset{(13)}{\leqslant} 4\Big(\frac{4\lambda}{\alpha}\Big)^{1/(1-\beta)} + \frac{4a+x}{\lambda} + \frac{9}{2}\lambda E^2 x + z.
\end{aligned}$$

This yields an inequality similar to Inequality (14). Optimizing in $\lambda > 0$, as we did for Inequality (14) gives:

$$\lambda = \min\left\{\frac{1}{2E}, \left(\frac{4a+x}{4T}\right)^{\frac{1-\beta}{2-\beta}}\left(\frac{\alpha}{4}\right)^{\frac{1}{2-\beta}}\right\},$$

and

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{t-1}\big[\ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta^*)\big] \leqslant 4\left(\frac{16a+4x}{\alpha T}\right)^{\frac{1}{2-\beta}} + \frac{(4a+x)E}{T} + \frac{9Ex}{4T} + \frac{z}{T}\,.$$

where we recall that $a = \ln(K) + \ln\ln(ET^2)$, $z = 16\varepsilon\bar{G}_T\sqrt{aT} + \frac{2}{T} + 6aE$ and $\bar{G}_T := \sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\nabla\ell_t(\widehat{\theta}_{t-1})\|_\infty^2} \leqslant G$. Replacing z with its definition and simplifying yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{t-1}\big[\ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta^*)\big] \leqslant 4\left(\frac{16a+4x}{\alpha T}\right)^{\frac{1}{2-\beta}} + \frac{(10a+4x)E}{T} + 16\varepsilon\bar{G}_T\sqrt{\frac{a}{T}} + \frac{2}{T^2}\,. \qquad (21)$$

Keeping the main terms only and replacing $\varepsilon := D(\theta, \Theta_0)$ concludes the proof.

C.3. Proof of Lemma B.1. Let $\pi := \|\theta' - \theta\|_1/(\|\theta' - \theta\|_1 + 1 - \|\theta\|_1)$. Then, thanks to the triangular inequality, we have

$$\begin{aligned}
\big\|\theta - (1-\pi)\theta'\big\|_1 = \big\|(1-\pi)(\theta-\theta') + \pi\theta\big\|_1 &\leqslant (1-\pi)\|\theta-\theta'\|_1 + \pi\|\theta\|_1 \\
&= \frac{(1-\|\theta\|_1)\|\theta-\theta'\|_1 + \|\theta-\theta'\|_1\|\theta\|_1}{\|\theta-\theta'\|_1 + 1 - \|\theta\|_1} = \pi\,.
\end{aligned}$$

The Definition 3.1 of $D(\theta, \theta')$ concludes the proof.

C.4. Proof of Lemma B.2. Denote $\pi := 1 - \min_{1\leqslant i\leqslant d}|\theta_i|/|\theta'_i|$. Then, for any $1 \leqslant i \leqslant d$, $|\theta_i| \geqslant (1-\pi)|\theta'_i|$. Because θ'_i and θ_i have same signs, this yields $\big|\theta_i - (1-\pi)\theta'_i\big| = \big|\theta_i\big| - (1-\pi)\big|\theta'_i\big|$ for all $1 \leqslant i \leqslant d$. Summing over $i = 1, \ldots, d$, entails

$$\begin{aligned}
\big\|\theta - (1-\pi)\theta'\big\|_1 = \sum_{i=1}^{d}\Big|\theta_i - (1-\pi)\theta'_i\Big| = \sum_{i=1}^{d}\big|\theta_i\big| - (1-\pi)\big|\theta'_i\big| & \\
= \|\theta\|_1 - (1-\pi)\|\theta'\|_1 \overset{\|\theta'\|_1\geqslant\|\theta\|_1}{\leqslant} \pi\|\theta\|_1 \leqslant \pi. \qquad (22)&
\end{aligned}$$

Therefore, the Definition 3.1 of $D(\theta,\theta')$ concludes the proof of the first inequality. Now, let $1 \leqslant i \leqslant d$, if $|\theta'_i| \leqslant |\theta_i|$ then $1 - |\theta_i|/|\theta'_i| \leqslant 0$ and the second inequality holds. Otherwise, we have

$$1 - \frac{|\theta_i|}{|\theta'_i|} = \frac{|\theta'_i| - |\theta_i|}{|\theta'_i|} \overset{|\theta'_i|\geqslant|\theta_i|}{=} \frac{|\theta'_i - \theta_i|}{|\theta'_i|} \overset{|\theta'_i|\geqslant|\theta_i|}{\leqslant} \frac{|\theta'_i - \theta_i|}{|\theta_i|} \leqslant \frac{\|\theta' - \theta\|_\infty}{\Delta}\,,$$

which concludes the proof of the Lemma.

C.5. Proof of Lemma B.3. Let $\theta, \theta' \in \mathcal{B}_1$ such that $\|\theta - \theta'\|_\infty \leqslant \varepsilon$. First, we check that $\tilde{\theta}$ satisfies the assumptions of Lemma B.2. Since $\|\theta' - \theta\|_\infty \leqslant \varepsilon$, for all coordinates $1 \leqslant i \leqslant d$, we have $S_\varepsilon(\theta')_i = 0$ or $\operatorname{sign}(S_\varepsilon(\theta'))_i = \operatorname{sign}(\theta_i)$. Therefore, $\operatorname{sign}(\tilde{\theta}_i) = \operatorname{sign}(S_\varepsilon(\theta')_i) \in \{0, \operatorname{sign}(\theta_i)\}$. Furthermore,

$$\|S_\varepsilon(\theta')\|_1 \geqslant \sum_{i\in\operatorname{Supp}(\theta)} \big|S_\varepsilon(\theta')_i\big| \geqslant \sum_{i\in\operatorname{Supp}(\theta)} \big(\big|\theta'_i\big| - \varepsilon\big) \geqslant \sum_{i\in\operatorname{Supp}(\theta)} \big(\big|\theta_i\big| - 2\varepsilon\big) \geqslant \|\theta\|_1 - 2d_0\varepsilon\,. \quad (23)$$

If $S_\varepsilon(\theta') = 0$, then $\|\tilde{\theta}\|_1 = 0$ and $\|\theta\|_1 \leqslant 2d_0\varepsilon$ so that $D(\theta,\tilde{\theta}) \leqslant 1 \leqslant 2d_0\varepsilon/\|\theta\|_1$. Therefore, we can assume from now that $S_\varepsilon(\theta') \neq 0$. By definition of $\tilde{\theta}$, Inequality (23) yields $\|\tilde{\theta}\|_1 = \big(\|S_\varepsilon(\theta')\|_1 + 2d_0\varepsilon\big) \wedge 1 \geqslant \|\theta\|_1$. Then $\tilde{\theta}$ satisfies the assumptions of Lemma B.2, which we can apply

$$D(\theta,\tilde{\theta}) \leqslant 1 - \min_{1\leqslant i\leqslant d} \frac{|\theta_i|}{|\theta'_i|} = \max_{i\in\operatorname{Supp}(\tilde{\theta})} \frac{|\tilde{\theta}_i| - |\theta_i|}{|\tilde{\theta}_i|}\,. \quad (24)$$

We consider two cases:

- $\|S_\varepsilon(\theta')\|_1 \geqslant 1 - 2d_0\varepsilon$ in which case for $i \in \operatorname{Supp}(\tilde{\theta})$

  $$\tilde{\theta}_i = \frac{S_\varepsilon(\theta')_i}{\|S_\varepsilon(\theta')\|_1} = \frac{(|\theta'_i| - \varepsilon)\operatorname{sign}(\theta'_i)}{\|S_\varepsilon(\theta')\|_1}$$

  so that $|\tilde{\theta}_i| = (|\theta'_i| - \varepsilon)/\|S_\varepsilon(\theta')\|_1$ and upper-bounding $-|\theta_i| \leqslant -|\theta'_i| - \varepsilon$ we get

$$\frac{|\tilde{\theta}_i| - |\theta_i|}{|\tilde{\theta}_i|} = \frac{|\theta'_i| - \varepsilon - |\theta_i|\|S_\varepsilon(\theta')\|_1}{|\theta'_i| - \varepsilon} \leqslant \frac{|\theta'_i| - \varepsilon - (|\theta'_i| - \varepsilon)\|S_\varepsilon(\theta')\|_1}{|\theta'_i| - \varepsilon}$$
$$\leqslant 1 - \|S_\varepsilon(\theta')\|_1 \leqslant 2d_0\varepsilon \leqslant \frac{2d_0\varepsilon}{\|\theta\|_1}\,.$$

  Substituting into Inequality (24) concludes this case.
- Otherwise $\|S_\varepsilon(\theta')\|_1 \leqslant 1 - 2d_0\varepsilon$ and for $i \in \operatorname{Supp}(\tilde{\theta}) = \operatorname{Supp}(S_\varepsilon(\theta'))$

  $$|\tilde{\theta}_i| = |S_\varepsilon(\theta')_i|\Big(1 + \frac{2d_0\varepsilon}{\|S_\varepsilon(\theta')\|_1}\Big) = (|\theta'_i| - \varepsilon)\Big(1 + \frac{2d_0\varepsilon}{\|S_\varepsilon(\theta')\|_1}\Big)\,,$$

  which implies upper-bounding $-|\theta_i| \leqslant -|\theta'_i| - \varepsilon$,

$$\begin{aligned}
\frac{|\tilde{\theta}_i| - |\theta_i|}{|\tilde{\theta}_i|} &= \frac{(|\theta'_i| - \varepsilon)\Big(1 + \frac{2d_0\varepsilon}{\|S_\varepsilon(\theta')\|_1}\Big) - |\theta_i|}{(|\theta'_i| - \varepsilon)\Big(1 + \frac{2d_0\varepsilon}{\|S_\varepsilon(\theta')\|_1}\Big)} \\
&\leqslant \frac{(|\theta'_i| - \varepsilon)\frac{2d_0\varepsilon}{\|S_\varepsilon(\theta')\|_1}}{(|\theta'_i| - \varepsilon)\Big(1 + \frac{2d_0\varepsilon}{\|S_\varepsilon(\theta')\|_1}\Big)} \\
&= \frac{2d_0\varepsilon}{\|S_\varepsilon(\theta')\|_1 + 2d_0\varepsilon} \\
&\leqslant \frac{2d_0\varepsilon}{\|\tilde{\theta}\|_1} \leqslant \frac{2d_0\varepsilon}{\|\theta\|_1}\,.
\end{aligned}$$

Substituting the obtained bounds in each cases into Inequality (24) concludes the proof.

C.6. Proof of Theorem 3.3. For technical reasons, we perform the proof for $\theta \in \mathcal{B}_{1/2}$ only. However, optimization on $\mathcal{B}_1$ can be obtained by renormalizing the losses considering $\ell_t(2\theta)$ instead of ℓ_t. We leave this generalization for the reader. For simplicity, we also assume that $T = 2^I - 1$. Let $\theta \in \mathcal{B}_{1/2}$ and denote $d_0 = \|\theta\|_0$.

Part 1 ($\tilde{\mathcal{O}}(\sqrt{T})$ regret – logarithmic dependence on d_0 and d) First, we prove the slow rate bound obtained by Algorithm 1. Let $i \geqslant 0$. Denote by $\theta_1, \dots, \theta_{3d}$ the $3d$ elements of $\Theta^{(i)}$. For any distribution $\pi \in \Delta_{3d}$ over $\Theta^{(i)}$, we have from Inequality (8):

$$\sum_{t=t_i}^{t_{i+1}-1} \mathbb{E}_{k\sim\pi}[r_{k,t}] \leqslant \mathbb{E}_{k\sim\pi}\left[4\sqrt{a_k V_k} + 6a_k E\right] + \frac{2}{T}. \tag{25}$$

where we recall $r_{k,t} \leqslant \nabla\ell_t(\widehat{\theta}_{t-1})^\top(\widehat{\theta}_{t-1} - \theta_k)$, $a_k := \ln(\pi_k/\pi_{k,0}) + \ln\ln(ET^2) \leqslant \ln(3d) + \ln\ln(ET^2) =: a$ and $V_k \leqslant \sum_{t=t_i}^{t_{i+1}-1} r_{k,t}^2 \leqslant t_i G^2$. Let π such that $\theta = \sum_{k=1}^{3d} \pi_k \theta_k$, then thanks to the convexity assumption on the losses, we have

$$\ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta) \leqslant \nabla\ell_t(\widehat{\theta}_{t-1})^\top(\widehat{\theta}_{t-1} - \theta_k) = \mathbb{E}_{k\sim\pi}[r_{k,t}]\,.$$

Therefore, Inequality (25) yields

$$\sum_{t=t_i}^{t_{i+1}-1} \ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta) \leqslant 4G\sqrt{at_i} + 6Ea + \frac{2}{T}\,.$$

Summing over $i = 0, \dots, j-1$ and substituting $t_i = 2^i$ we get for any $j \geqslant 1$:

$$\mathrm{Reg}_j(\theta) := \sum_{t=1}^{t_j-1} \ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta) \leqslant 4G\sqrt{a}\sum_{i=0}^{j-1} 2^{i/2} + \underbrace{6Eaj + \frac{2j}{T}}_{=:z} \leqslant 10G\sqrt{a}2^{j/2} + z\,, \tag{26}$$

where $a = \ln(2d) + \ln\ln(ET^2)$. In particular for $j = I \lesssim \ln T$ we obtain the first inequality stated by the theorem:

$$R_T(\theta) \leqslant \mathcal{O}\left(G\sqrt{\frac{\ln d + \ln\ln(ET)}{T}}\right).$$

Part 2 ($\tilde{\mathcal{O}}(T^{1/4})$ regret – logarithmic dependence on d) We prove by induction the second bound of the Theorem: that for some $c > 0$ and all $j \geqslant 0$, we have

$$\mathrm{Reg}_j(\theta) \leqslant 48\frac{ad_0c^2G^2}{\mu} + j\frac{caG^2}{\mu} + 16\sqrt{5}c\sqrt{\frac{d_0\left(G\sqrt{a}\right)^3}{\mu}}\sum_{k=0}^{j} 2^{-\frac{3j}{4}}\,. \tag{27}$$

Indeed, decomposing the cumulative regret, we have

$$\mathrm{Reg}_{j+1}(\theta) = \mathrm{Reg}_j(\theta) + \sum_{t=t_j}^{t_{j+1}-1} \ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta)\,.$$

Note that Assumption (A2) is satisfied with $\beta = 1$, $\alpha = \mu/(2G^2)$ and without the expectation $\mathbb{E}_t$. It is worth to notice that the transformation of the second-order term into a cumulative risk performed in (20) was not needed here since Assumption (A2) holds on the losses without the expectation $\mathbb{E}_t$. Therefore, the result of Theorem 3.2, that we can apply from time instance $t_j = 2^j$ to $t_{j+1} - 1$, holds almost surely with $x = 0$, $\beta = 1$ and $\alpha = \mu/(2G^2)$. We get that it exists some constant $c > 0$ such that

$$\sum_{t=t_j}^{t_{j+1}-1} \ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta) \leqslant cGD(\theta, [\theta_j^*]_{d_0})\sqrt{a2^j} + \frac{caG^2}{\mu}\,,$$

with $a = \ln(3d) + \ln\ln(ET^2)$. Replacing into the preceding inequality, it yields

$$\mathrm{Reg}_{j+1}(\theta) \leqslant \mathrm{Reg}_j(\theta) + cGD(\theta, [\theta_j^*]_{d_0})\sqrt{a2^j} + \frac{caG^2}{\mu} \tag{28}$$

Because $\theta \in \mathcal{B}_{1/2}$, we obtain from Lemma B.1

$$D\big(\theta, [\theta_j^*]_{d_0}\big) \overset{\text{(Lem. B.1)}}{\leqslant} 2\big\|\theta - [\theta_j^*]_{d_0}\big\|_1 \overset{\|\theta\|_0 = \|[\theta_j^*]_{d_0}\|_0 = d_0}{\leqslant} 2\sqrt{2d_0}\big\|\theta - [\theta_j^*]_{d_0}\big\|_2 \leqslant 2\sqrt{2d_0}\big(\big\|\theta - \theta_j^*\big\|_2 + \big\|\theta_j^* - [\theta_j^*]_{d_0}\big\|_2\big)\,.$$

By definition of the hard threshold, for any θ such that $\|\theta\|_0 = d_0$, we have

$$\big\|\theta_j^* - [\theta_j^*]_{d_0}\big\|_2 \leqslant \big\|\theta_j^* - \theta\big\|_2\,.$$

Therefore, plugging into the previous inequality

$$D\big(\theta, [\theta_j^*]_{d_0}\big) \leqslant 4\sqrt{2d_0}\big\|\theta - \theta_j^*\big\|_2\,. \tag{29}$$

But because the losses are μ-strongly convex, the average loss over several rounds is also μ-strongly convex. And since $\theta_j^* := \arg\min_{\theta\in\mathcal{B}_{1/2}} \sum_{t=1}^{t_j-1} \ell_t(\theta)$, we have for all $\theta \in \mathcal{B}_{1/2}$

$$\begin{aligned}\mu\big\|\theta - \theta_j^*\big\|_2^2 &\leqslant \frac{1}{2^j - 1}\sum_{t=1}^{t_j-1} \ell_t(\theta) - \ell_t(\theta_j^*) \\ &= \frac{\mathrm{Reg}_j(\theta_j^*) - \mathrm{Reg}_j(\theta)}{2^j - 1} \leqslant \frac{\mathrm{Reg}_j(\theta_j^*) - \mathrm{Reg}_j(\theta)}{2^{j-1}}\,.\end{aligned} \tag{30}$$

Thus, from Inequality (29), we obtain

$$D\big(\theta, [\theta_j^*]_{d_0}\big) \leqslant 8\sqrt{\frac{d_0\big(\mathrm{Reg}_j(\theta_j^*) - \mathrm{Reg}_j(\theta)\big)}{\mu 2^j}}\,.$$

Plugging into Inequality (28) gives

$$\mathrm{Reg}_{j+1}(\theta) \leqslant \mathrm{Reg}_j(\theta) + 8cG\sqrt{\frac{ad_0}{\mu}\big(\mathrm{Reg}_j(\theta_j^*) - \mathrm{Reg}_j(\theta)\big)} + \frac{caG^2}{\mu}\,. \tag{31}$$

We can upper-bound $\mathrm{Reg}_j(\theta_j^*)$ using Inequality (26). This entails

$$\mathrm{Reg}_{j+1}(\theta) \leqslant \mathrm{Reg}_j(\theta) + 8cG\sqrt{\frac{ad_0}{\mu}\big(10G\sqrt{a}2^{j/2} + z - \mathrm{Reg}_j(\theta)\big)} + \frac{caG^2}{\mu}\,.$$

Now we have an inequality of the form

$$\mathrm{Reg}_{j+1}(\theta) \leqslant \mathrm{Reg}_j(\theta) + x_1\sqrt{x_2 - \mathrm{Reg}_j(\theta)} + x_3$$

with $x_1 = 8cG\sqrt{ad_0/\mu}$, $x_2 = 10G\sqrt{a}2^{j/2} + z$ and $x_3 = (caG^2)/\mu$. If $\mathrm{Reg}_j(\theta) \geqslant 0$, $\mathrm{Reg}_{j+1}(\theta)$ is increased by at most $x_1\sqrt{x_2} + x_3$. Otherwise $\mathrm{Reg}_j(\theta) \leqslant 0$ and the right-hand side is at most $3x_1^2/4 + x_3$ (considering the maximum over $\mathrm{Reg}_j(\theta) \leqslant 0$). Therefore,

$$\begin{aligned}\mathrm{Reg}_{j+1}(\theta) &\leqslant \max\big\{3x_1^2/4, (\mathrm{Reg}_j(\theta))_+ + x_1\sqrt{x_2}\big\} + x_3\,. \\ &= \max\left\{48\frac{ad_0c^2G^2}{\mu},\ (\mathrm{Reg}_j(\theta))_+ + 16\sqrt{5}c\sqrt{\frac{d_0}{\mu}}\left(G\sqrt{\frac{a}{2^j}}\right)^{3/2}\right\} + \frac{caG^2}{\mu}\,.\end{aligned} \tag{32}$$

This concludes the induction, using the hypothesis (27). In particular, considering $j = I = \ln_2(T-1)$, we proved that

$$R_T(\theta) = \frac{\mathrm{Reg}_I(\theta)}{T} \leqslant \mathcal{O}\left(\sqrt{\frac{d_0}{\mu}}\left(G\sqrt{\frac{\ln d + \ln\ln(ET)}{T}}\right)^{\frac{3}{2}}\right)\,.$$

Part 3. ($\tilde{\mathcal{O}}(1)$ regret – square root dependence on d) Now, we prove a faster rate but at the price of a square root dependence in the total dimension d. The proof follows the same lines as the preceding part except that one changes the induction hypothesis and that one uses it to bound the regret of θ_j^*. We prove by induction: it exists $c_0 > 0$ such that for any $\theta \in \mathcal{B}_{1/2}$

$$\mathrm{Reg}_j(\theta) \leqslant j\frac{ac_0\sqrt{\|\theta\|_0 d}G^2}{\mu T}$$

where $a = \ln(3d) + \ln\ln(ET^2)$. We start from Inequality (31) obtained in Part 2:

$$\mathrm{Reg}_{j+1}(\theta) \leqslant \mathrm{Reg}_j(\theta) + 8cG\sqrt{\frac{ad_0}{\mu}\big(\mathrm{Reg}_j(\theta_j^*) - \mathrm{Reg}_j(\theta)\big)} + \frac{caG^2}{\mu}\,.$$

Now, instead of upper-bounding $\mathrm{Reg}_j(\theta_j^*)$ using Inequality (26), we use the induction hypothesis itself. Since θ_j^* is not necessarily sparse, we have

$$\mathrm{Reg}_j(\theta_j^*) \leqslant j\frac{ac_0dG^2}{\mu}\,,$$

which entails

$$\mathrm{Reg}_{j+1}(\theta) \leqslant \mathrm{Reg}_j(\theta) + 8cG\sqrt{\frac{ad_0}{\mu}\Big(j\frac{ac_0dG^2}{\mu} - \mathrm{Reg}_j(\theta)\Big)} + \frac{caG^2}{\mu}\,.$$

We obtain a regret bound of the same form than in Part 2:

$$\mathrm{Reg}_{j+1}(\theta) \leqslant \mathrm{Reg}_j(\theta) + x_1\sqrt{x_2 - \mathrm{Reg}_j(\theta)} + x_3,$$

with $x_1 = 8cG\sqrt{ad_0/\mu}$, $x_2 = (jac_0dG^2)/\mu$ and $x_3 = caG^2/\mu$. Similarly to Inequality (32), we have

$$\begin{aligned}
\mathrm{Reg}_{j+1}(\theta) &\leqslant \max\left\{3x_1^2/4, (\mathrm{Reg}_j(\theta))_+ + x_1\sqrt{x_2}\right\} + x_3 \\
&= \max\left\{48\frac{ad_0c^2G^2}{\mu}, (\mathrm{Reg}_j(\theta))_+ + \frac{8c\sqrt{c_0}a\sqrt{d_0d}G^2}{\mu}\right\} + \frac{caG^2}{\mu} \\
&\leqslant (\mathrm{Reg}_j(\theta))_+ + \frac{(49 + 8c\sqrt{c_0})a\sqrt{d_0d}G^2}{\mu}\,.
\end{aligned}$$

Choosing $c_0 > 0$ such that $49 + 8c\sqrt{c_0} \leqslant c_0$ concludes the induction. In particular, considering $j = I = \ln_2(T-1)$, we proved that

$$R_T(\theta) \leqslant \mathcal{O}\left(\frac{\sqrt{d_0d}G^2(\ln d + \ln\ln(ET))\ln T}{\mu T}\right).$$

C.7. Proof of Theorem 3.4. We recall that $\Theta^* = \arg\min_{\theta\in\mathcal{B}_1}\mathbb{E}[\ell_t(\theta)]$. The idea of the proof is to show that at each session i, SABOA performs BOA by adding sparse estimators in $\Theta^{(i)}$ that are exponentially closer to Θ^*.

Let $x > 0$. We prove by induction on $i \geqslant 0$ that with probability at least $1 - ie^{-x}$, it exists $\theta^* \in \Theta^*$ such that

$$D\big(\theta^*, \Theta^{(i)}\big) \leqslant \varepsilon 2^{-\tau i}\,, \tag{$\mathcal{H}_i$}$$

where D is defined in Definition 3.1,

$$\varepsilon := \max_{\theta^*\in\Theta^*}\left((8\sqrt{a}G)^\beta \max\left\{\frac{2}{\alpha G^2}, \frac{8\|\theta^*\|_0}{\mu}\min\left\{\frac{8\|\theta^*\|_0}{\|\theta^*\|_1^2}, \frac{1}{(1-\|\theta^*\|_1)^2}\right\}\right\}\right)^{\frac{1}{2-\beta}}, \tag{33}$$

and $\tau = \frac{1}{2-\beta} - \frac{1}{2}$. Remark that θ^* in $(\mathcal{H}_i)$ depends on i when Θ^* is not a singleton.

Initialization. For $i = 0$, by definition (see Algorithm 2), $\Theta^{(0)} := \{0\}$ and $D(\theta^*, \{0\}) \leqslant \|\theta^*\|_1 \leqslant 1$. The initialization thus holds true as soon as $\varepsilon > 1$.

Induction step. Let $i \geqslant 0$ and assume $(\mathcal{H}_i)$. We start from Theorem 3.2 (see Inequality (21) for the precise constants that we upper-bound here) that we apply for $t = t_{i-1}, \ldots, t_i - 1$ and $\theta^* \in \Theta^*$ satisfying $(\mathcal{H}_i)$: with probability $1 - e^{-x}$

$$\frac{1}{2^{i-1}}\sum_{t=t_{i-1}}^{t_i-1}\mathbb{E}_{t-1}\big[\ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta^*)\big] \leqslant \frac{2\sqrt{a}GD(\theta^*, \Theta^{(i)})}{2^{(i-1)/2}} + 4\left(\frac{a}{\alpha 2^{i-1}}\right)^{\frac{1}{2-\beta}} + \frac{aE}{2^{i-1}} + \frac{2}{2^{2i-2}}\,,$$

where for simplicity of notation we define $a := 16(1 + \ln(K_i)) + 16\ln\ln(ET^2) + 4x$, where $K_i := \mathrm{Card}(\Theta^{(i)}) + 2d$ denotes the number of experts used during the doubling session i, and where

we used $t_i = t_{i-1} + 2^{i-1}$. Using $(\mathcal{H}_i)$ together with Jensen's inequality and recalling $\bar{\theta}^{(i)} := 2^{-i+1}\sum_{t=t_{i-1}}^{t_i-1}\widehat{\theta}_{t-1}$, we obtain

$$\mathbb{E}\big[\ell_t(\bar{\theta}^{(i)}) - \ell_t(\theta^*)\big] \leqslant 2\sqrt{2a}G\varepsilon 2^{-(\frac{1}{2}+\tau)i} + 4\Big(\frac{a}{\alpha}\Big)^{\frac{1}{2-\beta}} 2^{-\frac{i}{2-\beta}} + aE2^{1-i} + 2^{3-2i}\,. \tag{34}$$

Now, we simplify this expression by showing that the last three terms of the right-hand side are negligible with respect to the first one. First, because $a \geqslant 16$ and $E \geqslant 1$, we have $16 \leqslant aE$ and thus $2^{3-2i} \leqslant aE2^{-1-i}$. Then, because $\varepsilon \geqslant \sqrt{a}$, $aE \leqslant \sqrt{a}E\varepsilon = \frac{4}{3}\sqrt{a}\varepsilon G$ and thus

$$aE2^{1-i} + 2^{3-2i} \leqslant \frac{3}{2}aE2^{-i} \leqslant 2\sqrt{a}\varepsilon G 2^{-i} \overset{\tau\leqslant 1/2}{\leqslant} 2\sqrt{a}\varepsilon G 2^{-(\frac{1}{2}+\tau)i}\,. \tag{35}$$

The second term is also dominated thanks to the definition of ε in (33)

$$\varepsilon \overset{(33)}{\geqslant} \left(\frac{2(\sqrt{8a}G)^\beta}{\alpha G^2}\right)^{\frac{1}{2-\beta}} \quad \Rightarrow \quad 2\sqrt{2a}G\varepsilon \geqslant \Big(\frac{16a}{\alpha}\Big)^{\frac{1}{2-\beta}} \overset{0\leqslant\beta\leqslant 1}{\geqslant} 4\Big(\frac{a}{\alpha}\Big)^{\frac{1}{2-\beta}}$$

and

$$\tau \overset{(33)}{=} \frac{1}{2-\beta} - \frac{1}{2} \quad \Rightarrow \frac{1}{2-\beta} \geqslant \frac{1}{2} + \tau$$

which yields

$$4\Big(\frac{a}{\alpha}\Big)^{\frac{1}{2-\beta}} 2^{-\frac{i}{2-\beta}} \leqslant 2\sqrt{2a}G\varepsilon 2^{-(\frac{1}{2}+\tau)i}\,. \tag{36}$$

Thus replacing Inequalities (35) and (36) into Inequality (34) and upper-bounding $4\sqrt{2}+2 \leqslant 8$, we get for any $\theta^* \in \Theta^*$

$$\mathbb{E}\big[\ell_t(\bar{\theta}^{(i)}) - \ell_t(\theta^*)\big] \leqslant 8\sqrt{a}G\varepsilon 2^{-(\frac{1}{2}+\tau)i}\,. \tag{37}$$

Using Assumption (A3), there exists at least one $\theta^* \in \Theta^*$ (which can be different from the preceding session), which satisfies

$$\big\|\bar{\theta}^{(i)} - \theta^*\big\|_\infty \leqslant \big\|\bar{\theta}^{(i)} - \theta^*\big\|_2 \overset{(37)+(A3)}{\leqslant} (8\sqrt{a}G\varepsilon)^{\frac{\beta}{2}}\mu^{-\frac{1}{2}}2^{-(\frac{1}{2}+\tau)\frac{\beta}{2}i} =: \varepsilon'\,. \tag{38}$$

Now, we want to apply Lemma B.3 if $\|\theta^*\|_1$ is close to 1 and Lemma B.1 if $\|\theta^*\|_1 < 1$. In order to apply Lemma B.1, we consider hard-truncated estimators $[\bar{\theta}^{(i)}]_{\tilde{d}_0}$, canceling the $d - \tilde{d}_0$ smallest components of $\bar{\theta}^{(i)}$ for $\tilde{d}_0 \in \{1,\dots,d\}$. For the (unknown) choice $\tilde{d}_0 = d_0$, since $\|[\bar{\theta}^{(i)}]_{d_0}\|_0 = \|\theta^*\|_0 = d_0$, we have $\|[\bar{\theta}^{(i)}]_{d_0} - \theta^*\|_0 \leqslant 2d_0$ and

$$\begin{aligned}\big\|[\bar{\theta}^{(i)}]_{d_0} - \theta^*\big\|_1 \leqslant \sqrt{2d_0}\big\|[\bar{\theta}^{(i)}]_{d_0} - \theta^*\big\|_2 &\leqslant \sqrt{2d_0}\big(\big\|[\bar{\theta}^{(i)}]_{d_0} - \bar{\theta}^{(i)}\big\|_2 + \big\|\bar{\theta}^{(i)} - \theta^*\big\|_2\big)\\ &\leqslant 2\sqrt{2d_0}\big\|\bar{\theta}^{(i)} - \theta^*\big\|_2 \leqslant 2\sqrt{2d_0}\varepsilon'\,.\end{aligned}$$

Applying Lemma B.1, we get

$$D\big(\theta^*, [\bar{\theta}^{(i)}]_{d_0}\big) \leqslant \frac{\big\|[\bar{\theta}^{(i)}]_{d_0} - \theta^*\big\|_1}{1 - \|\theta^*\|_1} \leqslant \frac{2\sqrt{2d_0}\varepsilon'}{1 - \|\theta^*\|_1}\,. \tag{39}$$

This bound is only useful for $\|\theta^*\|_1 < 1$. Otherwise, we want to apply Lemma B.3. However the values of ε' and $d_0 = \|\theta^*\|_0$ are unknown. We approximate them with $\tilde{\varepsilon}$ and $\tilde{d}_0$ on exponential grids, which we define now:

$$\mathcal{G}_{\varepsilon'} = \big\{2^{-k}, \quad k = 0,\dots,i\big\} \quad \text{and} \quad \mathcal{G}_{d_0} = \big\{1, 2, \dots 2^{-\lfloor \ln d \rfloor}, d\big\}\,.$$

We define for all $\tilde{\varepsilon} \in \mathcal{G}_{\varepsilon'}$ and $\tilde{d}_0 \in \mathcal{G}_{d_0}$ the dilated soft-threshold

$$\tilde{\theta}(\tilde{\varepsilon}, \tilde{d}_0) := S_{\tilde{\varepsilon}}(\bar{\theta}^{(i)})\left(1 + \frac{2\tilde{d}_0\tilde{\varepsilon}}{\|S_{\tilde{\varepsilon}}(\bar{\theta}^{(i)})\|_1}\right) \wedge \frac{1}{\|S_{\tilde{\varepsilon}}(\bar{\theta}^{(i)})\|_1}\,, \tag{40}$$

with the convention $\frac{0}{0} = 0$, recalling the definition of the soft-threshold operator $S_\varepsilon(x)_i = \mathrm{sign}(x_i)(|x_i| - \varepsilon)_+$ for all $1 \leqslant i \leqslant d$. Because $\varepsilon' \geqslant 2^{-i}$ (using $\varepsilon \geqslant \sqrt{a} \geqslant 4$, $G \geqslant 1$ and $\tau \leqslant 1/2$ and $\mu \geq 1$) and $\big\|\bar{\theta}^{(i)} - \theta^*\big\|_\infty \leqslant 1$, it exists $\tilde{\varepsilon} \in \mathcal{G}_{\varepsilon'}$ such that $\tilde{\varepsilon} \leqslant 2\varepsilon'$ and $\big\|\bar{\theta}^{(i)} - \theta^*\big\|_\infty \leqslant \tilde{\varepsilon}$. Furthermore, it exists also $\tilde{d}_0 \in \mathcal{G}_{d_0}$ such that $d_0 \leqslant \tilde{d}_0 \leqslant 2d_0$. We can thus apply Lemma B.3, which yields

$$D(\theta^*, \tilde{\theta}(\tilde{\varepsilon}, \tilde{d}_0)) \leqslant \frac{2\tilde{d}_0\tilde{\varepsilon}}{\|\theta^*\|_1} \leqslant \frac{8d_0\varepsilon'}{\|\theta^*\|_1}\,. \tag{41}$$

We define the new approximation grid

$$\Theta^{(i+1)} := \left\{\tilde{\theta}(\tilde{\varepsilon}, \tilde{d}_0), \tilde{\varepsilon} \in \mathcal{G}_{\varepsilon'}, \tilde{d}_0 \in \mathcal{G}_{d_0}\right\} \cup \left\{[\bar{\theta}^{(i)}]_{\tilde{d}_0}, \quad \tilde{d}_0 = 1, \dots, d\right\}, \tag{42}$$

where $\tilde{\theta}(\tilde{\varepsilon}, \tilde{d}_0)$ is defined in Equation (40) and $[\cdot]_k$ are hard-truncations to k coordinates. We get from Inequality (39) and (41) that

$$\begin{aligned} D\big(\theta^*, \Theta^{(i+1)}\big) &\leqslant \min\left\{\frac{\sqrt{8d_0}}{1-\|\theta^*\|_1}, \frac{8d_0}{\|\theta^*\|_1}\right\}\varepsilon' \\ &\overset{(38)}{=} (8\sqrt{a}G\varepsilon)^{\frac{\beta}{2}}\mu^{-\frac{1}{2}} \min\left\{\frac{\sqrt{8d_0}}{1-\|\theta^*\|_1}, \frac{8d_0}{\|\theta^*\|_1}\right\} 2^{-(\frac{1}{2}+\tau)\frac{\beta}{2}i}. \end{aligned}$$

To conclude the induction, it suffices to show that this is smaller then $\varepsilon 2^{-\tau(i+1)}$. Our choices of ε and τ defined in (33) was done in that purpose, so that the induction is completed.

Conclusion. Substituting the values of ε and τ into Inequality (37) and using the choice $i = \ln_2 T$ (which upper-bound the number of sessions after T times steps) concludes the proof:

$$\begin{aligned} \mathbb{E}\big[\ell_t(\bar{\theta}^{(i)}) - \ell_t(\theta^*)\big] &\overset{\text{Jensen}}{\leqslant} \frac{R_T^{(i)}}{2^i} \\ &\overset{(37)}{\leqslant} 8\sqrt{a}G\varepsilon 2^{-(\frac{1}{2}+\tau)i} \\ &\overset{(33)}{\leqslant} \max_{\theta^*\in\Theta^*}\left(\frac{128a}{T}\max\left\{\frac{1}{\alpha}, \frac{4G^2\|\theta^*\|_0}{\mu}\min\left\{\frac{1}{(1-\|\theta^*\|_1)^2}, \frac{8\|\theta^*\|_0}{\|\theta^*\|_1^2}\right\}\right\}\right)^{\frac{1}{2-\beta}}, \end{aligned}$$

where we recall that $a := 16(1 + \ln(K_i) + \ln\ln(ET^2)) + 4x$, where $K_i := \mathrm{Card}(\Theta^{(i)}) + 2d \leqslant (1 + \ln_2 d)(1 + \ln_2 T) + d$. Summing over $i = 1, \dots, \ln_2(T)$, we get the upper-bound for the cumulative risk.