
Efficient online algorithms for fast-rate regret bounds under sparsity

Pierre Gaillard

INRIA, ENS, PSL Research University
Paris, France
pierre.gaillard@inria.fr

Olivier Wintenberger

Sorbonne Université, CNRS, LPSM
Paris, France
olivier.wintenberger@upmc.fr

Abstract

We consider the problem of online convex optimization in two different settings: arbitrary and i.i.d. sequence of convex loss functions. In both settings, we provide efficient algorithms whose cumulative excess risks are controlled with fast-rate sparse bounds. First, the excess risks bounds depend on the sparsity of the objective rather than on the dimension of the parameters space. Second, their rates are faster than the slow-rate $1/\sqrt{T}$ under additional convexity assumptions on the loss functions. In the adversarial setting, we develop an algorithm BOA+ whose cumulative excess risks is controlled by several bounds with different trade-offs between sparsity and rate for strongly convex loss functions. In the i.i.d. setting under the Łojasiewicz’s assumption, we establish new risk bounds that are sparse with a rate adaptive to the convexity 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 under weak assumptions on the risk.

1 Introduction

We consider the following setting of online convex optimization where a sequence of random convex loss functions $(\ell_t : \mathbb{R}^d \rightarrow \mathbb{R})_{t \geq 1}$ is sequentially observed. At each iteration $t \geq 1$, a learner chooses a point $\hat{\theta}_{t-1} \in \mathbb{R}^d$ based on past observations $\mathcal{F}_{t-1} = \sigma(\{\ell_1, \dots, \ell_{t-1}\})$. The learner aims at minimizing the average excess risk defined as $\hat{L}_T := (1/T) \sum_{t=1}^T \mathbb{E}_{t-1}[\ell_t(\hat{\theta}_{t-1})]$ where $\mathbb{E}_{t-1} = \mathbb{E}[\cdot | \mathcal{F}_{t-1}]$. For any parameter θ in some reference set $\Theta \subset \mathbb{R}^d$, the average excess risk can be decomposed as the sum of the approximation-estimation errors:

$$\hat{L}_T = \underbrace{\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{t-1}[\ell_t(\theta)]}_{\text{approximation error}} + \underbrace{\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{t-1}[\ell_t(\hat{\theta}_{t-1})] - \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{t-1}[\ell_t(\theta)]}_{\text{estimation error} \quad := R_T(\theta)}. \quad (1)$$

Though the final goal is to minimize \hat{L}_T , a common proxy is to upper-bound the estimation term $R_T(\theta)$ (also referred to as average excess risk¹) simultaneously for all $\theta \in \Theta$. If the loss functions are exp-concave and Θ is bounded, several sequential algorithms achieve the uniform bound² on the estimation term $R_T := \sup_{\theta \in \Theta} R_T(\theta) \leq \mathcal{O}(d/T)$; see [13]. In this paper, we are interested with non-uniform bounds on $R_T(\theta)$ increasing with the complexity of θ . Such non-uniform bounds are called oracle inequalities and state that the learner achieves the best approximation-estimation

¹The average excess risk $R_T(\theta)$ generalizes the average regret more commonly used in the online learning literature by considering the Dirac masses on $\{\ell_t\}$ as conditional distributions so that $\ell_t = \mathbb{E}_{t-1}[\ell_t]$, $t \geq 1$.

²Throughout the paper \lesssim denotes an approximate inequality which holds up to universal constants and $\tilde{\mathcal{O}}$ denotes an asymptotic inequality up to logarithmic terms.

trade-off of (1). Using the ℓ_0 -norm to measure the complexity of θ , we are looking for fast-rate sparse bounds of the form

$$R_T(\theta) \leq \tilde{O} \left(\left(\frac{\|\theta\|_0}{T} \right)^{\frac{1}{2-\beta}} \right), \quad \text{for any } \theta \in \Theta.$$

The parameter $\beta \in [0, 1]$ depends on the convexity properties of the loss functions and will be specified later. We call fast-rate bound any bound which provides a better rate than $1/\sqrt{T}$ and sparse bounds any bound where some dependence on d has been replaced with $\|\theta\|_0$. Our analysis starts from a careful study of the finite case $\Theta = \{\theta_1, \dots, \theta_K\}$. We consider then online averaging algorithms on adaptive finite discretization grids that achieve sparse oracle bounds on $\Theta = \mathcal{B}_1 = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1\}$.

First contribution: fast-rate high probability quantile bound (finite Θ , adversarial data) The case of finite reference set $\Theta = \{\theta_1, \dots, \theta_K\}$ corresponds to the setting of prediction with expert advice (see Section 2.2 or [5]) where a learner makes sequential predictions over a series of rounds with the help of K experts. Hedge introduced by [19] and [26] achieves the rate $R_T \leq \mathcal{O}(\sqrt{(\ln K)/T})$. The latter is optimal for general convex loss functions but better performance can be obtained in favorable scenarios. The rate $R_T \leq \mathcal{O}((\ln K)/T)$ is for instance obtained for strongly convex loss functions in [28]. Another improvement (see [16] and references therein) is devoted to quantile bounds, i.e. bounds on $\mathbb{E}_{k \sim \pi}[R_T(\theta_k)]$ for any probability distribution $\pi \in \Delta_K^3$. The latter improve the dependence on the number of experts from $\ln K$ to the Kullback divergence $\mathcal{K}(\pi, \hat{\pi}_0)$ for any prior $\hat{\pi}_0$. They are smaller whenever many experts perform well or when a good prior knowledge is available. Squint [16] achieves a fast-rate quantile bound for adversarial data. Such a bound is obtained in high-probability by [20] but it suffers an additional gap term.

In Section 2, we extend the analysis of [16] to remove the gap term of [20]. We introduce a weak version of exp-concavity; see Assumption (A2). It depends on a parameter $\beta \in [0, 1]$ which goes from $\beta = 0$ for general convex loss functions to $\beta = 1$ for exp-concavity. We show in Theorem 2.1 that BOA [28] and Squint [16] achieve a fast rate quantile bound with high probability: i.e. $\mathbb{E}_\pi[R_T(\theta_k)] \leq \tilde{O}((\mathcal{K}(\pi, \hat{\pi}_0)/T)^{1/(2-\beta)})$.

Second contribution: efficient sparse oracle bound ($\Theta = \mathcal{B}_1$, adversarial data) The extension from finite reference sets to convex sets is natural. The seminal paper [15] introduced the Exponentiated Gradient algorithm (EG), a version of Hedge using the sub-gradients of the loss functions. The latter guarantees $R_T \leq \mathcal{O}(\sqrt{(\ln d)/T})$ for $\Theta = \mathcal{B}_1$ which is optimal for convex loss functions. Recently, fast rate $R_T \leq \tilde{O}((d/T)^{1/(2-\beta)})$ are obtained by [17] under a slightly different assumption than (A2). Here our purpose is to improve the dependence on d under the sparsity condition $\|\theta\|_0$ small. The literature on learning under sparsity with i.i.d. data is vast; we refer to [12] for a review. Yet, little work was done on sparsity bounds under adversarial data; see Table 1 for a summary. The papers [7; 18; 29] focus on providing sparse estimators $\hat{\theta}_t$ rather than sparse guarantees. More recent works [8; 14] consider sparse approximations of the sub-gradients. Though they also compare themselves with sparse parameters, they incur a bound larger than $\mathcal{O}(1/\sqrt{T})$ which is optimal in their setting. Fast rate sparse regret bounds involving $\|\theta\|_0$ were, up to our knowledge, only obtained through non-efficient (exponential time) procedures (see [10]). In Section 3.3, we provide an efficient algorithm BOA+ which satisfies the oracle inequality

$$R_T(\theta) \leq \tilde{O}((\sqrt{d\|\theta\|_0}/T) \wedge (\sqrt{\|\theta\|_0}/T^{3/4})), \quad \text{for any } \theta \in \mathcal{B}_1,$$

for strongly-convex loss functions ($\beta = 1$). The gain $\sqrt{\|\theta\|_0/d} \wedge \sqrt{\|\theta\|_0}/T$ compared with the usual rate $\mathcal{O}(d/T)$ is significant for sparse parameters θ .

A crucial step of our analysis is an intermediate result which is interesting in its own. We define an efficient algorithm with input any finite grid $\Theta_0 \subset \mathcal{B}_1$. We provide in Theorem 3.2 a bound of the form $R_T(\theta) \leq \tilde{O}(D(\theta, \Theta_0)/\sqrt{T})$ for a pseudo-metric D and any $\theta \in \mathcal{B}_1$. We say that this bound is *acceleratable* as the rate may decrease if $D(\theta, \Theta_0)$ decreases with T . In particular, it yields an oracle bound of the form $R_T(\theta) \leq \mathcal{O}(\|\theta\|_1/\sqrt{T})$.

³Here and subsequently, $\Delta_K := \{\pi \in [0, 1]^K : \|\pi\|_1 = 1\}$ denotes the simplex of dimension $K \geq 1$.

Procedure	Rate	Polynomial	Assumption	Sparsity setting
Kale et al. [8; 14]	$\text{Poly}(d)/\sqrt{T}$	Yes	Convexity	Sparse observed gradients
[7; 18; 29]	$\sqrt{\frac{\ln d}{T}}$ or $\frac{d}{T}$	Yes	(Strong) Convexity	Produce sparse estimators
SeqSEW [11]	$\frac{d_0 \ln d}{T}$	No	Strong Convexity	Sparse bound
SABOA	$\sqrt{\frac{\ln d}{T}} \wedge \frac{\sqrt{d_0 d}}{T} \ln d$	Yes	Strong Convexity	Sparse bound

Table 1: Comparison of sequential optimization procedures in sparse adversarial environment.

Third contribution: sparse regret bound under Łojasiewicz assumption ($\Theta = \mathcal{B}_1$, i.i.d. data)

In Section 3.4 we turn to a stochastic setting where the loss functions ℓ_1, \dots, ℓ_T are i.i.d.. This setting extends the regression one with random design to general loss functions. The classical Lasso procedure satisfies, in the regression setting for the quadratic risk ($\beta = 1$), $R_T(\theta) \leq \tilde{\mathcal{O}}(\|\theta\|_0/T)$ where θ is a sparse approximation of $\theta^* = \arg \min_{\theta \in \mathbb{R}^d} R_T(\theta)$, see [3]. Yet, few procedures satisfying sparse bounds are sequential; we can cite [1; 8; 9; 14; 23]. We compare in Table 2 their results and settings.

The first line of work [1; 9; 23] provides sparse rates of order $\tilde{\mathcal{O}}(\|\theta^*\|_0 \ln d/T)$. Their settings are close to the one of [3] but their methods differ; the one of [23] uses a ℓ_1 -penalized gradient descent whereas the one of [1] and [9] are based on restarting a subroutine centered around the current estimate on sessions of exponentially growing length. A common limitation of these works is that they do not provide oracle inequality. They only compete with the global optimum over \mathbb{R}^d only, which is assumed to be (approximately in [1]) sparse with a known ℓ_1 -bound. In other words, they assume that the global optimum also realizes the approximation-estimation errors trade-off in (1). In order to avoid this restriction, our first objective is to obtain the sparse bounds $R_T(\theta^*(U)) \leq \tilde{\mathcal{O}}(\|\theta^*(U)\|_0/T)$ where $\theta^*(U) \in \arg \min_{\|\theta\|_1 \leq U} R_T(\theta)$ for any $U > 0$. For U well chosen so that $\|\theta^*(U)\|_1 = U$, $\theta^*(U)$ is sparse and the approximation-estimation errors trade-off in (1) is achieved. We restrict to the case $U = 1$ suppressing the dependence on U in θ^* for the ease of notation. We leave the adaptation in $U > 0$ for future research.

The second line of works [14; 8] considers sparse approximation of sub-gradients. Yet, they provide a sparse regret bound of order $\mathcal{O}(\|\theta^*\|_0^2 \ln d/T)$ where θ^* is the optimum in \mathcal{B}_1 when the loss functions are strongly convex. Our second objective is to relax the strong convexity assumption which is too restrictive in the sequential regression setting. Indeed, the usual restricted eigenvalues conditions on the Gram matrix cannot hold uniformly for small t 's. We work under Łojasiewicz's Assumption introduced by [32; 33]: There exist $\beta > 0$ and $\mu > 0$ such that for all $\theta \in \mathcal{B}_1$, there exists a minimizer θ^* of the risk over \mathcal{B}_1 satisfying

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

The Łojasiewicz assumption depends on a parameter $\beta \in [0, 1]$ that ranges from general convex risk function ($\beta = 0$) to generalized strongly convex risk function ($\beta = 1$). In Theorem 3.4 we show that our new efficient procedure SABOA achieves a fast rate upper-bound on the average excess risk of order $\tilde{\mathcal{O}}((\|\theta^*\|_0 \ln(d)/T)^{1/(2-\beta)})$ when the optimal parameters have ℓ_1 -norm bounded by $1 - \gamma < 1$. Then we recover the optimal rate of [1; 9; 23] in a similar setting, when the global optimum is assumed to be sparse. When $\|\theta^*\|_1 = 1$, guaranteeing a good approximation-estimation trade-off in (1), the bound suffers an additional factor $\|\theta^*\|_0$. Notice that Łojasiewicz's Assumption (A3) allows multiple optima which is important when we are dealing with degenerated co-linear design (allowing zero eigenvalues in the covariance matrix). It is an open question whether the fast rate $\tilde{\mathcal{O}}((\|\theta^*\|_0^2 \ln(d)/T))$ is optimal for efficient $\mathcal{O}(dT)$ -complex procedures such as SABOA under Łojasiewicz's Assumption.

Outline of the paper To summarize our contributions, we provide

- the first high-probability quantile bound achieving a fast rate in Theorem 2.1;
- an accelerable bound on $R_T(\theta)$ that is small whenever θ is close to a prior grid Θ_0 (Thm. 3.2);
- two efficient algorithms with sparse regret bounds in the adversarial setting with strongly convex loss functions (BOA+, Thm. 3.3) and in the i.i.d. setting (SABOA, Thm. 3.4). In the latter setting, the results are obtained under the Łojasiewicz's assumption. This generalizes the usual necessary conditions for obtaining sparse bounds that are too restrictive in our sequential setting.

Procedure	Setting	Rate	Assumptions/Setting	Optimum over
Lasso [3]	B	$d_0 \ln d/T$	Mutual Coherence	\mathbb{R}^d
Kale et al. [8; 14]	S	$d_0^2 \ln d/T$	Strong Convexity + Sparse Gradients	\mathcal{B}_1
[1; 9; 23]+SABOA	S	$d_0 \ln d/T$	Strong convexity or Łojasiewicz ($\beta = 1$)	\mathbb{R}^d
SABOA	S	$d_0^2 \ln d/T$	Łojasiewicz ($\beta = 1$)	\mathcal{B}_1

Table 2: Comparison of sequential (S) and batched (B) optimization procedures in i.i.d. environment.

2 Finite reference set

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

- (A1) *Convex Lipschitz*⁴: the loss functions ℓ_t are convex on \mathcal{B}_1 and there exists $G > 0$ such that $\|\nabla \ell_t(\theta)\|_\infty \leq G$ for all $t \geq 1, \theta \in \mathcal{B}_1$.
- (A2) *Weak exp-concavity*: There exist $\alpha > 0$ and $\beta \in [0, 1]$ such that for all $t \geq 1, \theta_1, \theta_2 \in \mathcal{B}_1$, almost surely

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

For convex loss functions (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 ($\mathbb{E}_{t-1}[\ell_t]$), $t \geq 1$.
- *Exp-concavity of the loss*: Lemma 4.2, Hazan [13] states that (A2) with $\alpha \leq \frac{1}{4} \min\{\frac{1}{8G}, \kappa\}$ is implied by κ -exp-concavity of the loss functions ℓ_t , $t \geq 1$. Our assumption is slightly weaker since it holds in conditional expectation.

2.1 Fast-rate quantile bound with high probability

For prediction with $K \geq 1$ expert advice, [28] showed that a fast rate $\mathcal{O}((\ln K)/T)$ can be obtained by the BOA algorithm under the LIST condition (i.e., Lipschitz and strongly convex loss functions). In this section, we show that Assumption (A2) is enough and we improve the dependence on the total number of experts with a quantile bound.

Our algorithm is described in Algorithm 1 and corresponds to a particular case of two algorithms: the Squint algorithm of [16] used with a discrete prior over a finite set of learning rates and the BOA algorithm of [28] where each expert is replicated multiple times with different constant learning rates. The proof (with the exact constants) is deferred to Appendix C.1.

Theorem 2.1. *Let $T \geq 1$. Assume (A1) and (A2). Apply Algorithm 1, parameter $E = 4G/3$ and initial weight vector $\hat{\pi}_0 \in \Delta_K$. Then, for all $\pi \in \Delta_K$, with probability at least $1 - 2e^{-x}$, $x > 0$,*

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

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

A fast rate of this type (without quantiles property) can be obtained in expectation by using Hedge for exp-concave loss functions. However, Theorem 2.1 is stronger. First, Assumption (A2) is weaker than the exp-concavity of the loss functions ℓ_t as it holds for absolute or quantile loss functions in a sufficiently regular regression setting. Second, the algorithm uses the so-called gradient trick; See [24]. Therefore, simultaneously with the fast rate $\mathcal{O}(T^{-1/(2-\beta)})$ with respect to the experts (θ_k),

⁴Throughout 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 [6] for instance) under sub-Gaussian conditions.

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

Parameters: $\Theta_0 = \{\theta_1, \dots, \theta_K\} \subset \mathcal{B}_1$, $E > 0$ and $\hat{\pi}_0 \in \Delta_K$.

Initialization: For $1 \leq i \leq \ln(ET^2)$, define $\eta_i := (e^i E)^{-1}$.

For each iteration $t = 1, \dots, T$ do:

- Choose $\hat{\theta}_{t-1} = \sum_{k=1}^K \hat{\pi}_{k,t-1} \theta_k$ and observe $\nabla \ell_t(\hat{\theta}_{t-1})$,
- Update component-wise for all $1 \leq k \leq K$

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

the algorithm achieves the slow rate $\mathcal{O}(1/\sqrt{T})$ with respect to any convex combination $\mathbb{E}_{k \sim \pi}[\theta_k]$ (similarly to EG). Finally, high-probability regret bounds as ours are not satisfied by Hedge (see [2]).

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

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

Thanks to the quantile bounds, we pay the proportion of good experts $\ln(K/\text{Card}(\Theta'))$ in the regret instead of the total number of experts $\ln(K)$. We refer to [16] for more interesting applications. Such quantile bounds on the risk were studied by Mehta [20, Section 7] in a batch i.i.d. setting (i.e., ℓ_t are i.i.d.). A standard online to batch conversion shows that Theorem 2.1 yields with high probability

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

This improves the bound obtained by [20] who suffers the additional gap

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

2.2 Prediction with expert advice

The framework of prediction with expert advice is widely considered in the literature (see [5] 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 $K \geq 1$ experts predict $\mathbf{f}_t = (f_{1,t}, \dots, f_{K,t}) \in [0, 1]^K$ from the history \mathcal{F}_{t-1} . The learner then chooses a weight vector θ_{t-1} in the simplex Δ_K and produces a prediction $\hat{f}_t := \theta_{t-1}^\top \mathbf{f}_t \in \mathbb{R}$ as a convex combination of the experts. Its performance at time t is evaluated by a loss function $g_t : \mathbb{R} \rightarrow \mathbb{R}$. The goal of the learner is to approach the performance of the best expert on a long run. This can be done by minimizing the average excess risk $R_{k,T} := \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{t-1}[g_t(\hat{f}_t)] - \mathbb{E}_{t-1}[g_t(f_{k,t})]$, 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 $\ell_t : \theta \mapsto g_t(\theta^\top \mathbf{f}_t)$ 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 $\theta_k^\top \mathbf{f}_t = f_{k,t}$, 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:

$$\mathbb{E}_{k \sim \pi} [R_{k,T}] \lesssim \left(\frac{\kappa(\pi, \hat{\pi}_0) + \ln \ln(GT) + x}{\alpha T} \right)^{\frac{1}{2-\beta}}.$$

An important point to note here is 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. We do not assume in this section that there is a single fixed expert $k^* \in \{1, \dots, K\}$ which is always the best, i.e., $\mathbb{E}_{t-1}[g_t(f_{k^*,t})] \leq \min_k \mathbb{E}_{t-1}[g_t(f_{k,t})]$. Hence, we cannot replace (A2) with the closely related Bernstein assumption (see Ass. (A2') or [17, Cond. 1]).

Actually one can reformulate Assumption (A2) on the one dimensional loss functions g_t as follows: there exist $\alpha > 0$ and $\beta \in [0, 1]$ such that for all $t \geq 1$, for all $0 \leq f_1, f_2 \leq 1$,

$$\mathbb{E}_{t-1}[g_t(f_1) - g_t(f_2)] \leq \mathbb{E}_{t-1}[g'_t(f_1)(f_1 - f_2)] - \mathbb{E}_{t-1} \left[\left(\alpha (g'_t(f_1)(f_1 - f_2))^2 \right)^{1/\beta} \right], \quad a.s.$$

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 $\Theta_0 = \{\theta \in \mathbb{R}^d : \|\theta\|_0 = 1, \|\theta\|_1 = 1\}$ the $2d$ corners of the ℓ_1 -ball suffers a slow rate $\mathcal{O}(1/\sqrt{T})$ on the average excess risk with respect to any $\theta \in \mathcal{B}_1$. The goal is to exhibit algorithms that go beyond $\mathcal{O}(1/\sqrt{T})$. In Section 3.1 we investigate non-adaptive discretization grids of the space that yield optimal upper-bounds but suffer exponential time complexity. 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 derive the adaptive points to add to the $2d$ corners in the adversarial case (Section 3.3) and in the i.i.d. case (Section 3.4) in order to obtain two efficient procedures (BOA+ and SABOA respectively) with sparse guarantees.

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 so that $\text{Card}(\mathcal{B}_1(\varepsilon)) \lesssim (1/\varepsilon)^d$. We obtain a nearly optimal regret for this procedure.

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

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

with probability at least $1 - e^{-x}$, $x > 0$.

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: $R_T(\theta) \leq R_T(\tilde{\theta}) + G\varepsilon$ and by applying Theorem 2.1 on $R_T(\tilde{\theta})$. \square

One can improve d to $\|\theta\|_0 \ln d$ by carefully choosing the prior $\hat{\pi}_0$ as in [21]; 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 Oracle bound for arbitrary fixed discretization grid

Let $\Theta_0 \subset \mathcal{B}_1$ be a finite set. The aim of this Section is to study the regret of Algorithm 1 with respect to any $\theta \in \mathcal{B}_1$. Similarly to Proposition 3.1, the average excess risk may be bounded as

$$R_T(\theta) \lesssim \left(\frac{\ln \text{Card}(\Theta_0) + \ln \ln T + x}{\alpha T} \right)^{\frac{1}{2-\beta}} + G\|\theta' - \theta\|_1, \quad (3)$$

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 (3)) that decreases with T . This property will be crucial in obtaining fast rates by adapting the grid Θ_0 sequentially. The regret bound (3) 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 (3). We give the intuition behind this notion in the sketch of the proof of Theorem 3.2.

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

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

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 $D(\theta, \theta') \leq \|\theta - \theta'\|_1 / (1 - \|\theta'\|_1 \wedge \|\theta\|_1)$. 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 $\Theta_0 \subset \mathcal{B}_1$ such that $\{\theta : \|\theta\|_1 = 1, \|\theta\|_0 = 1\} \subseteq \Theta_0$. Let Assumption (A1) and (A2) be satisfied. Then, Algorithm 1 applied with uniform prior $\hat{\pi}_0$ over the elements of Θ_0 and $E = 8G/3$, satisfies with probability $1 - e^{-x}$, $x > 0$,

$$R_T(\theta) \lesssim \left(\frac{a}{\alpha T}\right)^{\frac{1}{2-\beta}} + GD(\theta, \Theta_0) \sqrt{\frac{a}{T}} + \frac{aG}{T}, \quad \theta \in \mathcal{B}_1,$$

where $a = \ln \text{Card}(\Theta_0) + \ln \ln(GT) + x$ 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. We give here the high-level ideas. Let $\theta' \in \Theta_0$ be 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. The regret bound can be decomposed into two terms:

- The first term quantifies the cost of picking the correct $\theta' \in \Theta_0$, 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 excess risk $R_T(\theta)$ is thus of order

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

□

Note that the bound of Theorem 3.2 is *accelerable* as its second term vanishes to zero on the contrary to Inequality (3). 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 the grid Θ_0 , we get an oracle-bound of order $\mathcal{O}(\|\theta\|_1/\sqrt{T})$ for any $\theta \in \mathcal{B}_1$. Moreover a bound of order $R_T(\theta) \leq \mathcal{O}(\|\theta - \theta_k\|_1/(\gamma\sqrt{T}))$ is obtained for all $\theta_k \in \Theta_0$ and $\|\theta\|_1 \leq 1 - \gamma < 1$.

It is worth pointing out that the bound on the gradient G can be substituted with the average gradient observed by the learner. The constant G can be improved to the level of the noise in certain situations with vanishing gradients (see for instance Theorem 3 of [9]).

3.3 Fast-rate sparsity regret bound in the adversarial setting

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 each doubling session $i \geq 0$, BOA+ chooses $\hat{\theta}_t$ 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 :

$$\Theta^{(i)} = \{[\theta_i^*]_k, k = 0, \dots, d\} \cup \{\theta : \|\theta\|_1 = 2, \|\theta\|_0 = 1\},$$

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 approximately as the solution of a strongly convex optimization problem.

Theorem 3.3. Assume the loss functions are μ -strongly convex on $\mathcal{B}_2 := \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 2\}$ with gradients bounded by G in ℓ_∞ -norm on \mathcal{B}_2 . The average regret of BOA+ satisfies the oracle bound

$$R_T(\theta) \leq \tilde{\mathcal{O}} \left(\min \left\{ G \sqrt{\frac{\ln d}{T}}, \sqrt{\frac{\|\theta\|_0}{\mu}} \left(G \sqrt{\frac{\ln d}{T}} \right)^{\frac{3}{2}}, \frac{\sqrt{\|\theta\|_0 d G^2 \ln d}}{\mu T} \right\} + \frac{G^2 \ln d}{\mu T} \right), \quad \theta \in \mathcal{B}_1.$$

The proof is deferred to Appendix C.6. We emphasize that the bound can be rewritten as follows:

$$R_T(\theta) \leq \tilde{\mathcal{O}} \left(\min \left\{ G \sqrt{\frac{\ln d}{T}}, \frac{\|\theta\|_0 G^2 \ln d}{\mu T} \right\} \min \left\{ G \sqrt{\frac{\ln d}{T}}, \frac{d G^2 \ln d}{\mu T} \right\} \right)^{1/2}, \quad \theta \in \mathcal{B}_1 \setminus \{0\}.$$

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 (for non-efficient procedures only) $\tilde{\mathcal{O}}(\|\theta\|_0/T)$. If all θ_i^* are approximately d_0 -sparse it is possible to achieve the optimal rate of order $\tilde{\mathcal{O}}(d_0/T)$, for any $\|\theta\|_0 \leq d_0$. We leave for future work whether it is possible to achieve it in general.

Remark 3.1. The strongly convex assumption on the loss functions can be relaxed (see Inequality (33) in the proof of Theorem 3.3) by assuming (A2) on \mathcal{B}_2 and that there exists $\mu > 0$ and $\beta \in [0, 1]$ such that for all $t \geq 1$ and $\theta \in \mathcal{B}_1$

$$\mu \|\theta - \theta_t^*\|_2^2 \leq \left(\frac{1}{t} \sum_{s=1}^t (\ell_s(\theta) - \ell_s(\theta_t^*)) \right)^\beta, \quad \text{where } \theta_t^* \in \arg \min_{\theta \in \mathcal{B}_1} \sum_{s=1}^t \ell_s(\theta). \quad (4)$$

The rates will depend on β as it is 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 (4) can be restricted in this case. Indeed any $\theta \in \mathcal{B}_1$ satisfies $\|\theta\|_1 \leq \|\theta_t^*\|_1$, which from Lemma 6 of [1] yields $\|\theta - \theta_t^*\|_1 \leq 2\|[\theta - \theta_t^*]_{S_t^*}\|_1$ where $[\theta]_S = (\theta_i \mathbb{1}_{i \in S})_{1 \leq i \leq d}$. One can restrict Assumption (4) to hold on S_t^* only. Such restricted conditions for $\beta = 1$ are common in the sparse learning literature and essentially necessary for the existence of efficient and optimal sparse procedures, see [31]. For obtaining regret bounds on BOA+, the restricted condition (4) with $\beta = 1$ should hold at any time $t \geq 1$, which is unlikely in the regression setting.

3.4 Fast-rate sparse excess risk bound in the i.i.d. setting

In this section, we assume the loss functions ℓ_t to be i.i.d. We provide an algorithm with fast-rate sparsity risk-bound on \mathcal{B}_1 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 the i.i.d. setting, a close inspection of the proof of Theorem 3.4 shows that we can replace Assumption (A2) with the Bernstein condition: there exists $\alpha' > 0$ and $\beta \in [0, 1]$, such that for all $\theta \in \mathcal{B}_1$, all $\theta^* \in \Theta^*$ and all $t \geq 1$,

$$\alpha' \mathbb{E} \left[\left(\nabla \ell_t(\theta)^\top (\theta - \theta^*) \right)^2 \right] \leq \mathbb{E} \left[\nabla \ell_t(\theta)^\top (\theta - \theta^*) \right]^\beta. \quad (\text{A2}')$$

This fast-rate type stochastic condition is equivalent to the *central condition* (see [25, Condition 5.2]) and was already considered to obtain faster rates of convergence for the regret (see [17, Condition 1]).

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

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

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

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

- multiple optima are allowed, which seems to be new when combined with sparsity bounds. An exception is [21] that provides the optimal sparse rate under a low-rank Gram matrix setting for the non-efficient ES algorithm;
- on the contrary to [23] or [9], 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 design dependent in the regression setting; The constant μ corresponds to the smallest non-zero eigenvalue of the covariance matrix while $\alpha = 1/G^2$ for the square loss functions. If $\Theta^* = \{\theta^*\}$ is a singleton than Assumption (A3) implies Assumption (A2') with $\alpha' \geq \mu/G^2$.

Algorithm and excess risk bound Our new procedure called SABOA is described in Algorithm 2. Again it starts from the accelerable bound provided in Theorem 3.2 which is small if one of the points in Θ_0 is close to Θ^* . As BOA+, SABOA restarts BOA by adding current estimators of Θ^* into an updated grid Θ_0 . The new points added to the grid are slightly different between the two algorithms. They are truncated versions of the average of past iterates $\bar{\theta}_{t-1}$ for SABOA and of the leader for BOA+. Remark that restart schemes under Łojasiewicz's Assumption is natural and was already used by [22]. We get the following upper-bound on the average excess risk. The proof that computes the exact constants is postponed to Appendix C.7.

Algorithm 2 SABOA – Sparse Acceleration of BOA

Parameters: $E > 0$

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

For each session $i = 0, \dots$ do:

- Define $\bar{\theta}^{(i-1)} := 0$ if $i = 0$ and $\bar{\theta}^{(i-1)} := 2^{-i+1} \sum_{t=t_{i-1}}^{t_i-1} \hat{\theta}_{t-1}$ otherwise,
- Define $\Theta^{(i)}$ a set of hard-truncated and dilated soft-thresholded versions of $\bar{\theta}^{(i-1)}$ as in (45),
- Denote $K_i := \text{Card}(\Theta^{(i)}) + 2d \leq (i+1)(1 + \ln d) + 3d$,
- 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 $\hat{\pi}_0$.

In other words, for time steps $t = t_i, \dots, t_{i+1} - 1$:

- Choose $\hat{\theta}_{t-1} = \sum_{k=1}^{K_i} \hat{\pi}_{k,t-1} \theta_k$ and observe $\nabla \ell_t(\hat{\theta}_{t-1})$,
- Define component-wise for all $1 \leq k \leq K_i$, denoting $\eta_j := (e^j E)^{-1}$,

$$\hat{\pi}_{k,t} = \frac{\sum_{j=1}^{\ln(ET^2)} \eta_j e^{\eta_j \sum_{s=t_i}^t (r_{k,s} - \eta_j r_{k',s}^2)} \hat{\pi}_{k,0}}{\sum_{j=1}^{\ln(ET^2)} \mathbb{E}_{k' \sim \hat{\pi}_0} [\eta_j e^{\eta_j \sum_{s=t_i}^t (r_{k',s} - \eta_j r_{k',s}^2)}]},$$

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

Theorem 3.4. *Under Assumptions (A1), (A2) and (A3), Algorithm 2 with $E = 4/3G \geq 1$ satisfies with probability at least $1 - e^{-x}$, $x > 0$, the average excess risk bound*

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

where $d_0 = \max_{\theta^* \in \Theta^*} \|\theta^*\|_0$ and $0 \leq \gamma \leq 1$ satisfies $\Theta^* \subseteq \mathcal{B}_{1-\gamma}$.

We conclude with some important remarks about Theorem 3.4. First, we point out that SABOA adapts automatically to unknown parameters $\delta, \beta, \alpha, \mu$ and d_0 to fulfill the rate of Theorem 3.4.

On the radius of ℓ_1 ball. We provide the analysis into \mathcal{B}_1 , the ℓ_1 -ball of radius $U = 1$ only. However, one might need to compare with points into $\mathcal{B}_1(U)$, the ℓ_1 -ball of radius $U > 0$, in order to obtain a good approximation-estimation trade-off. This can be done by rescaling the loss functions $\theta \in \mathcal{B}_1 \mapsto \ell_t(U\theta)$ and applying our results with $UG, U^2\mu$ and α under Assumptions (A1), (A2) and (A3) on $\mathcal{B}_1(U)$. The main rate of convergence of Theorem 3.4 is unchanged. The optimal choice of the radius, if it is not imposed by the application, is left for future research.

Support recovery. When all $\theta^* \in \Theta^*$ lie on the border of the ℓ_1 -ball, they are likely to be sparse. One can relax Assumption (A3) to hold in sup-norm and in a restricted version similar as done in the end of Remark 3.1. In this interesting setting, we could not avoid a factor d_0^2 . The reason is that our sequential algorithm recovers the (largest) support of θ^* (see Configuration 3 of Figure 1) in a framework where the necessary (for the rate $\|\theta^*\|_0$) Irrepressibility Condition [27] does not hold.

Conclusion In this paper, we show that BOA is an optimal online algorithm for aggregating experts under very weak conditions on the loss. Then we aggregate sparse versions of the leader (BOA+) or of the average of BOA's iterates (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 ideal one $\tilde{O}((\|\theta\|_0/T)^{1/(2-\beta)})$ that requires restrictive assumption for efficient algorithm. Our main condition (A3) is weaker and more realistic than the usual ones when seeking for sequential sparse rate bounds for any $t \geq 1$.

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, and M. Wegkamp. 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. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, pages 14–26, 2010.
- [8] D. J. Foster, S. Kale, and H. Karloff. Online sparse linear regression. In *Conference on Learning Theory*, pages 960–970, 2016.
- [9] P. Gaillard and O. Wintenberger. Sparse Accelerated Exponential Weights. In *20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Apr. 2017.
- [10] 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.
- [11] S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *The Journal of Machine Learning Research*, 14(1):729–769, 2013.
- [12] C. Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2014.
- [13] E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.
- [14] 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.
- [15] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.
- [16] W. M. Koolen and T. Van Erven. Second-order quantile methods for experts and combinatorial games. In *COLT*, volume 40, pages 1155–1175, 2015.
- [17] 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.
- [18] J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10(Mar):777–801, 2009.
- [19] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.
- [20] N. A. Mehta. Fast rates with high probability in exp-concave statistical learning. In *Artificial Intelligence and Statistics*, pages 1085–1093, 2017.
- [21] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, pages 731–771, 2011.
- [22] V. Roulet and A. d’Aspremont. Sharpness, restart and acceleration. In *Advances in Neural Information Processing Systems*, pages 1119–1129, 2017.
- [23] J. Steinhardt, S. Wager, and P. Liang. The statistics of streaming sparse regression. *arXiv preprint arXiv:1412.4182*, 2014.
- [24] I. Steinwart and A. Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

- [25] 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.
- [26] V. G. Vovk. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990.
- [27] 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.
- [28] O. Wintenberger. Optimal learning with bernstein online aggregation. *Machine Learning*, 106 (1):119–141, 2017.
- [29] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.
- [30] Y. Yang. Combining forecasting procedures: some theoretical results. *Econometric Theory*, 20 (01):176–222, 2004.
- [31] 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.
- [32] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, pages 87–89, 1963.
- [33] S. Łojasiewicz. Sur la géométrie semi-et sous-analytique. *Annales de l’institut Fourier*, 43(5): 1575–1595, 1993.

Supplementary material

A Sparse oracle inequality by discretizing the space

Inspired by the work of [21], one can improve d to $\|\theta\|_0 \ln d$ in Proposition 3.1 by carefully choosing the prior $\hat{\pi}_0$. To do so, we cover \mathcal{B}_1 by the subspaces

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

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^τ 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 $\hat{\pi}_0$. 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 $\tau \in \{0, 1\}^d$,

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

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

$$R_T(\theta) \lesssim \left(\frac{\|\theta\|_0 \ln(dT/\|\theta\|_0) + x}{\alpha T} \right)^{\frac{1}{2-\beta}} + \frac{\|\theta\|_0 G}{T^2}, \quad (5)$$

by noting that $\binom{d}{\|\theta\|_0} \leq d^{\|\theta\|_0}$ and choosing $\varepsilon = \|\theta\|_0/T^2$. Similar optimal oracle inequalities for mixing arbitrary regressions functions are obtained by Yang [30] and Catoni [4].

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

$$\|\theta''\|_1 = 1 \quad \text{and} \quad \theta'' = \lambda(\theta - \theta') + \theta' \quad (6)$$

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

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

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

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

As the maximum is achieved, the averaging accelerability corresponds to the inverse of λ in the definition (6) of the extrapolation point θ'' .

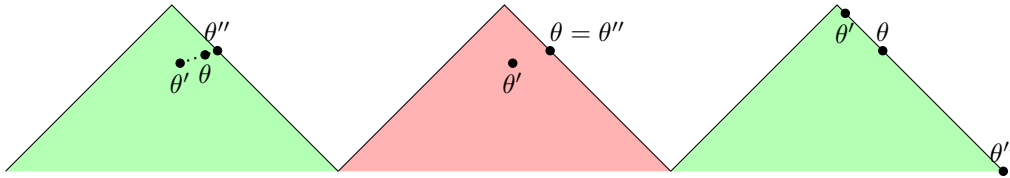


Figure 1: Averaging accelerability for 3 different configurations.

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 [9], 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 $\theta, \theta' \in \mathcal{B}_1$*

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

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 $\|\theta'\|_1 \geq \|\theta\|_1$ and $\text{sign}(\theta'_i) \in \{0, \text{sign}(\theta_i)\}$ for all $1 \leq i \leq d$. Then,*

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

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

We want to emphasize 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 [27]. 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 $S_\varepsilon(x)_i = \text{sign}(x_i)(|x_i| - \varepsilon)_+$ for all $1 \leq i \leq d$. 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 $\|\theta - \theta'\|_\infty \leq \varepsilon$ and $\|\theta\|_0 \leq d_0$. Then, define the dilated soft-threshold*

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

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

- (i) $\|\tilde{\theta}\|_1 \geq \|\theta\|_1$ if $\tilde{\theta} \neq 0$
- (ii) $\text{sign}(\tilde{\theta}_i) \in \{0, \text{sign}(\theta_i)\}$ for all $1 \leq i \leq d$
- (iii) $D(\theta, \tilde{\theta}) \leq 2d_0\varepsilon / \|\theta\|_1$.

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 $\ln(UT)$.

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 [28]⁵. We make more clear the connexion thereafter. We will start our proof with Theorem 3.2 of [28] that we recall now together with the definition of BOA. For each expert $j \in \mathcal{K}$ and each instance $t \geq 1$, from Equation (9) of [28], BOA assigns the weight

$$w_{j,t} := \frac{\exp\left(\eta_j \sum_{s=1}^t r_{j,s} - \eta_j^2 r_{j,s}^2\right) w_{j,0}}{\sum_{k \in \mathcal{K}} \exp\left(\eta_k \sum_{s=1}^t r_{k,s} - \eta_k^2 r_{k,s}^2\right) w_{k,0}} \quad (7)$$

where (η_j) and $w_{j,0}$ are parameters of BOA which respectively correspond to the learning rates and the initial weight associated with each expert; and where $r_{j,t}$ are the instantaneous linearized regrets (denoted $\ell_{j,t}$ in [28]). In our case, $r_{j,t} = \nabla \ell_t(\theta_{t-1})^\top (\theta_{t-1} - \theta_j)$. Now, Theorem 3.2 of [28] states that for any distribution $\tilde{\pi}$ over the set of experts $1 \leq j \leq K$:

$$\sum_{t=1}^T \mathbb{E}_{j \sim \tilde{\pi}} [r_{j,t}] \leq \mathbb{E}_{j \sim \tilde{\pi}} \left[\eta_j \sum_{t=1}^T r_{j,t}^2 + \frac{\ln(\tilde{\pi}_j / w_{j,0})}{\eta_j} + \frac{\ln(\mathbb{E}_{k \sim \pi_0} [\eta_k^{-1}] / \mathbb{E}_{k \sim \tilde{\pi}} [\eta_k^{-1}])}{\eta_j} \right]. \quad (8)$$

There are two main differences between BOA and Algorithm 1.

First, there is a subtle difference in the definition of the weights: we consider the weights $\pi_{j,t} = \eta_j w_{j,t} / \mathbb{E}_{k \sim \pi_t} [\eta_k w_{k,t}]$ instead of $w_{j,t}$. This only impacts the priors (that are multiplied by $\eta_j \mathbb{E}_{k \sim \pi_0} [\eta_k^{-1}]$) and allows to remove the last term in (8) as analyzed in the proof of [28, Theorem 3.2]. For this definition of weights, we thus get:

$$\sum_{t=1}^T \mathbb{E}_{j \sim \tilde{\pi}} [r_{j,t}] \leq \mathbb{E}_{j \sim \tilde{\pi}} \left[\eta_j \sum_{t=1}^T r_{j,t}^2 + \frac{\ln(\tilde{\pi}_j / \pi_{j,0})}{\eta_j} \right]. \quad (9)$$

We refer the reader to the last equation of the proof of Theorem 3.2 of [28] for this inequality.

Second, 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 replica being assigned a different learning rate $\eta_i = e^{-i} E^{-1}$ for $1 \leq i \leq \ln(ET^2)$. Algorithm 1 corresponds to applying BOA on this extended set where each expert k has $\ln(ET^2)$ replica indexed by i whose weights are cumulated into $\hat{\pi}_{k,t}$. The initial weight $\hat{\pi}_{k,0}$ of expert k is uniformly distributed among its $\ln(ET^2)$ replica; each gets the initial weight $\tilde{\pi}_{j,0} = \hat{\pi}_{k,0} / \ln(ET^2)$.

For each parameter θ_k , $k \in \{1, \dots, K\}$, let $1 \leq i_k \leq \ln(ET^2)$ 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 (9) to a specific distribution $\tilde{\pi}$ on the replica. We choose $\tilde{\pi}$ so that it assigns all the mass π_k on the replica (k, i_k) and no mass on the replica (k, i) for $i \neq i_k$. In other words, $\tilde{\pi}_j = \pi_k \mathbb{1}_{i=i_k}$. Then $\ln(\tilde{\pi}_j / \tilde{\pi}_{j,0}) = \ln(\pi_k / \hat{\pi}_{k,0} \ln(ET^2))$ and Inequality (9) entails

$$\begin{aligned} \sum_{t=1}^T \mathbb{E}_{k \sim \pi} [r_{k,t}] &\leq \mathbb{E}_{k \sim \pi} \left[\underbrace{e^{-i_k} E^{-1}}_{:= \lambda_k} \sum_{t=1}^T r_{k,t}^2 + e^{i_k} E (\ln(\pi_k / \hat{\pi}_{k,0}) + \ln \ln(ET^2)) \right] \\ &= \mathbb{E}_{k \sim \pi} \left[\lambda_k \sum_{t=1}^T r_{k,t}^2 + \frac{\ln(\pi_k / \hat{\pi}_{k,0}) + \ln \ln(ET^2)}{\lambda_k} \right], \end{aligned} \quad (10)$$

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 $1 \leq i_k \leq \ln(ET^2)$. 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 (10) are equal. This yields the optimal choice $\lambda_k \approx (V_k / a_k)^{-1/2}$, where $a_k := \ln(\pi_k / \hat{\pi}_{k,0}) + \ln \ln(ET^2)$. However, because of edge effects, this is only possible when $1/(ET)^2 \leq (V_k / a_k)^{-1/2} \leq 1/(Ee)$. We distinguish three cases:

⁵It is also a specific case of Squint of [16] with a discrete distribution over the learning rates

- if $\sqrt{a_k/V_k} > 1/(eE)$: then, we choose $\lambda_k = 1/(eE)$, which yields:

$$\lambda_k V_k + \frac{a_k}{\lambda_k} \leq \frac{2a_k}{\lambda_k} = 2ea_k E \leq 6a_k E$$

- if $1/(ET)^2 \leq (V_k/a_k)^{-1/2} \leq 1/(Ee)$: then, we can choose λ_k such that

$$\frac{\lambda_k}{\sqrt{e}} \leq (V_k/a_k)^{-1/2} \leq \sqrt{e}\lambda_k,$$

which entails $\lambda_k V_k + \frac{a_k}{\lambda_k} \leq 2\sqrt{e}\sqrt{a_k V_k} \leq 4\sqrt{a_k V_k}$

- if $\sqrt{a_k/V_k} < (ET)^{-2}$: then, the choice $\lambda_k = (ET)^{-2}$ gives

$$\lambda_k V_k + \frac{a_k}{\lambda_k} \leq 2\lambda_k V_k = \frac{2V_k}{E^2 T^2} \leq \frac{2}{T},$$

because $r_{k,t}^2 \leq E^2$.

Putting the three cases together and plugging into Inequality (10) yields

$$\sum_{t=1}^T \mathbb{E}_{k \sim \pi} [r_{k,t}] \leq \mathbb{E}_{k \sim \pi} \left[4\sqrt{a_k V_k} + 6a_k E \right] + \frac{2}{T}. \quad (11)$$

We recall Young's inequality.

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

Applying it, with $p = q = 2$, and $a = \sqrt{2\lambda_k V_k}$ and $b = \sqrt{8a_k/\lambda_k}$, we get $4\sqrt{a_k V_k} \leq \lambda_k V_k + 4a_k/\lambda_k$ for any $\lambda_k > 0$. Therefore, substituting into Inequality (11), for any distribution π over $\{1, \dots, K\}$, we have

$$\sum_{t=1}^T \mathbb{E}_{k \sim \pi} [r_{k,t}] \leq \mathbb{E}_{k \sim \pi} \left[\lambda_k V_k + \frac{4a_k}{\lambda_k} + 6a_k E \right] + \frac{2}{T}, \quad (12)$$

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}_π . Using Theorem 4.1 of [28] for $\eta_{j,t} = \lambda_j$ independent of t , we obtain with probability $1 - e^{-x}$ and integrating with respect to π

$$\begin{aligned} \sum_{t=1}^T \mathbb{E}_{t-1} [\mathbb{E}_\pi [r_{k,t}]] &\leq \sum_{t=1}^T \mathbb{E}_\pi [r_{k,t}] + \mathbb{E}_\pi \left[\lambda_k \sum_{t=1}^T r_{k,t}^2 + \frac{x}{\lambda_k} \right] \\ &\stackrel{(12)}{\leq} \mathbb{E}_\pi \left[2\lambda_k \sum_{t=1}^T r_{k,t}^2 + \frac{x + 4a_k}{\lambda_k} + 6a_k E \right] + \frac{2}{T}. \end{aligned} \quad (13)$$

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 [9]): with probability at least $1 - e^{-x}$

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

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

$$\sum_{t=1}^T \mathbb{E}_{t-1} [\mathbb{E}_\pi [r_{k,t}]] \leq \mathbb{E}_\pi \left[4\lambda_k \sum_{t=1}^T \mathbb{E}_{t-1} [r_{k,t}^2] + \frac{9}{2} \lambda_k E^2 x + \frac{4a_k + x}{\lambda_k} + 6a_k E \right] + \frac{2}{T}. \quad (14)$$

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 $t \geq 1$

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

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

$$\begin{aligned} \mathbb{E}_\pi \left[\sum_{t=1}^T \mathbb{E}_{t-1} [\ell_t(\hat{\theta}_{t-1}) - \ell_t(\theta_k)] \right] &\leq \mathbb{E}_\pi \left[\sum_{t=1}^T \mathbb{E}_{t-1} [r_{k,t}] - (\alpha \mathbb{E}_{t-1} [r_{k,t}^2])^{1/\beta} \right] \\ &\leq \mathbb{E}_\pi \left[4\lambda_k \sum_{t=1}^T \mathbb{E}_{t-1} [r_{k,t}^2] - \sum_{t=1}^T (\alpha \mathbb{E}_{t-1} [r_{k,t}^2])^{1/\beta} + \frac{9}{2} \lambda_k E^2 x + \frac{4a_k + x}{\lambda_k} + 6a_k E \right] + \frac{2}{T}. \end{aligned} \quad (15)$$

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

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

Thus,

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

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

$$\begin{aligned} 4\lambda_k \mathbb{E}_{t-1} [r_{k,t}^2] - (\alpha \mathbb{E}_{t-1} [r_{k,t}^2])^{1/\beta} &\leq 4\lambda_k (1-\beta) \gamma^{1/(1-\beta)} \\ &= 4\lambda_k (1-\beta) \left(\frac{(4\lambda_k \beta)^\beta}{\alpha} \right)^{1/(1-\beta)} \\ &= 4(1-\beta) (4\beta)^{\beta/(1-\beta)} \left(\frac{\lambda_k}{\alpha} \right)^{1/(1-\beta)} \\ &\leq 4 \left(\frac{4\lambda_k}{\alpha} \right)^{1/(1-\beta)}. \end{aligned} \quad (16)$$

Summing over t and substituting into Inequality (15), we get

$$\mathbb{E}_\pi \left[\sum_{t=1}^T \mathbb{E}_{t-1} [\ell_t(\hat{\theta}_{t-1}) - \ell_t(\theta_k)] \right] \leq E_\pi \left[\underbrace{4 \left(\frac{4\lambda_k}{\alpha} \right)^{1/(1-\beta)} T + \frac{4a_k + x}{\lambda_k}}_{=: R_k} + \frac{9}{2} \lambda_k E^2 x + 6a_k E \right] + \frac{2}{T}. \quad (17)$$

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

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

We express R_k in terms of λ_k using this identity

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

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

$$\frac{R_T}{T} \leq 4 \left(\frac{4\lambda_k}{\alpha} \right)^{1/(1-\beta)} + \frac{4a_k + x}{T\lambda_k} \leq \frac{(4a_k + x)E}{T}.$$

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

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

Substituting into Inequality (17) and upper-bounding $\lambda_k E^2 \leq E/2$, gives

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

Since $x \mapsto x^{1/(2-\beta)}$ is concave, using Jensen's inequality and replacing $a_k = \ln(\pi_k/\pi_{k,0}) + \ln \ln(ET^2)$ entails,

$$E_\pi \left[\left(\frac{16a_k + 4x}{\alpha T} \right)^{\frac{1}{2-\beta}} \right] \lesssim \left(\frac{\mathbb{E}_\pi[a_k] + x}{\alpha T} \right)^{\frac{1}{2-\beta}} \stackrel{(\text{def of } a_k)}{=} \left(\frac{\mathcal{K}(\pi, \hat{\pi}_0) + \ln \ln(ET^2) + x}{\alpha T} \right)^{\frac{1}{2-\beta}}$$

which 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 (11). We apply it to a Dirac distributions π on $\{1, \dots, K\}$. We get that for any $1 \leq k \leq K$, for any $\lambda_k > 0$,

$$\sum_{t=1}^T r_{k,t} \leq 4 \sqrt{a \sum_{t=1}^T r_{k,t}^2} + 6aE + \frac{2}{T}. \quad (18)$$

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

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

$$\theta^* = (1 - \varepsilon)\theta_{k^*} + \varepsilon\tilde{\theta}. \quad (19)$$

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 (19), there exists a distribution $\pi = (\pi_1, \dots, \pi_K) \in \Delta_K$ such that

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

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

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

and integrating Inequality (18) with respect to π , we obtain

$$\begin{aligned} \sum_{t=1}^T r_t &\leq \mathbb{E}_{k \sim \pi} \left[4 \sqrt{a \sum_{t=1}^T (\nabla \ell_t(\hat{\theta}_{t-1})^\top (\hat{\theta}_{t-1} - \theta^* + \theta^* - \theta_k))^2} \right] + \frac{2}{T} + 6aE \\ &\leq 4 \sqrt{a \sum_{t=1}^T r_t^2} + 4 \mathbb{E}_{k \sim \pi} \left[\sqrt{a \sum_{t=1}^T (\nabla \ell_t(\hat{\theta}_{t-1})^\top (\theta^* - \theta_k))^2} \right] + \frac{2}{T} + 6aE. \quad (20) \end{aligned}$$

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

$$\begin{aligned} &\mathbb{E}_{k \sim \pi} \left[\sqrt{\sum_{t=1}^T (\nabla \ell_t(\hat{\theta}_{t-1})^\top (\theta^* - \theta_k))^2} \right] \\ &\leq \sqrt{\sum_{t=1}^T \|\nabla \ell_t(\hat{\theta}_{t-1})\|_\infty^2 \sum_{k=1}^K \pi_k \|\theta^* - \theta_k\|_1} \\ &\leq \sqrt{\sum_{t=1}^T \|\nabla \ell_t(\hat{\theta}_{t-1})\|_\infty^2 \left(\pi_{k^*} \|\theta^* - \theta_{k^*}\|_1 + (1 - \pi_{k^*}) \max_{1 \leq k \leq K} \|\theta^* - \theta_k\|_1 \right)} \\ &\leq \sqrt{\sum_{t=1}^T \|\nabla \ell_t(\hat{\theta}_{t-1})\|_\infty^2 \left(\pi_{k^*} \|\theta^* - \theta_{k^*}\|_1 + 2(1 - \pi_{k^*}) \right)}, \quad (21) \end{aligned}$$

where the last inequality is because $\|\theta^* - \theta_k\|_1 \leq \|\theta_k\|_1 + \|\theta^*\|_1 \leq 2$. We also have from the definition of θ^* (see before (19))

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

Therefore, substituting into (21) we get

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

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

Therefore, substituting into Inequality (20), we have

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

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

$$\sum_{t=1}^T r_t \leq \lambda \sum_{t=1}^T r_t^2 + \underbrace{\frac{4a}{\lambda} + 16\varepsilon \bar{G}_T \sqrt{aT} + \frac{2}{T} + 6aE}_{=:z}. \quad (22)$$

Now, we recognize an inequality similar to Inequality (12). 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 (12)

$$\begin{aligned} \sum_{t=1}^T \mathbb{E}_{t-1}[r_t] &\stackrel{\text{Thm 4.1 of [28]}}{\leq} \sum_{t=1}^T r_t + \lambda \sum_{t=1}^T r_t^2 + \frac{x}{\lambda} \\ &\stackrel{(22)}{\leq} 2\lambda \sum_{t=1}^T r_t^2 + \frac{4a+x}{\lambda} + z \\ &\stackrel{\text{Thm 9 of [9]}}{\leq} 4\lambda \sum_{t=1}^T \mathbb{E}_{t-1}[r_t^2] + \frac{4a+x}{\lambda} + \frac{9}{2}\lambda E^2 x + z. \end{aligned} \quad (23)$$

Using Assumption (A2) then yields

$$\begin{aligned} \sum_{t=1}^T \mathbb{E}_{t-1}[\ell_t(\hat{\theta}_{t-1}) - \ell_t(\theta^*)] &\leq \sum_{t=1}^T \mathbb{E}_{t-1}[r_t] - (\alpha \mathbb{E}_{t-1}[r_t^2])^{1/\beta} \\ &\stackrel{(23)}{\leq} 4\lambda \sum_{t=1}^T \mathbb{E}_{t-1}[r_t^2] - (\alpha \mathbb{E}_{t-1}[r_t^2])^{1/\beta} + \frac{4a+x}{\lambda} + \frac{9}{2}\lambda E^2 x + z \\ &\stackrel{(16)}{\leq} 4\left(\frac{4\lambda}{\alpha}\right)^{1/(1-\beta)} + \frac{4a+x}{\lambda} + \frac{9}{2}\lambda E^2 x + z. \end{aligned}$$

This yields an inequality similar to Inequality (17). Optimizing in $\lambda > 0$, as we did for Inequality (17) 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

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

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

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{t-1} [\ell_t(\hat{\theta}_{t-1}) - \ell_t(\theta^*)] \leq 4 \left(\frac{16a + 4x}{\alpha T} \right)^{\frac{1}{2-\beta}} + \frac{(10a + 4x)E}{T} + 16\varepsilon \bar{G}_T \sqrt{\frac{a}{T} + \frac{2}{T^2}}. \quad (24)$$

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

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

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

C.4 Proof of Lemma B.2

Denote $\pi := 1 - \min_{1 \leq i \leq d} |\theta_i|/|\theta'_i|$. Then, for any $1 \leq i \leq d$, $|\theta_i| \geq (1 - \pi)|\theta'_i|$. Because θ'_i and θ_i have same signs, this yields $|\theta_i - (1 - \pi)\theta'_i| = |\theta_i| - (1 - \pi)|\theta'_i|$ for all $1 \leq i \leq d$. Summing over $i = 1, \dots, d$, entails

$$\begin{aligned} \|\theta - (1 - \pi)\theta'\|_1 &= \sum_{i=1}^d |\theta_i - (1 - \pi)\theta'_i| = \sum_{i=1}^d |\theta_i| - (1 - \pi)|\theta'_i| \\ &= \|\theta\|_1 - (1 - \pi)\|\theta'\|_1 \stackrel{\|\theta'\|_1 \geq \|\theta\|_1}{\leq} \pi\|\theta\|_1 \leq \pi. \end{aligned} \quad (25)$$

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

$$1 - \frac{|\theta_i|}{|\theta'_i|} = \frac{|\theta'_i| - |\theta_i|}{|\theta'_i|} \stackrel{|\theta'_i| \geq |\theta_i|}{\leq} \frac{|\theta'_i| - |\theta_i|}{|\theta'_i|} \stackrel{|\theta'_i| \geq |\theta_i|}{\leq} \frac{|\theta'_i| - |\theta_i|}{|\theta_i|} \leq \frac{\|\theta' - \theta\|_\infty}{\Delta},$$

which concludes the proof of the Lemma.

C.5 Proof of Lemma B.3

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

$$\|S_\varepsilon(\theta')\|_1 \geq \sum_{i \in \text{Supp}(\theta)} |S_\varepsilon(\theta')_i| \geq \sum_{i \in \text{Supp}(\theta)} (|\theta'_i| - \varepsilon) \geq \sum_{i \in \text{Supp}(\theta)} (|\theta_i| - 2\varepsilon) \geq \|\theta\|_1 - 2d_0\varepsilon. \quad (26)$$

If $S_\varepsilon(\theta') = 0$, then $\|\tilde{\theta}\|_1 = 0$ and $\|\theta\|_1 \leq 2d_0\varepsilon$ so that $D(\theta, \tilde{\theta}) \leq 1 \leq 2d_0\varepsilon/\|\theta\|_1$. Therefore, we can assume from now that $S_\varepsilon(\theta') \neq 0$. By definition of $\tilde{\theta}$, Inequality (26) yields $\|\tilde{\theta}\|_1 = (\|S_\varepsilon(\theta')\|_1 + 2d_0\varepsilon) \wedge 1 \geq \|\theta\|_1$. Then $\tilde{\theta}$ satisfies the assumptions of Lemma B.2, which we can apply

$$D(\theta, \tilde{\theta}) \leq 1 - \min_{1 \leq i \leq d} \frac{|\theta_i|}{|\theta'_i|} = \max_{i \in \text{Supp}(\tilde{\theta})} \frac{|\tilde{\theta}_i| - |\theta_i|}{|\tilde{\theta}_i|}. \quad (27)$$

We consider two cases:

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

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

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

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

Substituting into Inequality (27) concludes this case.

- Otherwise $\|S_\varepsilon(\theta')\|_1 \leq 1 - 2d_0\varepsilon$ and for $i \in \text{Supp}(\tilde{\theta}) = \text{Supp}(S_\varepsilon(\theta'))$

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

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

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

Substituting the obtained bounds in each cases into Inequality (27) concludes the proof.

C.6 Proof of Theorem 3.3

We perform the proof for $\theta \in \mathcal{B}_{1/2}$ only. However, optimization on \mathcal{B}_1 can be obtained by renormalizing the loss functions considering $\ell_t(2\theta)$ instead of ℓ_t . We leave this generalization to the reader. Let $\theta \in \mathcal{B}_{1/2}$ and denote $d_0 = \|\theta\|_0$. For simplicity, we also assume that $T = 2^I - 1$ and $d_0 \neq 0$.

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 $i \geq 0$. Denote by $\theta_1, \dots, \theta_{3d+1}$ the $3d+1$ elements of $\Theta^{(i)}$. For any distribution $\pi \in \Delta_{3d+1}$ over $\Theta^{(i)}$, we have from Inequality (11):

$$\sum_{t=t_i}^{t_{i+1}-1} \mathbb{E}_{k \sim \pi} [r_{k,t}] \leq \mathbb{E}_{k \sim \pi} \left[4\sqrt{a_k V_k} + 6a_k E \right] + \frac{2}{T}. \quad (28)$$

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

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

Therefore, Inequality (28) yields

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

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

$$\text{Reg}_j(\theta) := \sum_{t=1}^{t_j-1} \ell_t(\hat{\theta}_{t-1}) - \ell_t(\theta) \leq 4G\sqrt{a} \sum_{i=0}^{j-1} 2^{i/2} + \underbrace{6Eaj + \frac{2j}{T}}_{=:z} \leq 10G\sqrt{a}2^{j/2} + z, \quad (29)$$

where we recall $a = \ln(3d+1) + \ln \ln(ET^2)$. In particular for $j = I \lesssim \ln T$ we obtain the first inequality stated by the theorem:

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

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 $j \geq 0$, we have

$$\text{Reg}_j(\theta) \leq 48 \frac{ad_0 c^2 G^2}{\mu} + j \frac{caG^2}{\mu} + 16\sqrt{5}c \sqrt{\frac{d_0 (G\sqrt{a})^3}{\mu}} \sum_{k=0}^j 2^{-\frac{3k}{4}}. \quad (30)$$

Indeed, decomposing the cumulative regret, we have

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

Note that Assumption (A2) is satisfied with $\beta = 1$, $\alpha = \mu/(2G^2)$ and without the expectation \mathbb{E}_t . It is worth pointing out that the transformation of the second-order term into a cumulative risk performed in (23) was not needed here since Assumption (A2) holds on the loss functions 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 there exists some constant $c > 0$ such that

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

with $a = \ln(3d+1) + \ln \ln(ET^2)$. Replacing into the preceding inequality, it yields

$$\text{Reg}_{j+1}(\theta) \leq \text{Reg}_j(\theta) + cGD(\theta, [\theta_j^*]_{d_0}) \sqrt{a2^j} + \frac{caG^2}{\mu} \quad (31)$$

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

$$\begin{aligned} D(\theta, [\theta_j^*]_{d_0}) &\stackrel{(\text{Lem. B.1})}{\leq} 2 \|\theta - [\theta_j^*]_{d_0}\|_1 \stackrel{\|\theta\|_0 = \|\theta_j^*\|_{d_0} = d_0}{\leq} 2\sqrt{2d_0} \|\theta - [\theta_j^*]_{d_0}\|_2 \\ &\leq 2\sqrt{2d_0} (\|\theta - \theta_j^*\|_2 + \|\theta_j^* - [\theta_j^*]_{d_0}\|_2). \end{aligned}$$

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

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

Therefore, plugging into the previous inequality

$$D(\theta, [\theta_j^*]_{d_0}) \leq 4\sqrt{2d_0} \|\theta - \theta_j^*\|_2. \quad (32)$$

But because the loss functions 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 $\theta \in \mathcal{B}_{1/2}$

$$\begin{aligned} \mu \|\theta - \theta_j^*\|_2^2 &\leq \frac{1}{2^j - 1} \sum_{t=1}^{t_j-1} \ell_t(\theta) - \ell_t(\theta_j^*) \\ &= \frac{\text{Reg}_j(\theta_j^*) - \text{Reg}_j(\theta)}{2^j - 1} \leq \frac{\text{Reg}_j(\theta_j^*) - \text{Reg}_j(\theta)}{2^{j-1}}. \end{aligned} \quad (33)$$

Thus, from Inequality (32), we obtain

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

Plugging into Inequality (31) gives

$$\text{Reg}_{j+1}(\theta) \leq \text{Reg}_j(\theta) + 8cG\sqrt{\frac{ad_0}{\mu}(\text{Reg}_j(\theta_j^*) - \text{Reg}_j(\theta))} + \frac{caG^2}{\mu}. \quad (34)$$

We can upper-bound $\text{Reg}_j(\theta_j^*)$ using Inequality (29). This entails

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

Now we have an inequality of the form

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

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

$$\begin{aligned} \text{Reg}_{j+1}(\theta) &\leq \max\{3x_1^2/4, (\text{Reg}_j(\theta))_+ + x_1\sqrt{x_2}\} + x_3. \\ &= \max\left\{48\frac{ad_0c^2G^2}{\mu}, (\text{Reg}_j(\theta))_+ + 16\sqrt{5}c\sqrt{\frac{d_0}{\mu}}\left(G\sqrt{\frac{a}{2^j}}\right)^{3/2}\right\} + \frac{caG^2}{\mu}. \end{aligned} \quad (35)$$

This concludes the induction, using the hypothesis (30). In particular, considering $j = I = \ln_2(T-1)$, we proved that

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

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: there exists $c_0 > 0$ such that for any $\theta \in \mathcal{B}_{1/2}$

$$\text{Reg}_j(\theta) \leq j\frac{ac_0\sqrt{\|\theta\|_0}dG^2}{\mu T}$$

where $a = \ln(3d+1) + \ln \ln(ET^2)$. We start from Inequality (34) obtained in Part 2:

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

Now, instead of upper-bounding $\text{Reg}_j(\theta_j^*)$ using Inequality (29), we use the induction hypothesis itself. Since θ_j^* is not necessarily sparse, we have

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

which entails

$$\text{Reg}_{j+1}(\theta) \leq \text{Reg}_j(\theta) + 8cG\sqrt{\frac{ad_0}{\mu}\left(j\frac{ac_0dG^2}{\mu} - \text{Reg}_j(\theta)\right)} + \frac{caG^2}{\mu}.$$

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

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

with $x_1 = 8cG\sqrt{ad_0/\mu}$, $x_2 = (jac_0dG^2)/\mu$ and $x_3 = caG^2/\mu$. Similarly to Inequality (35), we have

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

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

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

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 $i \geq 0$ that with probability at least $1 - ie^{-x}$, there exists $\theta^* \in \Theta^*$ such that

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

where D is defined in Definition 3.1,

$$\varepsilon := \max_{\theta^* \in \Theta^*} \left((8\sqrt{a}G)^\beta \max \left\{ \frac{2}{\alpha G^2}, \frac{8\|\theta^*\|_0}{\mu} \min \left\{ \frac{8\|\theta^*\|_0}{\|\theta^*\|_1^2}, \frac{1}{(1 - \|\theta^*\|_1)^2} \right\} \right\} \right)^{\frac{1}{2-\beta}}, \quad (36)$$

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 $D(\theta^*, \{0\}) \leq \|\theta^*\|_1 \leq 1$. The initialization thus holds true as soon as $\varepsilon > 1$.

Induction step. Let $i \geq 0$ and assume (\mathcal{H}_i) . We start from Theorem 3.2 (see Inequality (24) for the precise constants that we upper-bound here) that we apply for $t = t_{i-1}, \dots, t_i - 1$ and $\theta^* \in \Theta^*$ satisfying (\mathcal{H}_i) : with probability $1 - e^{-x}$

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

where for simplicity of notation we define $a := 16(1 + \ln(K_i)) + 16 \ln \ln(ET^2) + 4x$, where $K_i := \text{Card}(\Theta^{(i)}) + 2d$ 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 $\bar{\theta}^{(i)} := 2^{-i+1} \sum_{t=t_{i-1}}^{t_i-1} \hat{\theta}_{t-1}$, we obtain

$$\mathbb{E}[\ell_t(\bar{\theta}^{(i)}) - \ell_t(\theta^*)] \leq 2\sqrt{a}G\varepsilon 2^{-(\frac{1}{2}+\tau)i} + 4 \left(\frac{a}{\alpha} \right)^{\frac{1}{2-\beta}} 2^{-\frac{i}{2-\beta}} + aE2^{1-i} + 2^{3-2i}. \quad (37)$$

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 $a \geq 16$ and $E \geq 1$, we have $16 \leq aE$ and thus $2^{3-2i} \leq aE2^{1-i}$. Then, because $\varepsilon \geq \sqrt{a}$, $aE \leq \sqrt{a}E\varepsilon = \frac{4}{3}\sqrt{a}\varepsilon G$ and thus

$$aE2^{1-i} + 2^{3-2i} \leq \frac{3}{2}aE2^{-i} \leq 2\sqrt{a}\varepsilon G2^{-i} \stackrel{\tau \leq 1/2}{\leq} 2\sqrt{a}\varepsilon G2^{-(\frac{1}{2}+\tau)i}. \quad (38)$$

The second term is also dominated thanks to the definition of ε in (36)

$$\varepsilon \stackrel{(36)}{\geq} \left(\frac{2(\sqrt{8a}G)^\beta}{\alpha G^2} \right)^{\frac{1}{2-\beta}} \Rightarrow 2\sqrt{2a}G\varepsilon \geq \left(\frac{16a}{\alpha} \right)^{\frac{1}{2-\beta}} \stackrel{0 \leq \beta \leq 1}{\geq} 4 \left(\frac{a}{\alpha} \right)^{\frac{1}{2-\beta}}$$

and

$$\tau \stackrel{(36)}{=} \frac{1}{2-\beta} - \frac{1}{2} \Rightarrow \frac{1}{2-\beta} \geq \frac{1}{2} + \tau$$

which yields

$$4 \left(\frac{a}{\alpha} \right)^{\frac{1}{2-\beta}} 2^{-\frac{i}{2-\beta}} \leq 2\sqrt{2a}G\varepsilon 2^{-(\frac{1}{2}+\tau)i}. \quad (39)$$

Thus replacing Inequalities (38) and (39) into Inequality (37) and upper-bounding $4\sqrt{2} + 2 \leq 8$, we get for any $\theta^* \in \Theta^*$

$$\mathbb{E}[\ell_t(\bar{\theta}^{(i)}) - \ell_t(\theta^*)] \leq 8\sqrt{a}G\varepsilon 2^{-(\frac{1}{2}+\tau)i}. \quad (40)$$

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

$$\|\bar{\theta}^{(i)} - \theta^*\|_\infty \leq \|\bar{\theta}^{(i)} - \theta^*\|_2 \stackrel{(40)+(A3)}{\leq} (8\sqrt{a}G\varepsilon)^{\frac{\beta}{2}} \mu^{-\frac{1}{2}} 2^{-(\frac{1}{2}+\tau)\frac{\beta}{2}i} =: \varepsilon'. \quad (41)$$

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 $\|[\bar{\theta}^{(i)}]_{d_0} - \theta^*\|_0 \leq 2d_0$ and

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

Applying Lemma B.1, we get

$$D(\theta^*, [\bar{\theta}^{(i)}]_{d_0}) \leq \frac{\|[\bar{\theta}^{(i)}]_{d_0} - \theta^*\|_1}{1 - \|\theta^*\|_1} \leq \frac{2\sqrt{2d_0}\varepsilon'}{1 - \|\theta^*\|_1}. \quad (42)$$

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:

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

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

$$\tilde{\theta}(\tilde{\varepsilon}, \tilde{d}_0) := S_{\tilde{\varepsilon}}(\bar{\theta}^{(i)}) \left(1 + \frac{2\tilde{d}_0\tilde{\varepsilon}}{\|S_{\tilde{\varepsilon}}(\bar{\theta}^{(i)})\|_1} \right) \wedge \frac{1}{\|S_{\tilde{\varepsilon}}(\bar{\theta}^{(i)})\|_1}, \quad (43)$$

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

$$D(\theta^*, \tilde{\theta}(\tilde{\varepsilon}, \tilde{d}_0)) \leq \frac{2\tilde{d}_0\tilde{\varepsilon}}{\|\theta^*\|_1} \leq \frac{8d_0\varepsilon'}{\|\theta^*\|_1}. \quad (44)$$

We define the new approximation grid

$$\Theta^{(i+1)} := \{\tilde{\theta}(\tilde{\varepsilon}, \tilde{d}_0), \tilde{\varepsilon} \in \mathcal{G}_{\varepsilon'}, \tilde{d}_0 \in \mathcal{G}_{d_0}\} \cup \{[\bar{\theta}^{(i)}]_{\tilde{d}_0}, \quad \tilde{d}_0 = 1, \dots, d\}, \quad (45)$$

where $\tilde{\theta}(\tilde{\varepsilon}, \tilde{d}_0)$ is defined in Equation (43) and $[\cdot]_k$ are hard-truncations to k coordinates. We get from Inequality (42) and (44) that

$$\begin{aligned} D(\theta^*, \Theta^{(i+1)}) &\leq \min \left\{ \frac{\sqrt{8d_0}}{1 - \|\theta^*\|_1}, \frac{8d_0}{\|\theta^*\|_1} \right\} \varepsilon' \\ &\stackrel{(41)}{=} (8\sqrt{a}G\varepsilon)^{\frac{\beta}{2}} \mu^{-\frac{1}{2}} \min \left\{ \frac{\sqrt{8d_0}}{1 - \|\theta^*\|_1}, \frac{8d_0}{\|\theta^*\|_1} \right\} 2^{-(\frac{1}{2}+\tau)\frac{\beta}{2}i}. \end{aligned}$$

To conclude the induction, it suffices to show that this is smaller than $\varepsilon 2^{-\tau(i+1)}$. Our choices of ε and τ defined in (36) was done in that purpose, so that the induction is completed.

Conclusion. Substituting the values of ε and τ into Inequality (40) and using the choice $i = \ln_2 T$ (which upper-bound the number of sessions after T times steps) concludes the proof:

$$\begin{aligned} \mathbb{E}[\ell_t(\bar{\theta}^{(i)}) - \ell_t(\theta^*)] &\stackrel{\text{Jensen}}{\leq} \frac{R_T^{(i)}}{2^i} \\ &\stackrel{(40)}{\leq} 8\sqrt{a}G\varepsilon 2^{-(\frac{1}{2}+\tau)i} \\ &\stackrel{(36)}{\leq} \max_{\theta^* \in \Theta^*} \left(\frac{128a}{T} \max \left\{ \frac{1}{\alpha}, \frac{4G^2\|\theta^*\|_0}{\mu} \min \left\{ \frac{1}{(1 - \|\theta^*\|_1)^2}, \frac{8\|\theta^*\|_0}{\|\theta^*\|_1^2} \right\} \right\} \right)^{\frac{1}{2-\beta}}, \end{aligned}$$

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