Sparse Accelerated Exponential Weights

Pierre Gaillard, Olivier Wintenberger

Sparse Accelerated Exponential Weights

Pierre Gaillard*[1] and Olivier Wintenberger†[1,2]

[1]University of Copenhagen, Denmark
[2]Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France

October 11, 2016

Abstract

We consider the stochastic optimization problem where a convex function is minimized observing recursively the gradients. We introduce SAEW, a new procedure that accelerates exponential weights procedures with the slow rate $1/\sqrt{T}$ to procedures achieving the fast rate $1/T$. Under the strong convexity of the risk, we achieve the optimal rate of convergence for approximating sparse parameters in $\mathbb{R}^d$. The acceleration is achieved by using successive averaging steps in an online fashion. The procedure also produces sparse estimators thanks to additional hard threshold steps.

1 Introduction

Stochastic optimization procedures have encountered more and more success in the past few years. This common framework includes machine learning methods minimizing the empirical risk. LeCun and Bottou [2004] emphasized the utility of Stochastic Gradient Descent (SGD) procedures compared with batch procedures; the lack of accuracy in the optimization is balanced by the robustness of the procedure to any random environment. Zinkevich [2003] formalized this robustness property by proving a $d/\sqrt{T}$ rate of convergence in any possible convex environment for a d-dimensional parametric bounded space. This rate is optimal with no additional condition. However, under strong convexity of the risk, accelerated SGD procedures achieve the fast rate d/T, that is also optimal Agarwal et al. [2012]. One of the most popular acceleration procedure is obtained by a simple averaging step, see Polyak and Juditsky [1992] and Bach and Moulines [2013]. Other robust and adaptive procedures using exponential weights have been studied in the setting of individual sequences by Cesa-Bianchi and Lugosi [2006]. The link with the stochastic optimization problem has been done in Kivinen and Warmuth [1997], providing in the ℓ_1-ball algorithms with an optimal logarithmic dependence on the dimension d but a slow rate $1/\sqrt{T}$. The fast rate $\log(T)$ on the regret has been achieved in some strongly convex cases as in Theorem 3.3 of Cesa-Bianchi and Lugosi [2006]. Thus, the expectation of the risk of their averaging, studied under the name of progressive mixture rule by Catoni [2004], also achieves the fast rate $\log(T)/T$. However, progressive mixture rules do not achieve the fast rate with high probability, see Audibert [2008] and their complexity is prohibitive (of order d^T). The aim of this paper is to propose an efficient acceleration of exponential weights procedures that achieves the fast rate $1/T$ with high probability.

In parallel, optimal rates of convergence for the risk were provided by Bunea et al. [2007] in the sparse setting. When the optimal parameter θ^* is of dimension $d_0 = \|\theta^*\|_0$ smaller than the dimension of the parametric space d, the optimal rate of convergence is $d_0 \log(d)/T$. Such fast rates can be achieved for polynomial time algorithm only up to the multiplicative factor α^{-1} where α is the strong convexity constant of the risk, see Zhang et al. [2014]. For instance, the Lasso procedure achieves this optimal rate for least square linear regression, see Assumption (A3) (implying strong convexity of the risk) of Bunea et al. [2007]. Other more robust optimal batch procedures such as ℓ_0 penalization or exploration of the parametric space suffer serious complexity drawbacks and are known to be NP-hard. Most of the stochastic algorithms do not match this rate, with the exception of SeqSEW (in expectation only), see Gerchinovitz [2013]. As the strong convexity constant α does not appear in the bounds of

*pierre@gaillard.me
†wintenberger@math.ku.dk

Gerchinovitz [2013], one suspects that the algorithm is NP-hard.

The aim of this paper is to provide the first acceleration of exponential weights procedures achieving the optimal rate of convergence $d_0 \log(d)/(\alpha T)$ in the identically and independently distributed (i.i.d.) online optimization setting with sparse solution θ^*. The acceleration is obtained by localizing the exponential weights around their averages in an online fashion. The idea is that the averaging alone suffers too much from the exploration of the entire parameter space. The sparsity is achieved by an additional hard-truncation step, producing sparse approximations of the optimal parameter θ^*. The acceleration procedure is not computationally hard as its complexity is $\mathcal{O}(dT)$. We obtain theoretical optimal bounds on the risk similar to the Lasso for random design, see Bunea et al. [2007]. We also obtain optimal bounds on the cumulative risk of the exploration of the parameter space.

The paper is organized as follows. After some preliminaries in Section 2, we present our acceleration procedure and we prove that it achieves the optimal rate of convergence in Section 3. We refine the constants for least square linear regression in Section 4. Finally, we give some simulations in Section 5.

2 Preliminaries

We consider a sequence $\ell_t : \mathbb{R}^d \to \mathbb{R}, t \geqslant 1$ of i.i.d. random loss functions. We define the instantaneous risk as $\mathbb{E}[\ell_t] : \theta \mapsto \mathbb{E}[\ell_t(\theta)]$[†]. We assume that the risk is (2α)-strongly convex, i.e., for all $\theta_1, \theta_2 \in \mathbb{R}^d$

$$\mathbb{E}\big[\ell_t(\theta_1) - \ell_t(\theta_2)\big] \leqslant \mathbb{E}\big[\nabla \ell_t(\theta_1)\big]^\top (\theta_1 - \theta_2) - \alpha \big\|\theta_1 - \theta_2\big\|_2^2 . \quad \text{(SC)}$$

The (unique) risk minimizer in $\mathbb{R}^d$ is denoted θ^* and its effective dimension is $\|\theta^*\|_0 \leqslant d_0$. We insist on the fact that the strong convexity is only required on the risk and not on the loss function. This condition is satisfied for many non strongly convex loss functions such as the quantile loss (see Section 5) and necessary to obtain fast rates of convergence (see Agarwal et al. [2012]).

[†]Because the losses are i.i.d, the risk does not depend on $t \geqslant 1$. However, we still use the time index in the notation to emphasize that a quantity indexed by $s \geqslant 1$ cannot depend on ℓ_t for any $t > s$. The notation $\mathbb{E}[\ell_t](\widehat{\theta}_{t-1})$ denotes $\mathbb{E}\big[\ell_t(\widehat{\theta}_{t-1}) \big| \ell_1, \dots, \ell_{t-1}\big]$.

Online optimization setting For each $t \geqslant 1$, we provide two parameters $(\widehat{\theta}_{t-1}, \tilde{\theta}_{t-1}) \in \mathbb{R}^d \times \mathbb{R}^d$ having observed the past gradients of the first parameter $\nabla \ell_s(\widehat{\theta}_{s-1}) \in \mathbb{R}^d$ for $s \leqslant t-1$ only.

Our aim is to provide high-probability upper-bounds on the cumulative excess risk (also called cumulative risk for simplicity) of the sequence $(\widehat{\theta}_{t-1})$ and on the instantaneous excess risk of $\tilde{\theta}_{t-1}$:

- *Cumulative risk:* the online exploration vs. exploitation problem aims at minimizing the cumulative risk of the sequence $(\widehat{\theta}_{t-1})$ defined as

$$\mathrm{Risk}_{1:T}(\widehat{\theta}_{0:(T-1)}) := \sum_{t=1}^{T} \mathrm{Risk}(\widehat{\theta}_{t-1}) , \qquad (1)$$

  where $\mathrm{Risk}(\theta) := \mathbb{E}[\ell_t](\theta) - \mathbb{E}[\ell_t](\theta^*)$ is the instantaneous excess risk. This goal is useful in a predictive scenario when the observation of $\nabla \ell_t(\widehat{\theta}_{t-1})$ comes at the cost of $\mathrm{Risk}(\widehat{\theta}_{t-1})$.

- *Instantaneous excess risk*: simultaneously, at any time $t \geqslant 1$, we provide an estimator $\tilde{\theta}_{t-1}$ of θ^* that minimizes the instantaneous risk. This problem has been widely studied in statistics and the known solutions are mostly batch algorithms. Under the strong convexity of the risk, a small instantaneous risk ensures in particular that $\tilde{\theta}_{t-1}$ is close in ℓ_2-norm to the true parameter θ^* (by Lemma 5, Appendix A.1).

To make a parallel with the multi-armed bandit setting, minimizing the cumulative risk is related to minimizing the cumulative regret. In contrast, the second goal is related to simple regret (see Bubeck et al. [2009]): the cost of exploration only comes in terms of resources (time steps T) rather than of costs depending on the exploration.

By convexity of the risk, the averaging $\bar{\theta}_{T-1} := (1/T)\sum_{t=1}^{T} \widehat{\theta}_{t-1}$ has an instantaneous risk upper-bounded by the cumulative risk

$$\mathrm{Risk}(\bar{\theta}_{T-1}) \leqslant \mathrm{Risk}_{1:T}(\widehat{\theta}_{0:(T-1)})/T . \qquad (2)$$

Therefore, upper bounds on the cumulative risk lead to upper bounds on the instantaneous risk for $\tilde{\theta}_{T-1} = \bar{\theta}_{T-1}$. However, we will provide another solution to build $\tilde{\theta}_{T-1}$ with better guarantees than the one obtained by (2).

On the contrary, since each $\tilde{\theta}_{t-1}$ minimizes the instantaneous risk at time t, it is tempting to use them in the exploration vs. exploitation problem. However, it is impossible in our setting as the parameters $(\tilde{\theta}_t)$

are constructed upon the observation of the gradients $\nabla \ell_s(\widehat{\theta}_{s-1})$, $s < t$. Remark that our bounds on the cumulative risk are optimal as of the same order than $\sum_{t=1}^{T} \mathrm{Risk}(\tilde{\theta}_{t-1})$.

Our main contribution (see Theorems 1 and 2) is to introduce a new acceleration procedure that simultaneously ensures (up to loglog terms) both optimal risk for $\tilde{\theta}_{t-1}$ and optimal cumulative risk for $(\widehat{\theta}_{t-1})$. Up to our knowledge, this is the first polynomial time online procedure that recovers the minimax rate obtained in a sparse strongly convex setting. Its instantaneous risk achieves the optimal rate of convergence

$$\min\left\{\frac{B^2 d_0 \log(d)}{\alpha T}, UB\sqrt{\frac{\log(d)}{T}}\right\}, \tag{3}$$

where $B \geqslant \sup_{\theta:\|\theta\|_1 \leqslant 2U} \|\nabla \ell_t(\theta)\|_\infty$ is an almost sure bound on the gradients,

$$\big\|\theta^*\big\|_1 \leqslant U \quad \text{and} \quad \big\|\theta^*\big\|_0 \leqslant d_0\,. \tag{4}$$

For least square linear regression (see Theorem 3), B^2 is replaced in (3) with a term of order $\sigma^2 := \mathbb{E}[\ell_t(\theta^*)]$. In the batch setting, the Lasso achieves a similar rate under the slightly stronger Assumption (A3) of Bunea et al. [2007].

3 Acceleration procedure for known parameters

We propose SAEW (described in Algorithm 2) that depends on the parameters (d_0, α, U, B) and performs an optimal online optimization in the ℓ_1 ball of radius U. SAEW accelerates a convex optimization subroutine (see Algorithm 1). If the latter achieves a slow rate of convergence on its cumulative regret, SAEW achieves a fast rate of convergence on its cumulative and instantaneous risks. We describe first what is expected from the subroutine.

3.1 Convex optimization in the ℓ_1-ball with a slow rate of convergence

Assume that a generic subroutine (Algorithm 1), denoted by $\mathcal{S}$, performs online convex optimization into the ℓ_1-ball $\mathcal{B}_1\big(\theta_{\mathrm{center}}, \varepsilon\big) := \big\{\theta \in \mathbb{R}^d : \|\theta - \theta_{\mathrm{center}}\|_1 \leqslant \varepsilon\big\}$ of center $\theta_{\mathrm{center}} \in \mathbb{R}^d$ and radius $\varepsilon > 0$. Centers and radii will be settled online thanks to SAEW.

We assume that the subroutine $\mathcal{S}$ applied on any sequence of convex sub-differentiable losses $(\ell_t)_{t \geqslant t_{\mathrm{start}}}$

Algorithm 1: Subroutine $\mathcal{S}$: convex optimization in ℓ_1-ball

Parameters: $B > 0$, $t_{\mathrm{start}} > 0$, $\theta_{\mathrm{center}} \in \mathbb{R}^d$ and $\varepsilon > 0$.

For each $t = t_{\mathrm{start}}, t_{\mathrm{start}} + 1, \ldots,$

- predict $\widehat{\theta}_{t-1} \in \mathcal{B}_1(\theta_{\mathrm{center}}, \varepsilon)$ (thanks to some online gradient procedure)
- suffer loss $\ell_t\big(\widehat{\theta}_{t-1}\big) \in \mathbb{R}$ and observe the gradient $\nabla \ell_t\big(\widehat{\theta}_{t-1}\big) \in \mathbb{R}^d$

satisfies the following upper-bound on its cumulative regret: for all $t_{\mathrm{end}} \geqslant t_{\mathrm{start}}$ and for all $\theta \in \mathcal{B}_1\big(\theta_{\mathrm{center}}, \varepsilon\big)$

$$\sum_{t=t_{\mathrm{start}}}^{t_{\mathrm{end}}} \ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta) \leqslant a\varepsilon\sqrt{\sum_{t_{\mathrm{start}}}^{t_{\mathrm{end}}} \big\|\nabla \ell_t(\widehat{\theta}_{t-1})\big\|_\infty^2} + b\varepsilon B\,, \tag{5}$$

for some non-negative constants a, b that may depend on the dimension d.

Several online optimization algorithms do satisfy the regret bound (5) while being totally tuned, see for instance Gerchinovitz [2011, Corollary 2.1] or Cesa-Bianchi et al. [2007], Gaillard et al. [2014], Wintenberger [2014]. The regret bound is satisfied for instance with [‡] $a \lesssim \sqrt{\log d}$ and $b \lesssim \log d$ by a well online-calibrated Exponentiated Gradient (EG) forecaster combining the corners of $\mathcal{B}_1\big(\theta_{\mathrm{center}}, \varepsilon\big)$. This logarithmic dependence on the dimension is crucial here and possible because the optimization is performed in the ℓ_1-ball. SGD optimizing in the ℓ_2-ball, such as RDA of Xiao [2010], suffer a linear dependence on d. Therefore, they cannot be used as subroutines.

The regret bound yields the slow rate of convergence $\mathcal{O}\big(\sqrt{(\log d)(t_{\mathrm{end}} - t_{\mathrm{start}})}\big)$ (with respect to the length of the session) on the cumulative risk. Our acceleration procedure provides a generic method to also achieve a fast rate under sparsity.

3.2 The acceleration procedure

Our acceleration procedure (SAEW, described in Algorithm 2) performs the subroutine $\mathcal{S}$ on sessions of adaptive length optimizing in exponentially decreasing ℓ_1-balls. The sessions are indexed by $i \geqslant 0$ and denoted $\mathcal{S}_i$. The algorithm defines in an online fashion

[‡] As in the rest of the paper, the sign $\lesssim$ denotes an inequality which is fulfilled up to multiplicative constants.

Algorithm 2: SAEW

Parameters: $d_0 \geqslant 1$, $\alpha > 0$, $U > 0$, $B > 0$, $\delta > 0$ and a subroutine $\mathcal{S}$ that satisfies (5)

Initialization: $t_0 = t = 1, \varepsilon_0 = U$ and $\bar{\theta}_0 = 0$

For each $i = 0, 1, \ldots$

- define $[\bar{\theta}_{t_i-1}]_{d_0}$ by rounding to zero the $d - d_0$ smallest coefficients of $\bar{\theta}_{t_i-1}$
- start a new instance $\mathcal{S}_i$ of the subroutine $\mathcal{S}$ with parameters $t_{\text{start}} = t_i$, $\theta_{\text{center}} = [\bar{\theta}_{t_i-1}]_{d_0}$, $\varepsilon = U2^{-i/2}$ and B,
- for $t = t_i, t_i + 1, \ldots$ and while $\varepsilon_{t-1} > U2^{-(i+1)/2}$
  - forecast $\widehat{\theta}_{t-1}$ by using the subroutine $\mathcal{S}_i$
  - observe $\nabla \ell_t(\widehat{\theta}_{t-1})$
  - update the bound
$$\text{Err}_t := a_i' \sqrt{\sum_{s=t_i}^{t} \left\|\nabla \ell_s(\widehat{\theta}_{s-1})\right\|_\infty^2} + b_i' B$$
with a_i' and b_i' resp. defined in (10) and (11).
  - update the confidence radius
$$\varepsilon_t := 2\sqrt{\frac{2d_0 U 2^{-i/2}}{\alpha(t - t_i + 1)}\,\text{Err}_t}$$
  - update the averaged estimator
$$\bar{\theta}_t := (t - t_i + 1)^{-1} \textstyle\sum_{s=t_i}^{t} \widehat{\theta}_{s-1}$$
  - update the estimator
$$\tilde{\theta}_t := \bar{\theta}_{\arg\min_{0 \leqslant s \leqslant t} \varepsilon_s}$$
- stop the instance $\mathcal{S}_i$ and define $t_{i+1} := t + 1$

a sequence of starting times $1 = t_0 < t_1 < \ldots$ such that the instance $\mathcal{S}_i$ is used to perform predictions between times $t_{\text{start}} = t_i$ and $t_{\text{end}} = t_{i+1} - 1$. The idea is that our accuracy in the estimation of θ^* increases over time so that $\mathcal{S}_i$ can be a localized optimization subroutine in a small ball $\mathcal{B}_1\big([\bar{\theta}_{t_i-1}]_{d_0}, U2^{-i/2}\big)$ around the current sparse estimator $[\bar{\theta}_{t_i-1}]_{d_0}$ of θ^* at time t_i, see Algorithm 2 for the definition of $[\bar{\theta}_{t_i-1}]_{d_0}$.

The cumulative risk suffered during each session will remain constant: the increasing rate $\big(\sum_{t_i}^{t_{i+1}-1} \left\|\nabla \ell_t(\widehat{\theta}_{t-1})\right\|_\infty^2\big)^{1/2} \leqslant B\sqrt{t_{i+1} - t_i}$ due to the length of the session (see Equation (5)) will be shown to be of order $2^{i/2}$. But it will be offset by the decreasing radius $\varepsilon = U2^{-i/2}$.

By using a linear-time subroutine $\mathcal{S}$, the global time and storage complexities of SAEW are also $\mathcal{O}(dT)$.

Our main theorem is stated below. It controls the excess risk of the instantaneous estimators of SAEW. The proof is deferred to Appendix A.2.

Theorem 1. *Under Assumption* (SC)*, SAEW satisfies with probability at least* $1 - \delta$*,* $0 < \delta < 1$*, for all* $T \geqslant 1$

$$\text{Risk}\left(\tilde{\theta}_T\right) \leqslant \min\Bigg\{ UB\left(a'\sqrt{\frac{2}{T}} + \frac{4b'}{T}\right) + \frac{\alpha U^2}{8d_0 T}, \\ \frac{d_0 B^2}{\alpha}\left(\frac{2^7 a'^2}{T} + \frac{2^{11} b'^2}{T^2}\right) + \frac{2\alpha U^2}{d_0 T^2} \Bigg\},$$

where $a' = a + \sqrt{6\log(1 + 3\log T) - 2\log\delta}$ *and* $b' = b + 1/2 + 3\log(1 + 3\log t) - \log\delta$*.*

Remark 3.1. Using EG as the subroutines, the main term of the excess risk becomes of order

$$\text{Risk}\left(\tilde{\theta}_T\right) = \mathcal{O}_T\left(\frac{d_0 B^2}{\alpha T}\log\left(\frac{d\log T}{\delta}\right)\right). \qquad (6)$$

Remark 3.2. From the strong convexity assumption, Theorem 1 also ensures that, with probability $1 - \delta$, the estimator $\tilde{\theta}_T$ is close enough to θ^*:

$$\left\|\tilde{\theta}_T - \theta^*\right\|_2 \lesssim \frac{\sqrt{d_0}B}{\alpha\sqrt{T}}\sqrt{a'^2\log_2 T + \frac{b'^2}{T} + \frac{\alpha U^2}{d_0 T}}\,.$$

Theorem 2. *Under the assumptions and the notation of Theorem 1, the cumulative risk of SAEW is upper-bounded with probability at least* $1 - \delta$ *as*

$$\text{Risk}_{1:T}(\widehat{\theta}_{0:(T-1)}) \leqslant \min\Bigg\{ 4UB(a'\sqrt{T} + b' + 1), \\ \frac{2^5 d_0 B^2}{\alpha} a'^2 \log_2 T + 4UB(1 + b') + \frac{\alpha U^2}{8d_0} \Bigg\}.$$

Remark 3.3. Using EG as the subroutines, we get a cumulative risk of order

$$\text{Risk}_{1:T}(\widehat{\theta}_{0:(T-1)}) = \mathcal{O}_T\left(\frac{d_0 B^2}{\alpha T}\log\left(\frac{d\log T}{\delta}\right)\log T\right).$$

The averaged cumulative risk bound has an additional factor $\log T$ in comparison to the excess risk of $\tilde{\theta}_T$. This

logarithmic factor is unavoidable. Indeed, at time t, the rate stated in Equation (6) is optimal for any estimator. An optimal rate for the cumulative risk can thus be obtained by summing this rate of order $\mathcal{O}(1/t)$ over t introducing the log factor.

Remark 3.4. Adapting Corollary 13 of Gerchinovitz [2013], the boundedness of $\nabla\ell_t$ can be weakened to unknown B under the subgaussian condition. The price of this adaptation is a multiplicative factor of order $\log(dT)$ in the final bounds.

Remark 3.5. Using the strong convexity property, the averaging of SAEW has much faster rate ($\log T/T$ on the excess risk) than the averaging of the EG procedure itself (only slow rate $1/\sqrt{T}$ with high probability, see Audibert [2008]). But the last averaging $\tilde{\theta}_T$ achieves the best rate overall. Also note the difference of the impact of the ℓ_1-ball radius U on the rates: for the overall average $\bar{\theta}_T$ it is U^2/T whereas it is U^2/T^2 for the last averaging $\tilde{\theta}_T$. On the contrary to the overall averaging, the last averaging forgets the cost of the exploration of the initial ℓ_1-ball.

4 Square linear regression

Consider the common least square linear regression setting. Let (X_t, Y_t), $t \geqslant 1$ be i.i.d. random pairs taking values in $\mathbb{R}^d \times \mathbb{R}$. For simplicity, we assume that $\|X_t\|_\infty \leqslant X$ and $|Y_t| \leqslant Y$ almost surely for some constants $X, Y > 0$. We aim at estimating linearly the conditional mean of Y_t given X_t, by approaching $\theta^* = \arg\min_{\theta\in\mathbb{R}^d} \mathbb{E}\big[(Y_t - X_t^\top\theta)^2\big]$. Notice that the strong convexity of the risk is equivalent to the positivity of the covariance matrix of X_t as $\alpha \leqslant \lambda_{\min}\big(\mathbb{E}\big[X_t X_t^\top\big]\big)$, where $\lambda_{\min}$ is the smallest eigenvalue.

Applying the previous general setting to the square loss function $\ell_t : \theta \mapsto (Y_t - X_t^\top\theta)^2$, we get the following Theorem 3. It improves upon Theorem 1 the factor B^2 in the main term into a factor $X^2\sigma^2$, where $\sigma^2 := \mathbb{E}\big[(Y_t - X_t^\top\theta^*)^2\big]$ is the expected loss of the best linear predictor. This is achieved without the additional knowledge of σ^2. The proof of the theorem is highly inspired from the one of Theorem 1 and is deferred to Appendix A.6.

Theorem 3. *SAEW tuned with $B = 2X\big(Y + 2XU\big)$ satisfies with probability at least $1-\delta$ the bound*

$$\mathrm{Risk}(\tilde{\theta}_T) \lesssim \min\left\{ UX\left(\frac{\sigma a'}{\sqrt{T}} + \frac{(Y+XU)c'}{T}\right) + \frac{\alpha U^2}{d_0 T}, \right.$$
$$\left. \frac{X^2 d_0}{\alpha}\left(\frac{\sigma^2 a'^2}{T} + \frac{(Y+XU)^2 c'^2}{T^2}\right) + \frac{\alpha U^2}{d_0 T^2}\right\},$$

for all $T \geqslant 1$, where $a' \lesssim a + \sqrt{\log(1/\delta) + \log\log T}$ and $b' \lesssim b + \log(1/\delta) + \log\log T$.

Remark 4.1. Using a well-calibrated EG for the subroutines, the main term of the excess risk is of order

$$\mathrm{Risk}\left(\tilde{\theta}_T\right) = \mathcal{O}_T\left(\frac{d_0 X^2\sigma^2}{\alpha T}\log\left(\frac{d\log T}{\delta}\right)\right).$$

Remark 4.2. Similarly to Remark 3.4, if (X_t, Y_t) are subgaussian only (and not necessary bounded), classical arguments show that Theorem 3 still holds with X of order $\mathcal{O}(\log(dT))$ and $Y = \mathcal{O}(\log T)$.

Remark 4.3. The improvement from Theorem 1 to Theorem 3 (i.e., replacing B with $X^2\sigma^2$ in the main term) is less significant if we apply it to the cumulative risk (Theorem 2). This would improve $B^2\log T$ to $B^2 + X^2\sigma^2\log T$ and thus lead to a bound on the cumulative risk of order $\mathcal{O}(d_0\sigma^2\log(T)/\alpha)$.

Calibration of the parameters

To achieve the bound of Theorem 3, SAEW is given the parameters d_0, α, U, and B beforehand. We provide here how to tune these parameters in order to sequentially get an estimator achieving high rate on its excess risk. To do so, we use a combination of well-known calibration techniques: doubling trick, meta-algorithm, and clipping.

We only prove the calibration in the setting of linear regression with square loss (i.e., for Theorem 3 only and not for the general Theorem 1). It remains an open question whether the calibration of the parameters can be performed in the general setting of Section 3. We leave this question for future research. Furthermore, for the sake of clarity the adaption to Y (which is only necessary for clipping) is not considered here. However, it can be achieved simultaneously by updating the clipping range based on the past observations Y_s, $s \leqslant t-1$ (see [Gerchinovitz, 2013, Section 4.5]).

The calibration algorithm (Algorithm 3) works as follows. We define large enough grids of parameters for each doubling session $j \geqslant 0$

$$\mathcal{G}_j = \Big\{(d_0,\alpha,U,B) \in [1,\dots,d]\times\mathbb{R}_+^3 \quad \text{such that}$$
$$d_0 \in \{0\}\cup\big\{2^k, k=0,\dots,\lceil\log_2 d\rceil\big\}$$
$$\alpha \in \big\{2^k, k=-2j+\lceil\log_2(Bd_0/Y^2)\rceil,\dots,$$
$$j+\lceil\log_2 d_0\rceil\big\}$$
$$U \in \big\{2^k, k=-2j,\dots,2j+\lceil 2\log_2 Y\rceil\big\}$$

Algorithm 3: Calibration algorithm

Parameters: $Y > 0$, $\delta > 0$

Initialization: $t_0 = t = 1$ and $\bar{\theta}^{(0)} = 0$

For each $j = 0, 1, \dots$

- Define the grid $\mathcal{G}_j$ as in (7)
- For parameters $p = (d_0, \alpha, U, B) \in \mathcal{G}_j$:
  - Define $\delta_j = \delta/(2(j+1)^2)$
  - Run SAEW with parameter $(d_0, \alpha, U, B, \delta_j)$ for $t = 0, \dots, 2^j - 1$ and get the estimator $\tilde{\theta}_{2^j-1}$, denoted by $\tilde{\theta}_{p,j}$.
  - Define the clipped predictor

$$f_{p,j} : x \mapsto [x^\top \tilde{\theta}_{p,j}]_Y$$

    where $[\,\cdot\,]_Y := \max\big\{-Y, \min\{\,\cdot\,, Y\}\big\}$.
- For $t = 2^j, \dots, 2^{j+1} - 1$,
  - predict $\widehat{f}_{t-1}(X_t)$ by performing BOA with experts $(f_{p,j})_{p \in \mathcal{G}_j}$
  - output the estimator $\tilde{f}_{t-1} = \bar{f}_j$
- Define the average estimator

$$\bar{f}_{j+1} = 2^{-j} \sum_{t=2^j}^{2^{j+1}-1} \widehat{f}_{t-1}\,.$$

$$B \in \big\{2^k, k = -2j, \dots, 2j + \lceil 2\log_2 Y \rceil\big\}\Big\}. \tag{7}$$

For each set of parameters $p = (d_0, \alpha, U, B) \in \mathcal{G}_j$, we perform a local version of SAEW to obtain an estimator $\tilde{\theta}_{p,j}$ at time $t = 2^j - 1$. Then, the calibration algorithm uses the online aggregation procedure BOA of Wintenberger [2014] to make predictions from $t = 2^j$ to $2^{j+1} - 1$. Its predictions are based on online combinations of the (clipped) forecasts made by the $\tilde{\theta}_{p,j}$.

Theorem 4. *Let $Y > \max_{t=1,\dots,T} |Y_t|$ almost surely. With probability $1 - \delta$, the excess risk of the estimator $\tilde{f}_T$ produced by Algorithm 3 is of order*

$$\mathcal{O}_T\bigg(\frac{Y^2}{T}\log\Big(\frac{(\log d)(\log T + \log Y)}{\delta}\Big) + \frac{d_0 X^2 \sigma^2}{\alpha^* T}\log\Big(\frac{d\log T}{\delta}\Big)\bigg),$$

where $d_0 = \|\theta^*\|_0$ and $\alpha^* > 0$ is the largest value of α satisfying Inequality (SC).

The proof is postponed to Appendix A.7.

Remark 4.4. Similarly to the restricting eigenvalue condition of the Lasso, we believe that the strong convexity condition for α^* might be necessary on subspaces of dimension lower than d_0 only. However, to do so, SAEW should be used with a subroutine that produces sparse $\widehat{\theta}_{t-1}$. Up to our knowledge, such procedures do not exist for convex optimization in the ℓ_1-ball. As stated previously, sparse procedures such as RDA of Xiao [2010] cannot be used as subroutines since they perform optimization in the ℓ_2-ball and suffer a linear dependence on d. We leave this question for future work.

Remark 4.5. For the sake of clarity, the above result is only stated asymptotically. However the bound also holds in finite time up to universal multiplicative constant (as done in the proof). Additional negligible terms of order $\mathcal{O}(1/T^2)$ then appear in the bound. Furthermore, the finite time bound also achieves the best of the two regimes (slow rate vs fast rate) as in Theorem 3.

Remark 4.6. Theorem 4 has been proven only for square linear regression. However, it also holds for any strongly-convex loss function, with locally bounded gradients (i.e., with LIST condition, see Wintenberger [2014]).

Remark 4.7. To perform the calibration, we left the original framework of Section 2. First, because of the clipping, the estimators $\tilde{f}_{t-1}$ produced by Algorithm 3 are not linear any-more. Second, the meta-algorithm implies that we can observe the gradients of all subroutines SAEW simultaneously. Tuning the parameters in the original setting is left for future work.

5 Simulations

In this section, we provide computational experiments on simulated data. We compare three online aggregation procedures:

- RDA: a ℓ_1-regularized dual averaging method as proposed by Algorithm 2 of Xiao [2010]. The method was shown to produce sparse estimators. It obtained good performance on the MNIST data set of handwritten digits [LeCun et al., 1998]. We optimize the parameters γ, ρ, and λ in hindsight on the grid $\mathcal{E} := \{10^{-5}, \dots, 10^3\}$.
- BOA: the Bernstein Online Aggregation of Wintenberger [2014]. It proposes an adaptive calibration

of its learning parameters and achieves the fast rate for the model selection problem (see Nemirovski [2000]). BOA is initially designed to perform aggregation in the simplex, for the setting of prediction with expert advice (see Cesa-Bianchi and Lugosi [2006]). We use it together with the trick of Kivinen and Warmuth [1997] to extend it to optimization in the ℓ_1-ball $\mathcal{B}_1(0, \|\theta^*\|_1)$.

- SAEW: the acceleration procedure as detailed in Algorithm 2. We use BOA for the subroutines since it satisfies a regret bound of the form (5). For the parameters, we use $\delta = 0.95$, $U = \|\theta^*\|_1$ and $d_0 = \|\theta^*\|_0$. We calibrate α and B on the grid $\mathcal{E}$ in hindsight.

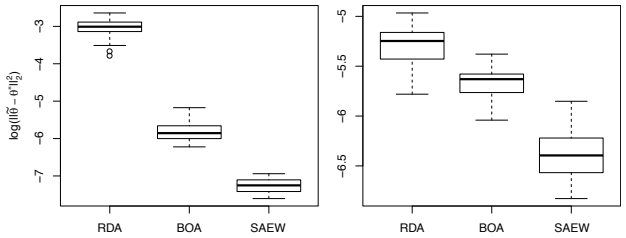
Our objective here is only to show the potential of the acceleration of BOA for a well-chosen set of parameters in the general setting of Section 3.

5.1 Application to square linear regression

We consider the square linear regression setting of Section 4. We simulate $X_t \sim \mathcal{N}(0, 1)$ for $d = 500$ and

$$Y_t = X_t^\top \theta^* + \varepsilon_t \qquad \text{with} \quad \varepsilon_t \sim \mathcal{N}(0, 0.01) \quad \text{i.i.d.},$$

where $d_0 = \|\theta^*\|_0 = 5$, $\|\theta^*\|_1 = 1$ with non-zero coordinates independently sampled proportional to $\mathcal{N}(0, 1)$.



(a) Square linear regression. (b) Quantile regression.

Figure 1: Boxplot of the logarithm of the ℓ_2 errors of the estimators $\tilde{\theta}_T$ at time $T = 2\,000$ with $d = 500$.

Figure 1a illustrates the results obtained by the different procedures after the observation of $T = 2\,000$ data points. It plots the box-plot of the ℓ_2 estimation errors of θ^*, which is also approximatively the instantaneous risk, over 30 experiments. In contrast to BOA and SAEW, RDA does not have the knowledge of $\|\theta^*\|_1$ in advance. This might explain the better performance obtained by BOA and SAEW. Another likely explanation comes from the theoretical guarantees of RDA, which is only linear in d (due to the sum of the squared gradients) though the ℓ_1-penalization.

In a batch context, the Lasso (together with cross-validation) may provide a better estimator for high dimensions d (its averaged error would be $\log \tilde{\theta}_T \approx -8.8$ in Figure 1a). This is mostly due to two facts. First, because of the online setting, our online procedures are here allowed to pass only once through the data. If we allowed multiple passes, their performance would be much improved. Second, although BOA satisfies theoretical guarantees in $\sqrt{\log d}$, its performance is deeply deteriorated when d becomes too large and does not converge before T being very large. We believe our acceleration procedure should thus be used with sparse online sub-procedures instead of BOA, but we leave this for future research.

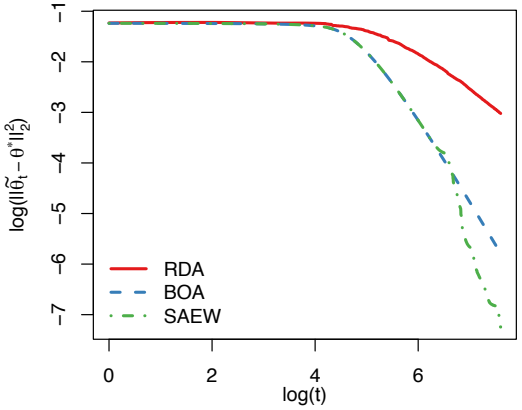


Figure 2: Averaged (over 30 experiments) evolution of the logarithm of the ℓ_2 error.

Figure 2 shows the decrease of the ℓ_2-error over time in log/log scale. The performance is averaged over the 30 experiments. We see that SAEW starts by following BOA, until it considers to be accurate enough to accelerate the process (around $\log t \approx 6.2$). Note that shortly after the acceleration start, the performance is shortly worse than the one of BOA. This can be explained by the doubling trick: the algorithm start learning again almost from scratch. The cumulative risks are displayed in Figure 3. SAEW and BOA seem to achieve logarithmic cumulative risk, in contrast to RDA which seems to be of order $\mathcal{O}(\sqrt{T})$.

In reality, the cumulative risk of BOA is of order $\mathcal{O}(\sigma^2 \sqrt{T \log d} + \log d)$. In the previous experiment, because of the small value of the noise $\sigma^2 = 0.01$, the first term is negligible in comparison to the second one unless T is very large. The behavior in $\sqrt{T}$ of BOA is thus better observed with higher noise and smaller di-

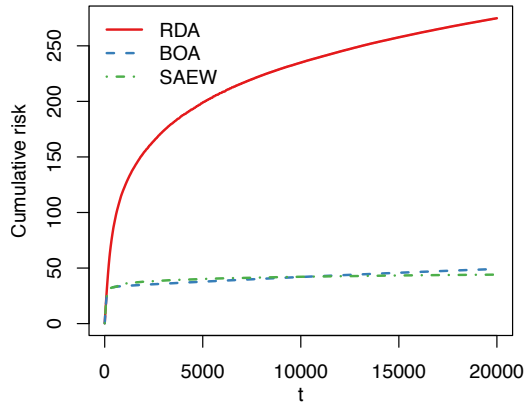


Figure 3: Averaged (over 30 runs) cumulative risk suffered by $\widehat{\theta}_t$ for square linear regression.

mension d, so that the first term becomes predominant. To illustrate this fact, we end the application on square linear regression with a simulation in small dimension $d_0 = d = 2$ with higher noise $\sigma = 0.3$. Our acceleration procedure can still be useful to obtain fast rates. Figure 4 shows that despite what seems on Figure 3, BOA does not achieve fast rate on its cumulative risk.

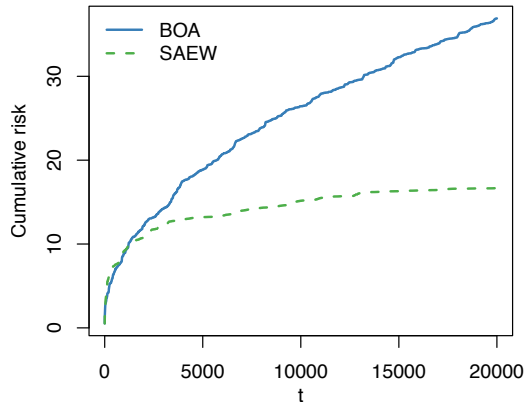


Figure 4: Cumulative risk suffered by $\widehat{\theta}_t$ for square linear regression with $d = d_0 = 2$.

5.2 Application to linear quantile regression

Let $\alpha \in (0,1)$. Here, we aim at estimating the conditional α-quantile of Y_t given X_t. A popular approach introduced by Koenker and Bassett [1978] consists in estimating the quantiles via the pinball loss defined for all $u \in \mathbb{R}$ by $\rho_\alpha(u) = u(\alpha - \mathbb{1}_{u<0})$. It can be shown that the conditional quantile $q_\alpha(Y_t|X_t)$ is the solution of the minimization problem

$$q_\alpha(Y_t|X_t) \in \arg\min_g \mathbb{E}\big[\rho_\alpha\big(Y_t - g(X_t)\big)\big|X_t\big] .$$

In linear quantile regression, we assume the conditional quantiles to be well-explained by linear functions of the covariates. Steinwart and Christmann [2011] proved that under some assumption the risk is strongly convex. We can thus apply our setting by using the loss functions $\ell_t : \theta \mapsto \rho_\alpha\big(Y_t - X_t^\top \theta\big)$.

We perform the same experiment as for linear regression (Y_t, X_t), but we aim at predicting the α-quantiles for $\alpha = 0.8$. To simulate an intercept necessary to predict the quantiles, we add a covariate 1 to the vector X_t. Figure 1b shows the improvements obtained by our accelerating procedure over the basic optimization algorithms.

In the next figures, to better display the dependence on T of the procedures, we run them during a longer time $T = 10^5$ with $d = 100$ only.

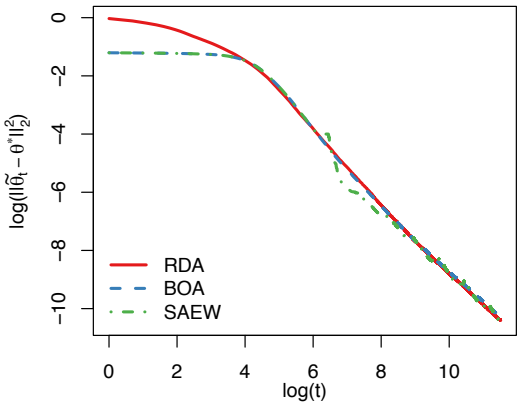


Figure 5: Averaged (over 30 runs) evolution of the logarithm of the ℓ_2-error for quantile regression ($d = 100$).

Figure 5 depicts the decreasing of the ℓ_2-errors of the different optimization methods (averaged over 30 runs). We see that unexpectedly most methods, although no theoretical properties, do achieve the fast rate $\mathcal{O}(1/T)$ (which corresponds to a slope -1 on the log/log scale). This explains why we do not really observe the acceleration on Figure 5. However, we only show here the dependence on t and not in d.

In Figure 6, we show how the slow rate high-probability bound on BOA (slope $-1/2$ in log/log scale) is transformed by SAEW into a fast rate bound (slope -1). To do so, it regularly restarts the algorithm to get smaller and smaller slow-rate bounds. Both BOA

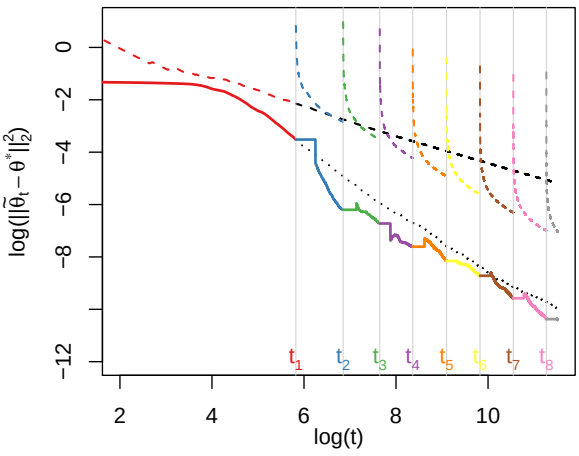


Figure 6: Logarithm of the ℓ_2-norm of the averaged estimator $\tilde{\theta}_t$ during one run. The dashed lines represent the high probability ℓ_2-bound estimated by SAEW on $\bar{\theta}_t$. The gray vertical lines are the stopping times t_i, $i \geqslant 1$. The first session is plotted in red, the second in blue,... The dotted and dashed black lines represent the performance (and the theoretical bound) that BOA would have obtained without acceleration.

(dotted black line) and SAEW do achieve fast rate here though only SAEW guarantees it. It would be interesting in the future to prove the fast rate convergence for the averaged estimator produced by BOA in this context. The classical proof technique that uses a cumulative risk to risk conversion (with Jensen's inequality) will have however to be changed since the fast rate is not achieved for the cumulative risk (see Figure 7).

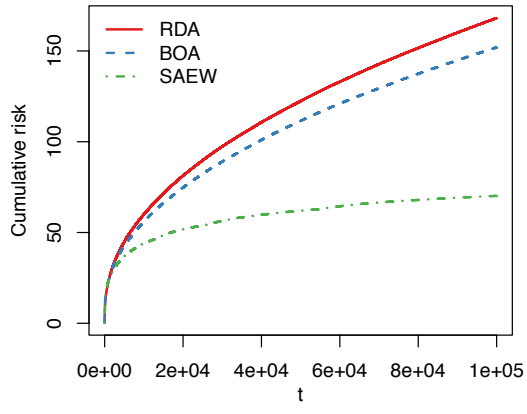


Figure 7: Averaged (over 30 runs) cumulative risk suffered by $\widehat{\theta}_t$ for quantile regression ($d = 100$).

References

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE TRANSACTIONS ON INFORMATION THEORY*, 58(5):3235, 2012.

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

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for gaussian regression. *The Annals of Statistics*, 35 (4):1674–1697, 2007.

O. Catoni. Statistical learning theory and stochastic optimization. Ecole d'Eté de probabilités de Saint-Flour 2001, Lectures Notes in Mathematics 1851, 2004.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

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

P. Gaillard, G. Stoltz, and T. van Erven. A second-order bound with excess losses. In *Proceedings of COLT'14*, volume 35, pages 176–196. JMLR: Workshop and Conference Proceedings, 2014.

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.

S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *The Journal of Machine Learning Research*, 14(1):729–769, 2013.

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

R. W. Koenker and G. W. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Y. LeCun and L. Bottou. Large scale online learning. *Advances in Neural Information Processing Systems*, 16:217, 2004.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

A. Nemirovski. Topics in non-parametric. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.

B. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

I. Steinwart and A. Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

O. Wintenberger. Optimal learning with bernstein online aggregation. Extended version available at arXiv:1404.1356 [stat. ML], 2014.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *COLT*, pages 921–948, 2014.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning, ICML 2003*, 2003.

SUPPLEMENTARY MATERIAL

A Proofs

A.1 Lemma 5

We first state Lemma 5, a classical result in strong convexity, as it will be useful in the proofs. It relates the ℓ_2-error of an estimator with its excess risk when the risk is strongly convex.

Lemma 5. *If the risk is 2α-strongly convex, then*

$$\|\theta - \theta^*\|_2^2 \leqslant \alpha^{-1} \operatorname{Risk}(\theta)$$

for all $\theta \in \mathbb{R}^d$.

Proof. Let $\theta \in \mathbb{R}^d$, by (SC) applied with $\theta_1 = \theta^*$ and $\theta_2 = \theta$, we get

$$\begin{aligned}\big\|\theta - \theta^*\big\|_2^2 \leqslant \alpha^{-1}\mathbb{E}\big[\ell_t(\theta) - \ell_t(\theta^*)\big] \\ + \alpha^{-1}\mathbb{E}\big[\nabla\ell_t(\theta^*)\big]^\top(\theta^* - \theta)\,.\end{aligned}$$

But, $\mathbb{E}\big[\nabla\ell_t(\theta^*)\big]^\top(\theta^* - \theta) \leqslant 0$. Otherwise, taking into account the convexity of the domain, the direction $d = \theta - \theta^*$ is a decreasing feasible direction, which contradicts the optimality of θ^*. □

A.2 Proof of Theorem 1

Let (δ_i) be a non-increasing sequence in $(0,1)$ such that $\sum_{i=1}^{\infty}\delta_i \leqslant \delta$.

Step 1. Proof by induction that the subroutines always perform the optimization in the correct ℓ_1-ball. We prove by induction on $i \geqslant 0$ that with probability at least $1 - \sum_{j=1}^{i}\delta_j$

$$\big\|\theta^* - [\bar\theta_{t_i-1}]_{d_0}\big\|_1 \leqslant U2^{-i/2}\,. \tag{8}$$

$\mathcal{H}_0$ is satisfied by assumption since $\|\theta^*\|_1 \leqslant U$ and $[\bar\theta_{t_0-1}]_{d_0} = [\bar\theta_0]_{d_0} = 0$ (see SAEW for the definition of $[\bar\theta_0]$).

Let $i \geqslant 0$ and assume (8). The following Lemma (whose proof is postponed to Appendix A.3) states that the gradients are indeed upper-bounded by B in sup-norm.

Lemma 6. *Let $i \geqslant 0$. Under (8), for all $t \in [t_i, t_{i+1} - 1]$, $\big\|\nabla\ell_t(\widehat\theta_{t-1})\|_\infty \leqslant B$ almost surely.*

Therefore, from the regret bound (5), the subroutine $\mathcal{S}_i$ satisfies for all $t \in [t_i, t_{i+1} - 1]$

$$\sum_{s=t_i}^{t} \ell_s(\widehat{\theta}_{s-1}) - \ell_s(\theta^*) \leqslant U 2^{-i/2} \left(a \sqrt{\sum_{s=t_i}^{t} \left\| \nabla \ell_s(\widehat{\theta}_{s-1}) \right\|_\infty^2} + bB \right).$$

Bounding the cumulative risk with the regret thanks to Theorem 10 in Appendix B.2, it yields with probability at least $1 - \sum_{j=1}^{i+1} \delta_j$,

$$\sum_{s=t_i}^{t} \mathbb{E}[\ell_s](\widehat{\theta}_{s-1}) - \mathbb{E}[\ell_s](\theta^*) \leqslant U 2^{-i/2} \operatorname{Err}_t \tag{9}$$

where $\operatorname{Err}_t := a_i' \sqrt{\sum_{t_i}^{t} \left\| \nabla \ell_s(\widehat{\theta}_{s-1}) \right\|_\infty^2} + b_i' B$ with

$$a_i' := a + \sqrt{2} \sqrt{\log\left(1 + \frac{1}{2} \log\left(\frac{t - t_i + 1}{2}\right)\right) - \log \delta_{i+1}}, \tag{10}$$

and

$$b_i' := b + \frac{1}{2} + \log\left(1 + \frac{1}{2} \log\left(\frac{t - t_i + 1}{2}\right)\right) - \log \delta_{i+1}. \tag{11}$$

Thus, recalling that by definition (see SAEW)

$$\bar{\theta}_t := (t - t_i + 1)^{-1} \sum_{s=t_i}^{t} \widehat{\theta}_{s-1},$$

and because the losses are i.i.d., Jensen's inequality yields

$$\begin{aligned} \operatorname{Risk}(\bar{\theta}_t) &= \mathbb{E}[\ell_{t+1}](\bar{\theta}_t) - \mathbb{E}[\ell_{t+1}](\theta^*) \\ &\overset{\text{Jensen}}{\leqslant} (t - t_i + 1)^{-1} \sum_{s=t_i}^{t} \mathbb{E}[\ell_s](\widehat{\theta}_{s-1}) - \mathbb{E}[\ell_s](\theta^*) \\ &\overset{(9)}{\leqslant} \frac{U \operatorname{Err}_t}{2^{i/2}(t - t_i + 1)}. \end{aligned} \tag{12}$$

Together with the strong convexity of the risk (Lemma 5), this entails

$$\left\| \bar{\theta}_t - \theta^* \right\|_2^2 \leqslant \frac{U \operatorname{Err}_t}{\alpha 2^{i/2}(t - t_i + 1)}. \tag{13}$$

We thus control the ℓ_2-error of $\bar{\theta}_t$. However, in order to control the ℓ_1-error without paying a factor d, we need to truncate coordinates of $\bar{\theta}_t$. By definition of $[\bar{\theta}_t]_{d_0}$ (see SAEW), we have

$$[\bar{\theta}_t]_{d_0} \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d : \|\theta\|_0 \leqslant d_0} \left\{ \left\| \bar{\theta}_t - \theta \right\|_2 \right\}. \tag{14}$$

Now, (14) together with $\|\theta^*\|_0 \leqslant d_0$ (by assumption) yields

$$\left\| \bar{\theta}_t - [\bar{\theta}_t]_{d_0} \right\|_2 \leqslant \left\| \bar{\theta}_t - \theta^* \right\|_2. \tag{15}$$

Furthermore, because both $\|\theta^*\|_0 \leqslant d_0$ and $\left\|[\bar{\theta}_t]_{d_0}\right\|_0 \leqslant d_0$, we have

$$\|[\bar{\theta}_t]_{d_0} - \theta^*\|_0 \leqslant 2d_0. \tag{16}$$

Therefore, with probability at least $1 - \sum_{j=0}^{i+1} \delta_j$

$$\begin{aligned} \left\| [\bar{\theta}_t]_{d_0} - \theta^* \right\|_1 &\overset{(16)}{\leqslant} \sqrt{2d_0} \left\| [\bar{\theta}_t]_{d_0} - \theta^* \right\|_2 \\ &\leqslant \sqrt{2d_0} \left(\left\| [\bar{\theta}_t]_{d_0} - \bar{\theta}_t \right\|_2 + \left\| \bar{\theta}_t - \theta^* \right\|_2 \right) \\ &\overset{(15)}{\leqslant} 2\sqrt{2d_0} \left\| \bar{\theta}_t - \theta^* \right\|_2 \\ &\overset{(13)}{\leqslant} 2 \sqrt{2d_0 \alpha^{-1} U \operatorname{Err}_t \, 2^{-i/2} (t - t_i + 1)^{-1}} \\ &=: \varepsilon_t, \end{aligned} \tag{17}$$

where the last equality holds by definition of ε_t (see SAEW). Finally, $(\mathcal{H}_{i+1})$ is fulfilled by definition of t_{i+1} (see SAEW), which satisfies $\varepsilon_{t_{i+1}-1} \leqslant U 2^{-(i+1)/2}$. The induction is thus completed.

In the rest of the proof, we consider that (8) are satisfied for all $i \geqslant 0$. This occurs with probability $1 - \sum_{j=1}^{\infty} \delta_j \geqslant 1 - \delta$ as stated by Step 1.

Step 2. Fast rate for the excess risk of $\tilde{\theta}_t$. First, we prove that the excess risk of $\tilde{\theta}_t$ is upper-bounded as

$$\operatorname{Risk}(\tilde{\theta}_t) \leqslant \frac{d_0 B^2}{\alpha} \left(\frac{2^7 a'^2}{t} + \frac{2^{11} b'^2}{t^2} \right) + \frac{2\alpha U^2}{d_0 t^2}, \tag{18}$$

for all $t \geqslant 1$, where $a' = a'_{\lfloor 2 \log_2 t \rfloor}$ and $b' = b'_{\lfloor 2 \log_2 t \rfloor}$.

To do so, we start from the risk inequality (12). From the definition of ε_t (see (17)), we get

$$\operatorname{Risk}(\bar{\theta}_t) \leqslant \frac{\alpha \varepsilon_t^2}{8 d_0}, \qquad t \geqslant 1. \tag{19}$$

Thus by definition of $\tilde{\theta}_t := \bar{\theta}_{\operatorname{arg\,min}_{s \leqslant t} \varepsilon_s}$, we have

$$\operatorname{Risk}(\tilde{\theta}_t) \leqslant \frac{\alpha \min_{s \leqslant t} \varepsilon_s^2}{8 d_0} \tag{20}$$

We conclude the proof with the following lemma proved in Appendix A.4

Lemma 7. *Let $i \geqslant 0$. Let $t_i - 1 \leqslant t \leqslant t_{i+1}$, then*

$$\min_{s \leqslant t} \varepsilon_s \leqslant U \left(\frac{\sqrt{2} \gamma a_i'}{\sqrt{t}} + \frac{2 + 4\gamma b_i'}{t} \right),$$

where $\gamma := 2^4 d_0 B / (\alpha U)$.

Let $i \geqslant 0$ such that $t_i - 1 \leqslant t \leqslant t_{i+1}$. Lemma 7 together with (20) and $(x+y)^2 \leqslant 2x^2+2y^2$ for $x, y \geqslant 0$, yields

$$\mathrm{Risk}(\tilde{\theta}_t) \leqslant \frac{\alpha U^2\gamma^2}{8d_0}\left(\frac{\sqrt{2}a'_i}{\sqrt{t}} + \frac{2\gamma^{-1}+4b'_i}{t}\right)^2 \qquad (21)$$

$$\leqslant \frac{\alpha U^2\gamma^2}{d_0}\left(\frac{a'^2_i}{2t} + \frac{2\gamma^{-2}+8b'^2_i}{t^2}\right). \qquad (22)$$

Now, remark that if $i \geqslant 2\log t$, then $\varepsilon_{t_i-1} \leqslant U2^{-i} \leqslant U/t$ and from (20), $\mathrm{Risk}(\tilde{\theta}_t) \leqslant \alpha U^2/(8d_0t^2)$. Together, with (22), we get

$$\mathrm{Risk}(\tilde{\theta}_t) \leqslant \frac{\alpha U^2\gamma^2}{d_0}\left(\frac{a'^2}{2t} + \frac{2\gamma^{-2}+8b'^2}{t^2}\right),$$

with $a' = a'_{\lfloor 2\log_2 t\rfloor}$ and $b' = b'_{\lfloor 2\log_2 t\rfloor}$. Substituting $\gamma = 2^4 d_0 B/(\alpha U)$ concludes the proof of Inequality (18).

Step 3. Slow rate for the excess risk of $\tilde{\theta}_t$. Now, we prove that

$$\mathrm{Risk}(\tilde{\theta}_t) \leqslant UB\left(\frac{a'}{\sqrt{t/2}} + \frac{4b'}{t}\right) + \frac{\alpha U^2}{8d_0 t},\ t \geqslant 1. \quad (23)$$

For small values of t, the slow rate will be satisfied from the initial bound of the subroutine during the first session. At some time $\tau > 0$, the fast rate becomes better than the slow rate. This splitting time is defined as the solution of the equality

$$\frac{\mathrm{Err}_{t_1-1}}{t_1-1} = B\Big(\frac{\sqrt{2}a'}{\sqrt{\tau}} + \frac{2\gamma^{-1}+4b'}{\tau}\Big). \qquad (24)$$

Let $t \geqslant 1$. To control $\mathrm{Risk}(\tilde{\theta}_t)$, we distinguish three cases:

- if $t \leqslant t_1 - 1$, then, since by definition of ε_s

  $$\arg\min_{s\leqslant t} \frac{\mathrm{Err}_s}{s} = \arg\min_{s\leqslant t} \varepsilon_s,$$

  we get from Inequality (12) that

  $$\mathrm{Risk}(\tilde{\theta}_t) = \mathrm{Risk}(\bar{\theta}_{\arg\min_{s\leqslant t}\varepsilon_s}) \leqslant U2^{-0/2}\min_{s\leqslant t}\frac{\mathrm{Err}_s}{s} \leqslant U\frac{\mathrm{Err}_t}{t}.$$

  By definition of Err_t (see (9)) and upper-bounding the gradients by B, we get

  $$\mathrm{Risk}(\tilde{\theta}_t) \leqslant UB\Big(\frac{a'_0}{\sqrt{t}} + \frac{b'_0}{t}\Big).$$

- if $t_1 \leqslant t \leqslant \tau$, then following the same reasoning as above, we have

  $$\mathrm{Risk}(\tilde{\theta}_t) \leqslant U\frac{\mathrm{Err}_{t_1-1}}{t_1-1},$$

  which yields by definition of τ (see Equality (24)) and by using $t \leqslant \tau$:

  $$\mathrm{Risk}(\tilde{\theta}_t) \leqslant UB\Big(\frac{\sqrt{2}a'}{\sqrt{\tau}} + \frac{2\gamma^{-1}+4b'}{\tau}\Big) \leqslant UB\Big(\frac{\sqrt{2}a'}{\sqrt{t}} + \frac{2\gamma^{-1}+4b'}{t}\Big).$$

- if $\tau \leqslant t$, since by definition of t_1 (see SAEW), $\varepsilon_{t_1-1} \leqslant U/2$, then by definition of ε_{t_1-1} (see (17)),

  $$2\sqrt{2d_0\alpha^{-1}U\frac{\mathrm{Err}_{t_1-1}}{t_1-1}} \leqslant \frac{U}{2},$$

  and thus taking the square and rearranging the terms

  $$\frac{d_0}{\alpha} \leqslant \frac{U}{2^5}\Big(\frac{t_1-1}{\mathrm{Err}_{t_1-1}}\Big).$$

  Using the definition of $\gamma = 2^4 d_0 B/(\alpha U)$ and substituting Err_{t_1-1} with Equality (24), this yields

  $$\frac{\alpha U^2\gamma^2}{8d_0} = \frac{2^5 d_0 B^2}{\alpha} \leqslant UB\Big(\frac{\sqrt{2}a'}{\sqrt{\tau}} + \frac{2\gamma^{-1}+4b'}{\tau}\Big)^{-1}.$$

  Finally from Inequality (21), and using $\tau \leqslant t$

  $$\mathrm{Risk}(\tilde{\theta}_t) \leqslant UB\left(\frac{\sqrt{2}a'}{\sqrt{t}} + \frac{2\gamma^{-1}+4b'}{t}\right).$$

Combining the three cases together and substituting $\gamma = 2^4 d_0 B/(\alpha U)$, concludes the proof of Inequality (23).

Step 4. Conclusion of the proof Combining Inequalities (18) and (23), we get the risk inequality stated in the theorem for $\tilde{\theta}_t$. It only remains to choose $\delta_j = \delta/(j+1)^2$ so that $\sum_{j=1}^{\infty}\delta_j \leqslant \delta$ and to control $a' = a'_{\lfloor 2\log_2 t\rfloor}$ and $b' = b'_{\lfloor 2\log_2 t\rfloor}$. From (10), we can use $\delta_{\lfloor 2\log_2 t\rfloor+1} \geqslant \delta/(1+2\log_2 t)^2$ and $T_i \leqslant t$. Straightforward calculation yields that $a' - a$ is lower than

$$\sqrt{2\big(\log(1+1/2\log(t/2)) - \log\delta + 2\log(1+2\log_2 t)\big)} \leqslant \sqrt{6\log(1+3\log t) - 2\log\delta}.$$

Similarly, for $b' - b$. It is upper-bounded by

$$\frac{1}{2} + \log\big(1+(1/2)\log(t/2)\big) - \log\delta + 2\log(1+2\log_2 t) \leqslant 1/2 + 3\log(1+3\log t) - \log\delta.$$

This concludes the proof.

A.3 Proof of Lemma 6

Since by assumption $B \geqslant \max_{\theta:\|\theta\|_1 \leqslant 2U} \|\nabla \ell_t(\theta)\|_\infty$ a.s. Therefore, it suffices to show that $\|\widehat{\theta}_{t-1}\|_1 \leqslant 2U$. By definition of the session $\mathcal{S}_i$,

$$\widehat{\theta}_{t-1} \in \mathcal{B}_1([\bar{\theta}_{t_i-1}]_{d_0}, U2^{-i/2})\,.$$

Thus:

- if $i = 0$, since $[\bar{\theta}_0]_{d_0} = 0$, $\|\widehat{\theta}_{t-1}\|_1 \leqslant U$.
- if $i = 1$, then since $\|[\bar{\theta}_{t_1-1}]\|_1 \leqslant U$ as a truncated average of vectors in $\mathcal{B}_1(0, U)$, we have

$$\begin{aligned}\|\widehat{\theta}_{t-1}\|_1 &\leqslant \|\widehat{\theta}_{t-1} - [\bar{\theta}_{t_1-1}]_{d_0}\|_1 + \|[\bar{\theta}_{t_1-1}]_{d_0}\|_1 \\ &\leqslant U/\sqrt{2} + U \leqslant 2U;\end{aligned}$$

- otherwise, $i \geqslant 2$ and $\|\widehat{\theta}_{t-1}\|_1$ is bounded by

$$\begin{aligned}&\|\widehat{\theta}_{t-1} - [\bar{\theta}_{t_i-1}]_{d_0}\|_1 + \|[\bar{\theta}_{t_i-1}]_{d_0} - \theta^*\|_1 + \|\theta^*\|_1 \\ &\qquad \overset{(8)}{\leqslant} U2^{-i/2} + U2^{-i/2} + U \leqslant 2U\,.\end{aligned}$$

Putting the tree cases together, $\|\widehat{\theta}_{t-1}\|_1 \leqslant 2U$, which concludes the proof.

A.4 Proof of Lemma 7

It is enough to control $\varepsilon_{t_i-1} \geqslant \min_{s \leqslant t} \varepsilon_s$. To do so, we prove that for every $j \geqslant 0$, $T_j := t_{j+1} - t_j$ cannot be too large, so that at time t, i will be at least of order $\log_2 t$.

Let $j \geqslant 0$. We can assume $t_{j+1} > t_j$, otherwise $T_j = 0$. Thus, from the bound on the gradients (Lemma 6) and from the definition of Err_t (see (9)) for all $t \in [t_j + 1, t_{j+1}]$,

$$\mathrm{Err}_{t-1} \leqslant B(a'_j\sqrt{t - t_j} + b'_j)\,, \tag{25}$$

and from the definition of ε_{t-1} (see (17))

$$\varepsilon_{t-1} \leqslant 2\sqrt{2d_0\alpha^{-1}UB\frac{a'_j\sqrt{t-t_j} + b'_j}{2^{j/2}(t-t_j)}}\,.$$

Since by definition, t_{j+1} is the smallest integer after t_j that satisfies $\varepsilon_{t_{j+1}-1} \leqslant U2^{-(j+1)/2}$, we have $\varepsilon_{t_{j+1}-2} \geqslant U2^{-(j+1)/2}$. This implies

$$\begin{aligned}&2\sqrt{2d_0\alpha^{-1}UB\frac{a'_j\sqrt{T_j - 1} + b'_j}{2^{j/2}(T_j - 1)}} \geqslant U2^{-(j+1)/2} \\ &\Leftrightarrow\ 2^{j/2}\underbrace{2^4 d_0\alpha^{-1}U^{-1}B}_{:=\gamma}\left(a'_j\sqrt{T_j - 1} + b'_j\right) \geqslant T_j - 1\end{aligned}$$

Then, by solving a second order equation in $\sqrt{T_j - 1}$ (see for instance [Gaillard et al., 2014, Lemma 10]), the above inequality entails

$$T_j \leqslant 1 + 2^j\,\gamma^2 a'^2_j + 2^{j/2}\gamma b'_j\,. \tag{26}$$

Therefore, summing over $j = 0, \ldots, i$

$$\begin{aligned}t_{i+1} &= \cancel{0} + \sum_{j=0}^{i} T_j \\ &\leqslant \sum_{j=0}^{i}\left(1 + 2^j\,\gamma^2 a'^2_j + 2^{j/2}\gamma b_j\right) \\ &\leqslant 2^{1+i}\,\gamma^2 a'^2_i + (1+\sqrt{2})2^{(i+1)/2}\gamma b'_i + i + 1 \\ &\leqslant 2^{1+i}\,\gamma^2 a'^2_i + 2^{(i+1)/2}\sqrt{2}\big(2\gamma b'_i + 1\big)\,,\end{aligned}$$

where the last inequality is because $2^{(i+1)/2} \geqslant \sqrt{2}(i+1)$ for $i \geqslant 0$. Solving the second-order inequality in $2^{(i+1)/2}$ we get

$$2^{-(i+1)/2} \leqslant \frac{\gamma a'_i}{\sqrt{t_{i+1}}} + \sqrt{2}\frac{1 + 2\gamma b'_i}{t_{i+1}}\,.$$

Thus, since $\varepsilon_{t_i-1} \leqslant U2^{-i/2}$, we have

$$\varepsilon_{t_i-1} \leqslant U\gamma\left(\frac{\sqrt{2}a'_i}{\sqrt{t_{i+1}}} + \frac{2\gamma^{-1} + 4b'_i}{t_{i+1}}\right)\,.$$

The proof of Lemma 7 finally follows using $t \leqslant t_{i+1}$.

A.5 Proof of Theorem 2

With probability $1 - \delta$, all inequalities provided in the proof of Theorem 1 are satisfied. We also consider the notation of the previous proof. Let $t \geqslant 1$.

Step 1. Slow rate We remark that for any $i \geqslant 0$,

$$\begin{aligned}\sum_{s=t_i}^{(t_{i+1}-1)\wedge t} \mathbb{E}[\ell_s](\widehat{\theta}_{s-1}) - \mathbb{E}[\ell_s](\theta^*) &\overset{(9)}{\leqslant} U2^{-i/2}\,\mathrm{Err}_{(t_{i+1}-1)\wedge t} \\ &\overset{(25)}{\leqslant} UB2^{-i/2}(a'_i\sqrt{t} + b'_i)\end{aligned} \tag{27}$$

where, in the last inequality, we use that $(t_{i+1}-1)\wedge t \leqslant t$ and $t_i \geqslant 1$. We will use this inequality for $i \leqslant \lfloor 2\log t \rfloor$. For $i > \lfloor 2\log t \rfloor$, we use the fact that the gradients are bounded by B, so that by convexity of the risk

$$\sum_{s=t_i}^{(t_{i+1}-1)\wedge t} \mathbb{E}[\ell_s](\widehat{\theta}_{s-1}) - \mathbb{E}[\ell_s](\theta^*)$$

$$\leqslant \sum_{s=t_i}^{(t_{i+1}-1)\wedge t} \|\mathbb{E}[\nabla \ell_s](\widehat{\theta}_{s-1})\|_\infty \|\widehat{\theta}_{s-1} - \theta^*\|_1$$
$$\leqslant UB 2^{-i/2} t\,. \tag{28}$$

Summing (27) over $i = 0, \dots, \lfloor 2\log_2 t \rfloor$ and (28) over $i = \lceil 2\log_2 t \rceil, \dots, \infty$, we get

$$\begin{aligned}\mathrm{Risk}_{1:t}(\widehat{\theta}_{0:(t-1)}) &:= \sum_{s=1}^{t} \mathbb{E}[\ell_s](\widehat{\theta}_{s-1}) - \mathbb{E}[\ell_s](\theta^*) \\ &\leqslant UB \sum_{i=0}^{\lfloor 2\log_2 t \rfloor} 2^{-i/2}(a_i'\sqrt{t} + b_i') \\ &\quad + UBt \sum_{i=\lceil 2\log_2 t \rceil}^{\infty} 2^{-i/2}\,. \end{aligned} \tag{29}$$

The second sum is controlled as

$$\sum_{i=\lceil 2\log_2 t \rceil}^{\infty} 2^{-i/2} \leqslant t^{-1} \sum_{i=0}^{\infty} 2^{-i/2}\,.$$

Thus, since $\sum_{i=0}^{\infty} 2^{-i/2} = 2 + \sqrt{2} \leqslant 4$, we have

$$\mathrm{Risk}_{1:t}(\widehat{\theta}_{0:(t-1)}) \leqslant 4UB(a'\sqrt{t} + b') + 4UB\,,$$

where we recall that $a' = a'_{\lfloor 2\log_2 t \rfloor}$ and $b' = b'_{\lfloor 2\log_2 t \rfloor}$. This concludes Step 1.

Step 2. Fast rate Let us now prove the fast rate

$$\mathrm{Risk}_{1:t}\left(\widehat{\theta}_{0:(t-1)}\right) \leqslant \frac{2^5 d_0 B^2}{\alpha} a'^2 \log_2 t + 4BU(1+b') + U^2 \frac{\alpha}{8d_0}\,,$$

for all $t \geqslant 1$.

First, we remark that similarly to (19), we get for all $i \geqslant 0$ that

$$\begin{aligned}\sum_{s=t_i}^{t_{i+1}-1} \mathbb{E}[\ell_s](\widehat{\theta}_{s-1}) - \mathbb{E}[\ell_s](\theta^*) &\overset{(9)}{\leqslant} U \frac{\mathrm{Err}_{t_{i+1}-1}}{2^{-i/2} T_i} T_i \\ &\overset{(17)}{\leqslant} \frac{\alpha \varepsilon_{t_{i+1}-1}^2}{8d_0} T_i \\ &\leqslant \frac{\alpha U^2 2^{-i}}{16 d_0} T_i \end{aligned} \tag{30}$$

where the last inequality is because $\varepsilon_{t_{i+1}-1} \leqslant U 2^{-(i+1)/2}$ by definition of t_{i+1} (see SAEW). We will use this inequality for $i \leqslant \lfloor 2\log t \rfloor$. Summing (30) over $i = 0, \dots, \lfloor 2\log_2 t \rfloor$ and (28) over $i = \lceil 2\log_2 t \rceil, \dots, \infty$, we get

$$\begin{aligned}\mathrm{Risk}_{1:t}(\widehat{\theta}_{0:(t-1)}) &:= \sum_{s=1}^{t} \mathbb{E}[\ell_s](\widehat{\theta}_{s-1}) - \mathbb{E}[\ell_s](\theta^*) \\ &\leqslant \frac{U^2 \alpha}{2^4 d_0} \sum_{i=0}^{\lfloor 2\log_2 t \rfloor} 2^{-i} T_i \\ &\quad + UBt \sum_{i=\lceil 2\log_2 t \rceil}^{\infty} 2^{-i/2}\,. \end{aligned} \tag{31}$$

We upper bound both sums. The second one is controlled as we did for (29). The first one is upper-bounded thanks to (26)

$$\begin{aligned}\sum_{i=0}^{\lfloor 2\log_2 t \rfloor} 2^{-i} T_i &\leqslant \sum_{i=0}^{\lfloor 2\log_2 t \rfloor} \left(\gamma^2 a_i'^2 + 2^{-i/2}\gamma b_i' + 2^{-i}\right) \\ &\leqslant 2\gamma^2 a'^2 \log_2 t + 4\gamma b' + 2\,. \end{aligned}$$

Therefore, substituting the two sums into (31), the cumulative risk $\mathrm{Risk}_{1:t}(\widehat{\theta}_{0:(t-1)})$ is upper-bounded by

$$\frac{U^2\alpha}{2^4 d_0}\left(2\gamma^2 a'^2 \log_2 t + 4\gamma b' + 2\right) + 4UB\,,$$

which, by substituting $\gamma = 2^4 d_0 B/(\alpha U)$, is equal to

$$\frac{2^5 d_0 B^2}{\alpha} a'^2 \log_2 t + 4BU(1+b') + \frac{\alpha U^2}{8d_0}\,.$$

This concludes the proof.

A.6 Proof of Theorem 3

Let first check that we are indeed in the setting of Theorem 1. The risk is strongly convex because for any $\theta_1, \theta_2 \in \mathbb{R}^d$

$$\begin{aligned}&\mathbb{E}[\ell_t(\theta_1) - \ell_t(\theta_2)] = \mathbb{E}\Big[(Y_t - X_t^\top \theta_1)^2 - (Y_t - X_t^\top \theta_2)^2\Big] \\ &= \mathbb{E}\Big[-2(Y_t - X_t^\top \theta_1) X_t^\top (\theta_1 - \theta_2) - \big(X_t^\top(\theta_1 - \theta_2)\big)^2\Big] \\ &= \nabla \mathbb{E}[\ell_t](\theta_1)^\top (\theta_1 - \theta_2) - (\theta_1 - \theta_2)^\top \mathbb{E}\big[X_t X_t^\top\big](\theta_1 - \theta_2)\,. \end{aligned}$$

Assumption (SC) is thus satisfied with $\alpha = \lambda_{\min}(\mathbb{E}\big[X_t X_t^\top\big])$. Besides, for all θ such that $\|\theta\|_1 \leqslant 2U$, we have

$$\|\nabla \ell_t(\theta)\|_\infty = \|2(Y_t - X_t^\top \theta) X_t\|_\infty \leqslant 2(Y + 2XU)X = B\,.$$

Now, we mimic the proof of Theorem 1. In the rest of the proof, we consider that (8) are satisfied for all $i \geqslant 0$.

This occurs with probability $1-\delta$ and all inequalities stated in the proof of Theorem 1 are satisfied.

The proof is based on the following Lemma that we substitute to Inequality (25) from the proof of Theorem 1.

Lemma 8. *For all $t \in [t_i, t_{i+1}-1]$, with probability $1-\delta_{i+1}$,*

$$\mathrm{Err}_{t-1} \leqslant 2\sqrt{2}X\sigma a'_i\sqrt{t-t_i} + Bc'_i\,,$$

where $c'_i := b'_i + a'_i\big(\sqrt{\log \delta_{i+1}^{-1}} + \sqrt{2b} + 2a\big)$. We recall that Err_{t-1} is defined in (9).

Proof of Lemma 8. In the particular case of the square loss, the gradients are given by $\nabla \ell_t(\theta) = 2X_t(X_t^\top \theta - Y_t)$, so that

$$\|\nabla \ell_t(\widehat{\theta}_{t-1})\|_\infty^2 \leqslant 4X^2 \ell_t(\widehat{\theta}_{t-1})\,. \tag{32}$$

Following [Gerchinovitz, 2013, Corollary 2.2], we get from Inequality (5) that

$$\sum_{t=t_{\mathrm{start}}}^{t_{\mathrm{end}}} \ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta^*) \leqslant 2aUX\sqrt{\sum_{t_{\mathrm{start}}}^{t_{\mathrm{end}}} \ell_t(\widehat{\theta}_{t-1})} + bUB\,.$$

Solving the second-order inequality (see [Gaillard et al., 2014, Lemma 10]), it yields the improvement for small losses

$$\sqrt{\sum_{t=t_{\mathrm{start}}}^{t_{\mathrm{end}}} \ell_t(\widehat{\theta}_{t-1})} \leqslant \sqrt{\sum_{t=t_{\mathrm{start}}}^{t_{\mathrm{end}}} \ell_t(\theta^*)} + \sqrt{bUB} + 2aUX\,.$$

Thus, from (32),

$$\sqrt{\sum_{s=t_i}^{t-1} \big\|\nabla \ell_s(\widehat{\theta}_{s-1})\big\|_\infty^2} \leqslant 2X\sqrt{\sum_{s=t_i}^{t-1} \ell_s(\theta^*)} + 2X\sqrt{bUB} + 4aUX^2\,.$$

But, with probability $1-\delta_{i+1}$, we have from Theorem 9

$$\begin{aligned}\sum_{s=t_i}^{t-1} \ell_s(\theta^*) &\leqslant (e-1)\sum_{s=t_i}^{t-1} \mathbb{E}[\ell_s(\theta^*)] + (Y+XU)^2 \log \delta_{i+1}^{-1} \\ &\leqslant 2\sigma^2(t-t_i) + (Y+XU)^2 \log \delta_{i+1}^{-1}\,,\end{aligned}$$

where $\sigma^2 = \mathbb{E}[\ell_t(\theta^*)]$. Plugging into the previous inequality and using $\sqrt{x+y} \leqslant \sqrt{x} + \sqrt{y}$ for $x, y > 0$, this yields

$$\begin{aligned}&2^{-1}X^{-1}\sqrt{\sum_{s=t_i}^{t-1} \big\|\nabla \ell_s(\widehat{\theta}_{s-1})\big\|_\infty^2} \\ &\leqslant \sqrt{2}\sigma\sqrt{t-t_i} + (Y+XU)\sqrt{\log \delta_{i+1}^{-1}} + \sqrt{bUB} + 2aUX \\ &\leqslant \sqrt{2}\sigma\sqrt{t-t_i} + 2^{-1}BX^{-1}\big(\sqrt{\log \delta_{i+1}^{-1}} + \sqrt{2b} + 2a\big)\,,\end{aligned} \tag{34}$$

where the second inequality is because $B/(2X) \geqslant (Y + XU) \geqslant XU$. The proof of Lemma 8 is concluded by using the definition of Err_{t-1} (see (9)). □

The proof of Theorem 3 is then completed following the one of Theorem 1 by using Lemma 8 instead of Inequality (25). Finally, it only suffices to substitute Ba'_i with $2\sqrt{2}X\sigma a'_i$ and b'_i with c'_i in the final results. At the end, b' of Theorem 1 must thus be substituted with

$$\begin{aligned}c' &:= b' + a'\big(\sqrt{2\log(1+2\log_2 T) - \log\delta} + \sqrt{2b} + 4a\big) \\ &\leqslant 1/2 + b + 3\log(1+3\log T) - \log\delta \\ &\qquad + \big(a + \sqrt{6\log(1+3\log t) - 2\log\delta}\big) \\ &\qquad\qquad \big(\sqrt{2\log(1+3\log T) - \log\delta} + \sqrt{2b} + 2a\big) \\ &\leqslant \frac{1}{2} + b + 3\log(1+3\log T) - \log\delta + 4a^2 \\ &\qquad\qquad + 2b + 6\log(1+3\log T) - 2\log\delta) \\ &\leqslant 1/2 + 3b + 4a^2 + 9\log(1+3\log T) - 3\log\delta\,. \\ &\lesssim 1 + b + a^2 + \log\log T - \log\delta\end{aligned}$$

However, in contrast to the bound B on the gradients, Lemma 8 only holds with probability $1-\delta_{i+1}$ (instead of almost surely). A union bound over all events states that the final result only holds with probability $1-\delta-\sum_{i=1}^{\infty}\delta_{i+1} = 1-2\delta$. To get a result with probability $1-\delta$, δ must thus be multiplied by 2 in the results.

This gives that, from the risk bound of Theorem 1, with probability $1-\delta$, $\mathrm{Risk}\big(\tilde{\theta}_t\big)$ is upper-bounded by

$$\min\left\{4U\left(\frac{X\sigma a'}{\sqrt{T}} + \frac{Bc'}{T}\right) + \frac{\alpha U^2}{8d_0T}, \right. \\ \left. \frac{d_0}{\alpha}\left(\frac{2^{10}X^2\sigma^2 a'^2}{T} + \frac{2^{11}B^2c'^2}{T^2}\right) + \frac{2\alpha U^2}{d_0T^2}\right\},$$

where $a' = 2a + 2\sqrt{6\log(1+3\log T) + 2\log(2/\delta)}$ and $c' = 1 + 3b + 4a^2 + 9\log(1+3\log T) + 3\log(2/\delta)$.

The bound of the theorem is then obtained by using that $B = 2X(Y + 2XU)$.

A.7 Proof of Theorem 4

For the sake of clarity, we only perform this proof up to universal constants. Let $B^* = 2X(Y + 2X\|\theta^*\|_1) \geqslant$

$\max_{\theta\in\mathcal{B}(0,2\|\theta^*\|_1)} \|\nabla\ell_t(\theta)\|_\infty$ almost surely. We also define by α^* the maximal number strong convexity parameter that satisfies (SC).

Let $T \geqslant 1$. Then, by definition (see Alg. 3), $\tilde{f}_{T-1} = \bar{f}_j$ for $j = \lfloor \log_2 T \rfloor - 1$.

We aim at controlling the excess risk of the average estimator $\bar{f}_j = \sum_{t=2^j}^{2^{j+1}-1} \widehat{f}_t$. To do so, we control the cumulative risk for $t = 2^j, \dots, 2^{j+1}-1$

$$\mathrm{Risk}^{(j)} := \sum_{t=2^j}^{2^{j+1}-1} \mathbb{E}_{t-1}\big[(Y_t - \widehat{f}_t(X_t))^2\big] - \mathbb{E}\big[(Y_t - X_t^\top\theta^*)^2\big]\,,$$

where $\mathbb{E}_{t-1}[\,\cdot\,] = \mathbb{E}\big[\,\cdot\,|(X_1,Y_1),\dots,(X_{t-1},Y_{t-1})\big]$. We will use that

$$\mathrm{Risk}(\bar{f}_j) \leqslant \mathrm{Risk}^{(j)}\, 2^{-j} \lesssim \frac{\mathrm{Risk}^{(j)}}{T}. \tag{35}$$

We first prove that it exists a predictor $f_{p,j}$ with $p \in \mathcal{G}_j$ that has a small excess risk. Then, we will apply Theorem 4.5 of Wintenberger [2014] to show that BOA almost achieves this performance.

Step 1. Either it exists a predictor $f_{p,j}$ with small excess risk or $\mathrm{Risk}^{(j)}$ is small. Since all predictions $\widehat{f}_t(X_t)$ lie in $[-Y,Y]$ almost surely,

$$\mathrm{Risk}^{(j)} \leqslant Y^2 2^j \leqslant Y^2 T\,. \tag{36}$$

Let d_0 in $\mathcal{G}_j$ (i.e., a power of 2) such that $d_0/2 \leqslant \|\theta^*\|_0 \leqslant d_0$. We show that if the conditions of Theorem 3 cannot be satisfied with any parameter of the grid $\mathcal{G}_j$, the cumulative risk $\mathrm{Risk}^{(j)}$ is small enough. We start with the choice of the parameter U, which should be of order $\|\theta^*\|_1$:

a) If $\|\theta^*\|_1 \leqslant 2^{-2j}$. It exists a predictor in $\mathcal{G}_j$ such that $f_{p,j} = 0$ (consider $d_0 = 0$). In this case,

$$\mathrm{Risk}(f_{p,j}) = \mathbb{E}[(Y_t - 0)^2] \leqslant B^*\|\theta^*\|_1 \leqslant B^* 2^{-2j} \lesssim B^* T^{-2}\,,$$

where we used that $2^{-j} \lesssim T^{-1}$.

b) If $\|\theta^*\|_1 \geqslant 2^{2j+\lceil 2\log Y\rceil}$, then $2^j \leqslant Y^{-2}\|\theta^*\|_1 2^{-j}$ and from Inequality (36),

$$\mathrm{Risk}^{(j)} \leqslant \|\theta^*\|_1 2^{-j} \lesssim \frac{\|\theta^*\|_1}{T} \lesssim \frac{\|\theta^*\|_0 (B^*)^2}{\alpha^* T}\,.$$

Otherwise, we can choose U in $\mathcal{G}_j$ such that $U/2 \leqslant \|\theta^*\|_1 \leqslant U$. Similarly for B:

c) if $B < 2^{-2j}$, then for $f_{p,j} = 0$,

$$\mathrm{Risk}(f_{p,j}) = \mathbb{E}[\ell(Y_t,0)] \leqslant B^*\|\theta^*\|_1 \leqslant \|\theta^*\|_1 2^{-2j} \lesssim \frac{\|\theta^*\|_1}{T^2}\,,$$

d) if $B > 2^{2j+\lceil 2\log Y\rceil}$, then from Inequality (36), $\mathrm{Risk}^{(j)} \leqslant B^* 2^{-j} \lesssim B^* T^{-1}$.

Otherwise, we can choose B in $\mathcal{G}_j$ such that $B/2 \leqslant B^* \leqslant B$. Finally, for α:

e) if $\alpha^* < 2^{-2j+\lceil \log_2(B^2 d_0/Y^2)\rceil} \leqslant d_0 B^2 2^{-2j}/Y^2$, then $2^j \leqslant d_0 B^2/(Y^2\alpha^* 2^j)$ and thus

$$\mathrm{Risk}^{(j)} \leqslant Y^2 2^j \leqslant Y^2 \frac{d_0 B^2}{Y^2\alpha^* 2^j} \lesssim \frac{\|\theta^*\|_0 (B^*)^2}{\alpha^* T}\,.$$

Otherwise, we can choose α in $\mathcal{G}_j$ such that $\min\{d_0/T, \alpha^*/2\} \leqslant \alpha \leqslant \alpha^*$.

f) Applying Theorem 3, with high probability the excess risk of the estimator $f_{p,j}$ with the choice (d_0, α, U, B) described above satisfies

$$\begin{aligned}\mathrm{Risk}(f_{p,j}) &\overset{\text{clipping}}{\leqslant} \mathrm{Risk}(\tilde{\theta}_{p,j}) \\ &\lesssim \min\left\{ \frac{X^2}{\gamma}\left(\frac{\sigma^2 a'^2}{T} + \frac{(Y + X\|\theta^*\|_1)^2 c'^2}{T^2}\right) + \frac{\gamma\|\theta^*\|_1^2}{T^2}, \right. \\ &\qquad \left. \|\theta^*\|_1 X\left(\frac{\sigma a'}{\sqrt{T}} + \frac{(Y + X\|\theta^*\|_1)c'}{T}\right) + \frac{\gamma\|\theta^*\|_1^2}{T} \right\},\end{aligned}$$

with $\gamma = \max\{d_0/\alpha, 1/T\}$.

Putting everything together, either (for cases b), d), and e))

$$\mathrm{Risk}^{(j)} \lesssim \Big(B^* + \frac{\|\theta^*\|_0 (B^*)^2}{\alpha^*}\Big) T^{-1} \tag{37}$$

or, for cases a), c), and f), there exists $p \in \mathcal{G}_j$ such that with high probability

$$\begin{aligned}\mathrm{Risk}(f_{p,j}) &\lesssim \min\left\{ \frac{1}{\gamma}\left(\frac{X^2\sigma^2 a'^2}{T} + \frac{(B^* c')^2}{T^2}\right) + \frac{\gamma\|\theta^*\|_1^2}{T^2}, \right. \\ &\left. \|\theta^*\|_1 X\left(\frac{\sigma a'}{\sqrt{T}} + \frac{(Y + X\|\theta^*\|_1)c'}{T}\right) + \frac{\gamma\|\theta^*\|_1^2}{T} \right\} + \frac{B^*}{T^2}\,,\end{aligned} \tag{38}$$

Step 2. Bound of the meta-algorithm. Using that the square loss is $4Y$-Lipschitz over the domain $[-2Y, 2Y]$ and 2-strongly convex, we can apply Theorem 4.5 of Wintenberger [2014] with $C_b = 4Y$, $C_\ell = 2$,

and $M = \#\mathcal{G}_j$. We get that with high enough probability

$$\mathrm{Risk}^{(j)} \lesssim T \min_{p\in\mathcal{G}_j} \mathrm{Risk}(f_{p,j}) + Y^2\big(\log \#\mathcal{G}_j + \log(\log T + \log Y) - \log\delta\big)\,.$$

Substituting

$$\#\mathcal{G}_j \lesssim (j+\log Y)^3 \log d \lesssim (\log T + \log Y)^3 \log d\,,$$

this yields

$$\mathrm{Risk}^{(j)} \lesssim \frac{\min_{p\in\mathcal{G}_j} \mathrm{Risk}(f_{p,j})}{T} + Y^2\big(\log\log d + \log(\log T + \log Y) - \log\delta\big)\,.$$

Combining with Inequality (38), we obtain that $\mathrm{Risk}^{(j)}$ is at most of order

$$\mathrm{Risk}^{(j)} \lesssim Y^2\big(\log\log d + \log(\log T + \log Y) - \log\delta\big) + \min\left\{\frac{1}{\gamma}\left(X^2\sigma^2 a'^2 + \frac{(B^* c')^2}{T}\right) + \frac{\gamma\|\theta^*\|_1^2}{T}, \|\theta^*\|_1 X\left(\sigma a'\sqrt{T} + (Y + X\|\theta^*\|_1)c'\right) + \gamma\|\theta^*\|_1^2\right\} + \frac{B^*}{T}.$$

Finally, using Inequality (35), keeping only the main asymptotic term in $1/T$, and substituting $a' \lesssim \log((d\log T)/\delta)$ concludes the proof.

B Martingale inequalities

In this section, we prove two martingale inequalities that are used in the analysis.

B.1 Poissonian inequality

First, we prove a Poissonian inequality which only works for nonnegative increments.

Theorem 9. *Let $T \geqslant 1$. Let $(X_t)_{t\geqslant 1}$ be a sequence of random variables such that $X_t \in [0,B]$ almost surely, then with probability at least $1-\delta$*

$$\sum_{t=1}^T X_t \leqslant (e-1)\sum_{t=1}^T \mathbb{E}_{t-1}[X_t] + B\log(1/\delta)\,.$$

Proof. Let $Z_t = X_t/B \in [0,1]$. From [Cesa-Bianchi and Lugosi, 2006, Lemma A.3], for all $t \geqslant 1$, and all $s > 0$

$$\mathbb{E}_{t-1}\Big[\exp\big(sZ_t - (e^s-1)\mathbb{E}_{t-1}[Z_t]\big)\Big] \leqslant 1\,.$$

Thus,

$$\begin{aligned}\mathbb{E}\Big[\exp\Big(s\sum_{t=1}^T Z_t - (e^s-1)\sum_{t=1}^T \mathbb{E}_{t-1}[Z_t]\Big)\Big] &= \mathbb{E}\Big[\mathbb{E}_{T-1}\Big[\exp\big(sZ_T - (e^s-1)\mathbb{E}_{T-1}[Z_T]\big)\Big] \exp\Big(s\sum_{t=1}^{T-1} Z_t - (e^s-1)\sum_{t=1}^{T-1}\mathbb{E}_{t-1}[Z_t]\Big)\Big] \\ &\leqslant \mathbb{E}\Big[\exp\Big(s\sum_{t=1}^{T-1} Z_t - (e^s-1)\sum_{t=1}^{T-1}\mathbb{E}_{t-1}[Z_t]\Big)\Big]\end{aligned}$$

By induction, we get

$$\mathbb{E}\Big[\exp\Big(s\sum_{t=1}^T Z_t - (e^s-1)\sum_{t=1}^T \mathbb{E}_{t-1}[Z_t]\Big)\Big] \leqslant 1\,.$$

We conclude thanks to Markov's inequality, with probability at least $1-\delta$

$$\sum_{t=1}^T Z_t \leqslant \frac{e^s-1}{s}\sum_{t=1}^T \mathbb{E}_{t-1}[Z_t] + \frac{1}{s}\log(1/\delta)\,.$$

The final result is obtained by substituting $Z_t = X_t/B$ and by choosing $s = 1$. □

B.2 From cumulative regret to cumulative risk

Theorem 10. *Let $x > 0$. Assume $\theta^* \in \mathcal{B}_1(\theta_{\mathrm{center}}, \varepsilon)$. The cumulative risk of any convex optimization procedure in $\mathcal{B}_1(\theta_{\mathrm{center}},\varepsilon)$ satisfies, with probability $1-\delta$*

$$\mathrm{Risk}_{1:T}(\widehat{\theta}_{0:(T-1)}) - \mathrm{Reg}_{1:T}(\widehat{\theta}_{0:(T-1)}) \leqslant \varepsilon\sqrt{2\log\Big(\frac{2+\log(T/2)}{2\delta}\Big)\sum_{t=1}^T \big\|\nabla\ell_t(\widehat{\theta}_{t-1})\big\|_\infty^2} + \Big(\frac{1}{2} + \log\big(1+\frac{1}{2}\log(T/2)\big) - \log\delta\Big)\varepsilon B\,,$$

where $B \geqslant \max_{\theta\in\mathcal{B}_1(\theta_{\mathrm{center}},\varepsilon)} \big\|\nabla\ell_t(\widehat{\theta}_{t-1})\big\|_\infty$ almost surely.

Proof. This is a consequence of Theorem 4.1 of Wintenberger [2014]. Let $\delta \in (0,1)$ and $(\eta_t)_{t\geqslant 0}$ be a sequence adapted to the filtration $(\mathcal{F}_t = \{\ell_1,\dots,\ell_{t-1}\})_{t\geqslant 0}$. Then, with the notation $\ell_{j,t}^2 \leqslant \varepsilon^2\|\nabla\ell_t(\widehat{\theta}_{t-1})\|_\infty^2$, applying Theorem 4.1 of Wintenberger [2014], we get that with probability $1-\delta$

$$
\begin{aligned}
R_T &:= \mathrm{Risk}_{1:T}(\widehat{\theta}_{0:(T-1)}) - \mathrm{Reg}_{1:T}(\widehat{\theta}_{0:(T-1)}) \\
&\leqslant \varepsilon^2 \sum_{t=1}^{T} \eta_{t-1} \|\nabla \ell_t(\widehat{\theta}_{t-1})\|_\infty^2 \\
&\quad + \frac{\log\Big(1 + \mathbb{E}\big[\log(\eta_1/\eta_T)\big]\Big) - \log\delta}{\eta_T}\,, \qquad (39)
\end{aligned}
$$

where $\mathrm{Risk}_{1:T}(\widehat{\theta}_{0:(T-1)}) := \sum_{t=1}^{T} \mathbb{E}[\ell_t](\widehat{\theta}_{t-1}) - \mathbb{E}[\ell_t](\theta^*)$ and $\mathrm{Reg}_{1:T}(\widehat{\theta}_{0:(T-1)}) := \sum_{t=1}^{T} \ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta^*)$.

We obtain the stated inequality from (39), by properly setting the tuning parameters

$$
\eta_t := \frac{1}{\varepsilon} \min\left\{ \frac{1}{B}, \frac{c\Gamma}{V_{t-1}} \right\},
$$

where c will be set by the analysis and

$$
\Gamma := \sqrt{\log\big(1 + \log(\sqrt{T}/c)\big) - \log\delta}\,,
$$

and

$$
V_{t-1} := \sqrt{\sum_{s=1}^{t-1} \big\|\nabla \ell_s(\widehat{\theta}_{s-1})\big\|_\infty^2}\,.
$$

Indeed, first we use that that $\eta_1/\eta_T \leqslant \sqrt{T}/c$ so that $\mathbb{E}[\log(\eta_1/\eta_T)] \leqslant \log(\sqrt{T}/c)$. Then, similarly to the proof of [Cesa-Bianchi et al., 2007, Theorem 5], we can show that the first term in the right-hand side of (39) is upper-bounded as

$$
\sum_{t=1}^{T} \eta_{t-1} \|\nabla \ell_t(\widehat{\theta}_{t-1})\|_\infty^2 \leqslant \frac{B}{2\varepsilon} + \frac{c\Gamma}{2\varepsilon} V_T.
$$

But, by definition of η_T, the second term is also controlled as

$$
\frac{\log\Big(1 + \mathbb{E}\big[\log(\eta_1/\eta_T)\big]\Big) - \log\delta}{\eta_T} \leqslant \varepsilon\Gamma \max\left\{ B\Gamma, \frac{1}{c} V_T \right\}.
$$

Plugging these two last inequalities into (39) leads to

$$
R_T \leqslant \frac{B\varepsilon}{2} + \frac{c\Gamma\varepsilon}{2} V_T + \varepsilon\Gamma \max\left\{ B\Gamma, \frac{V_T}{c} \right\}.
$$

We then need to distinguish two cases

- if $c\Gamma B \leqslant V_T$, then optimizing in $c = \sqrt{2}$

$$
R_T \leqslant \frac{B\varepsilon}{2} + \Big(\frac{c}{2} + \frac{1}{c}\Big)\Gamma\varepsilon V_T \leqslant \frac{B\varepsilon}{2} + \sqrt{2}\Gamma\varepsilon V_T
$$

- if $c\Gamma B \geqslant V_T$, then

$$
R_T \leqslant \frac{B\varepsilon}{2} + \frac{1}{\sqrt{2}}\Gamma\varepsilon V_T + \varepsilon B\Gamma^2\,.
$$

Therefore, putting the two cases together

$$
R_T \leqslant \frac{B\varepsilon}{2} + \sqrt{2}\Gamma\varepsilon V_T + \varepsilon B\Gamma^2\,.
$$

We conclude the proof by substituting Γ and V_T with their definitions. □