Sparse Accelerated Exponential Weights

Pierre Gaillard
pierre.gaillard@inria.fr
INRIA - Sierra project-team,
Département d'Informatique de
l'Ecole Normale Supérieure, Paris, France
University of Copenhagen, Denmark

Olivier Wintenberger
olivier.wintenberger@upmc.fr
Sorbonne Universités,
UPMC Univ Paris 06 LSTA, France
University of Copenhagen, Denmark

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 random environment with convex losses 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 T^d). 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) 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 Gerchinovitz (2013), one

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

suspects that the algorithm is NP-hard.

Procedure	Setting	Rate	Polynomial
Lasso (Bunea et al., 2007)	B	$\frac{d_0 \log d}{\alpha T}$	Yes
Rigollet and Tsybakov (2011)	B	$\frac{d_0 \log d}{T}$	No
SeqSEW (Gerchinovitz, 2013)	S	$\frac{d_0 \log(d/d_0)}{T}$	No
ℓ_1-RDA method (Xiao, 2010)	S	$\frac{d}{T}$	Yes
SAEW	S	$\frac{d_0 \log d}{\alpha T}$	Yes

Table 1: Comparison of sequential (S) and batched (B) sparse optimization procedures.

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. Table 1 summarizes the performance of existing algorithms in sparse regression.

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).

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 B.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

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

[†]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 $\mathbb{E}\big[\ell_t(\widehat{\theta}_{t-1}) \big| \ell_1, \ldots, \ell_{t-1}\big]$.

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 $\nabla \ell_s(\widehat{\theta}_{s-1})$, $s < t$.

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 weaker 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 $\mathcal{B}_1\big(\theta_{\text{center}}, \varepsilon\big) := \big\{\theta \in \mathbb{R}^d : \|\theta - \theta_{\text{center}}\|_1 \leqslant \varepsilon\big\}$ of center $\theta_{\text{center}} \in \mathbb{R}^d$ and radius $\varepsilon > 0$. Centers and radii will be settled online thanks to SAEW.

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

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

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

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

We assume that the subroutine $\mathcal{S}$ applied on any sequence of convex sub-differentiable losses $(\ell_t)_{t \geqslant t_{\text{start}}}$ satisfies the following upper-bound on its cumulative regret: for all $t_{\text{end}} \geqslant t_{\text{start}}$ and for all $\theta \in \mathcal{B}_1\big(\theta_{\text{center}}, \varepsilon\big)$

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

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), and 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 $\mathcal{B}_1\big(\theta_{\text{center}}, \varepsilon\big)$. 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.

The regret bound yields the slow rate of convergence $\mathcal{O}\big(\sqrt{(\log d)(t_{\text{end}} - t_{\text{start}})}\big)$ (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 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

[‡] 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 (9) and (10).
  - 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 $t_{i+1} := t + 1$

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 $\mathcal{O}(dT)$.

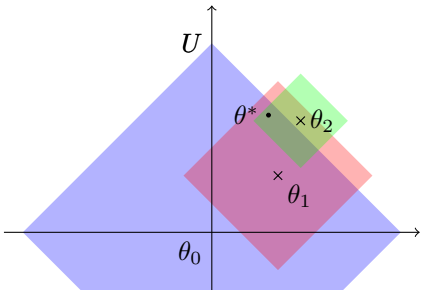


Figure 1: The acceleration procedure. First, the algorithm perfoms the optimization in the the blue ball of radius U and centered at the origin. Then, when the confidence is high enough, at time t_1, the algorithm is restarted to the red ball of radius $U/2$ and centered at the current estimator $\theta_1 := \big[\bar{\theta}_{t_1-1}\big]_{d_0}$. The process is repeated with $\theta_2 := \big[\bar{\theta}_{t_2-1}\big]_{d_0}$ and $U/4$ and so forth.

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

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

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 B.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. In Appendix A, We provide 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. The proof is however only done in the setting of linear regression with square loss.

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 method of Kivinen and Warmuth (1997) to extend it to the 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.

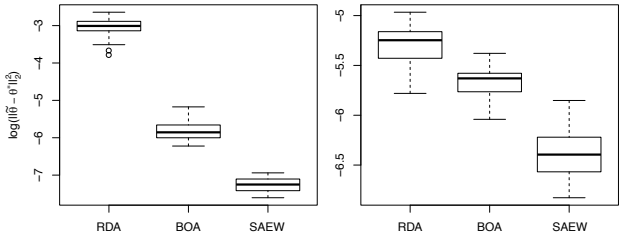
Our objective here is only to show the potential of the acceleration of BOA for a well-chosen set of parameters in the general setting of Section 3.

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

Figure 2a 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 2a). 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.

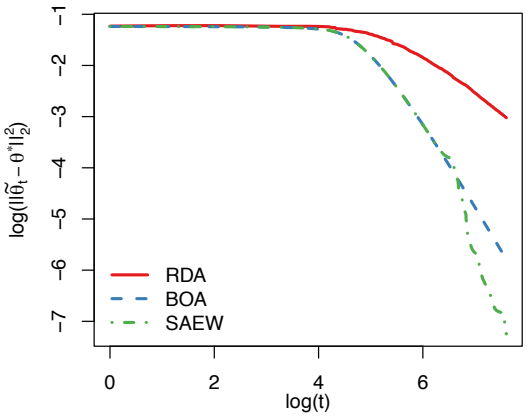


Figure 3: Averaged (over 30 experiments) evolution of the logarithm of the ℓ_2 error.

Figure 3 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 4. SAEW and BOA seem to achieve logarithmic cumulative risk, in contrast to RDA which seems to be of order $\mathcal{O}(\sqrt{T})$.

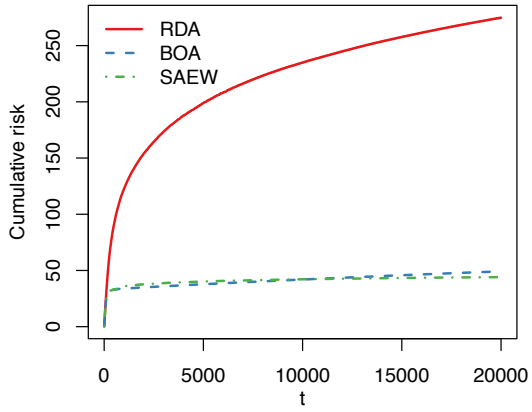


Figure 4: Averaged (over 30 runs) cumulative risk suffered by $\widehat{\theta}_t$ for square linear regression.

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 dimension 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 5 shows that despite what seems on Figure 4, BOA does not achieve fast rate on its cumulative risk.

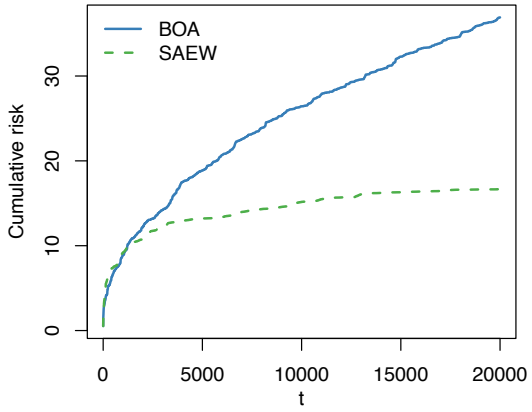


Figure 5: Cumulative risk suffered by $\widehat{\theta}_t$ for square linear regression with $d = d_0 = 2$.

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 2b 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 6 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 6. However, we only show here the dependence on t and not in d.

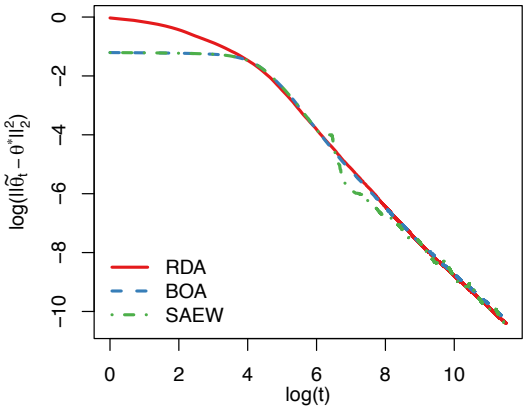


Figure 6: Averaged (over 30 runs) evolution of the logarithm of the ℓ_2-error for quantile regression ($d = 100$).

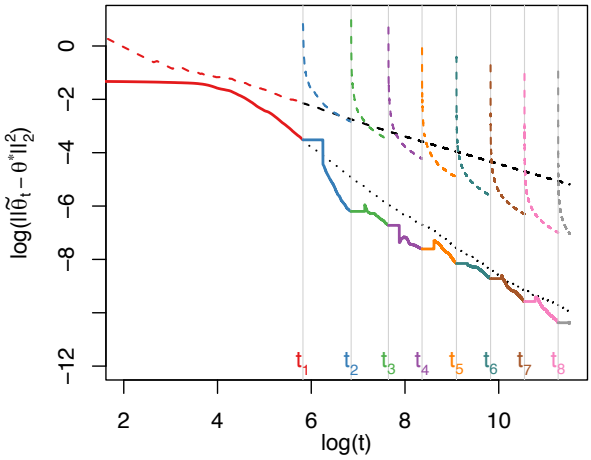


Figure 7: Logarithm of the ℓ_2-norm of the averaged estimator $\tilde{\theta}_t$ during one run. The dashed lines represent the high probability ℓ_2-bound estimated by SAEW on $\bar{\theta}_t$. The gray vertical lines are the stopping times t_i, $i \geqslant 1$. The first session is plotted in red, the second in blue,... The dotted and dashed black lines represent the performance (and the theoretical bound) that BOA would have obtained without acceleration.

In Figure 7, 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 (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. In order to control the risk of the

averaged estimator, the standard proof technique (in online learning) applies Jensen's inequality to the cumulative risk (see Inequality (2)). Since the later does not achieve the desired fast rate convergence for BOA (as observed in Figure 8), Jensen's inequality fails here to prove fast convergence for the averaged estimator and new proof techniques will be needed.

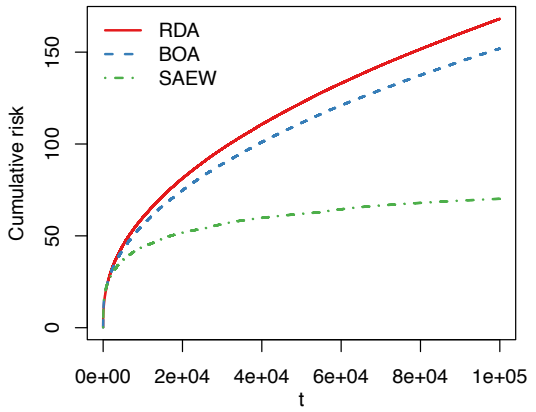


Figure 8: Averaged (over 30 runs) cumulative risk suffered by $\widehat{\theta}_t$ for quantile regression ($d = 100$).

References

Agarwal, A., P. L. Bartlett, P. Ravikumar, and M. J. Wainwright (2012). "Information-Theoretic Lower Bounds on the Oracle Complexity of Stochastic Convex Optimization." In: *IEEE TRANSACTIONS ON INFORMATION THEORY* 58.5, p. 3235.

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

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

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

Bunea, F., A. Tsybakov, and M. Wegkamp (2007). "Aggregation for Gaussian regression." In: *The Annals of Statistics* 35.4, pp. 1674–1697.

Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization. Ecole dEté de Probabilités de Saint-Flour 2001, Lectures Notes in Mathematics 1851*.

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

Cesa-Bianchi, N., Y. Mansour, and G. Stoltz (2007). "Improved second-order bounds for prediction with expert advice." In: *Machine Learning* 66.2-3, pp. 321–352.

Gaillard, P., G. Stoltz, and T. van Erven (2014). "A Second-order Bound with Excess Losses." In: *Proceedings of COLT'14*. Vol. 35. JMLR: Workshop and Conference Proceedings, pp. 176–196.

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

– (2013). "Sparsity regret bounds for individual sequences in online linear regression." In: *The Journal of Machine Learning Research* 14.1, pp. 729–769.

Kivinen, J. and M. K. Warmuth (1997). "Exponentiated Gradient Versus Gradient Descent for Linear Predictors." In: *Information and Computation* 132.1, pp. 1–63.

Koenker, R. W. and G. W. Bassett (1978). "Regression Quantiles." In: *Econometrica* 46.1, pp. 33–50.

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

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). "Gradient-based learning applied to document recognition." In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Nemirovski, A. (2000). "Topics in non-parametric." In: *Ecole dEté de Probabilités de Saint-Flour* 28, p. 85.

Polyak, B. and A. Juditsky (1992). "Acceleration of stochastic approximation by averaging." In: *SIAM Journal on Control and Optimization* 30.4, pp. 838–855.

Rigollet, P. and A. Tsybakov (2011). "Exponential screening and optimal rates of sparse estimation." In: *The Annals of Statistics*, pp. 731–771.

Steinwart, I. and A. Christmann (2011). "Estimating conditional quantiles with the help of the pinball loss." In: *Bernoulli* 17.1, pp. 211–225.

Wintenberger, O. (2014). "Optimal learning with Bernstein Online Aggregation." In: Extended version available at arXiv:1404.1356 [stat. ML].

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

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

Zinkevich, M. (2003). "Online Convex Programming and Generalized Infinitesimal Gradient Ascent." In: *Proceedings of the 20th International Conference on Machine Learning, ICML 2003*.