ONLINE ACCELERATION OF EXPONENTIAL WEIGHTS

Pierre Gaillard

February 10, 2017

INRIA Paris – SIERRA

This is joint work with Olivier Wintenberger

SETTING

Step by step minimization of i.i.d. convex loss functions[1] $\ell_1, \dots, \ell_n : \mathbb{R}^d \to \mathbb{R}$.

Assumption 1 (strongly convex risk)

$$\exists \alpha > 0, \theta^* \in \mathbb{R}^d,\ \forall \theta \in \mathbb{R}^d \qquad \alpha \big\|\theta - \theta^*\|_2^2 \leqslant \mathbb{E}\big[\ell_t(\theta) - \ell_t(\theta^*)\big]\,.$$

Setting: for each $t = 1, \dots, n$

- the learner provides $\widehat{\theta}_{t-1} \in \mathbb{R}^d$ based on past gradients $\nabla \ell_s(\widehat{\theta}_{s-1})$ for $s \leqslant t-1$
- the environment reveals $\nabla \ell_t(\widehat{\theta}_{t-1})$

Goal: minimize the average risk:

$$\mathrm{Risk}_n(\widehat{\theta}_{0:(n-1)}) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[\ell_t](\widehat{\theta}_{t-1}) - \mathbb{E}[\ell_t](\theta^*)\,.$$

Remark 1

- non-Lipschitz gradients
- only the risk needs to be strongly convex (pinball loss)

[1] Cesa-Bianchi and Lugosi, *Prediction, Learning, and Games*, 2006.

Step by step minimization of i.i.d. convex loss functions[1] $\ell_1, \dots, \ell_n : \mathbb{R}^d \to \mathbb{R}$.

Assumption 1 (strongly convex risk)

$$\exists \alpha > 0, \theta^* \in \mathbb{R}^d,\ \forall \theta \in \mathbb{R}^d \qquad \alpha \big\|\theta - \theta^*\|_2^2 \leqslant \mathbb{E}\big[\ell_t(\theta) - \ell_t(\theta^*)\big]\,.$$

Setting: for each $t = 1, \dots, n$

- the learner provides $\widehat{\theta}_{t-1} \in \mathbb{R}^d$ based on past gradients $\nabla \ell_s(\widehat{\theta}_{s-1})$ for $s \leqslant t-1$
- the environment reveals $\nabla \ell_t(\widehat{\theta}_{t-1})$

Goal: minimize the average risk:

$$\mathrm{Risk}_n(\widehat{\theta}_{0:(n-1)}) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[\ell_t](\widehat{\theta}_{t-1}) - \mathbb{E}[\ell_t](\theta^*)\,.$$

Remark 2 By convexity of the risk, the averaging $\bar{\theta}_{n-1} := (1/n)\sum_{t=1}^n \widehat{\theta}_{t-1}$ has an instantaneous risk upper-bounded by the cumulative risk

$$\mathrm{Risk}(\bar{\theta}_{n-1}) := \mathbb{E}[\ell_n](\bar{\theta}_{n-1}) - \mathbb{E}[\ell_n](\theta^*) \overset{\text{Jensen}}{\leqslant} \mathrm{Risk}_n(\widehat{\theta}_{0:(n-1)})\,.$$

[1] Cesa-Bianchi and Lugosi, *Prediction, Learning, and Games*, 2006.

Step by step minimization of i.i.d. convex loss functions[1] $\ell_1, \dots, \ell_n : \mathbb{R}^d \to \mathbb{R}$.

Assumption 1 (strongly convex risk)

$$\exists \alpha > 0, \theta^* \in \mathbb{R}^d, \; \forall \theta \in \mathbb{R}^d \qquad \alpha \big\| \theta - \theta^* \|_2^2 \leqslant \mathbb{E}\big[\ell_t(\theta) - \ell_t(\theta^*)\big] \,.$$

θ^* is d_0-sparse $\qquad \|\theta^*\|_1 \leqslant U$

Setting: for each $t = 1, \dots, n$

- the learner provides $\widehat{\theta}_{t-1} \in \mathbb{R}^d$ based on past gradients $\nabla \ell_s(\widehat{\theta}_{s-1})$ for $s \leqslant t-1$
- the environment reveals $\nabla \ell_t(\widehat{\theta}_{t-1})$

Goal: minimize the average risk:

$$\mathrm{Risk}_n(\widehat{\theta}_{0:(n-1)}) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[\ell_t](\widehat{\theta}_{t-1}) - \mathbb{E}[\ell_t](\theta^*) \,.$$

Remark 2 By convexity of the risk, the averaging $\bar{\theta}_{n-1} := (1/n) \sum_{t=1}^{n} \widehat{\theta}_{t-1}$ has an instantaneous risk upper-bounded by the cumulative risk

$$\mathrm{Risk}(\bar{\theta}_{n-1}) := \mathbb{E}[\ell_n](\bar{\theta}_{n-1}) - \mathbb{E}[\ell_n](\theta^*) \overset{\text{Jensen}}{\leqslant} \mathrm{Risk}_n(\widehat{\theta}_{0:(n-1)}) \,.$$

[1] Cesa-Bianchi and Lugosi, *Prediction, Learning, and Games*, 2006.

THE FRAMEWORK OF THIS TALK

Step by step minimization of i.i.d. convex loss functions $\ell_1, \dots, \ell_n : \mathbb{R}^d \to \mathbb{R}$.

Assumption 1 (strongly convex risk)

$$\exists \alpha > 0, \theta^* \in \mathbb{R}^d, \ \forall \theta \in \mathbb{R}^d \qquad \alpha \big\|\theta - \theta^*\|_2^2 \leqslant \mathbb{E}\big[\ell_t(\theta) - \ell_t(\theta^*)\big]\,.$$

θ^* is d_0-sparse $\qquad \|\theta^*\|_1 \leqslant U$

Setting: for each $t = 1, \dots, n$

- the learner provides $\widehat{\theta}_{t-1} \in \mathbb{R}^d$ based on past gradients $\nabla \ell_s(\widehat{\theta}_{s-1})$ for $s \leqslant t-1$
- the environment reveals $\nabla \ell_t(\widehat{\theta}_{t-1})$

Goal: minimize the average risk:

$$\mathrm{Risk}_n(\widehat{\theta}_{0:(n-1)}) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[\ell_t](\widehat{\theta}_{t-1}) - \mathbb{E}[\ell_t](\theta^*)\,.$$

Result

$$\mathrm{Risk}_n(\widehat{\theta}_{0:(n-1)}) \lesssim \min\left\{\frac{B^2 d_0 \log d \log n}{\alpha n}, UB\sqrt{\frac{\log d}{n}}\right\}$$

where $B \geqslant \max_{\|\theta\|_1 \leqslant 2U} \big\|\nabla \ell_t(\theta)\big\|_\infty$.

THE FRAMEWORK OF THIS TALK

Step by step minimization of i.i.d. convex loss functions $\ell_1, \dots, \ell_n : \mathbb{R}^d \to \mathbb{R}$.

Assumption 1 (strongly convex risk)

$$\exists \alpha > 0, \theta^* \in \mathbb{R}^d, \ \forall \theta \in \mathbb{R}^d \qquad \alpha \big\|\theta - \theta^*\|_2^2 \leqslant \mathbb{E}\big[\ell_t(\theta) - \ell_t(\theta^*)\big] \, .$$

θ^* is d_0-sparse $\qquad \|\theta^*\|_1 \leqslant U$

Setting: for each $t = 1, \dots, n$

- the learner provides $\widehat{\theta}_{t-1} \in \mathbb{R}^d$ based on past gradients $\nabla \ell_s(\widehat{\theta}_{s-1})$ for $s \leqslant t-1$
- the environment reveals $\nabla \ell_t(\widehat{\theta}_{t-1})$

Goal: minimize the average risk:

$$\mathrm{Risk}_n(\widehat{\theta}_{0:(n-1)}) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[\ell_t](\widehat{\theta}_{t-1}) - \mathbb{E}[\ell_t](\theta^*) \, .$$

Result

$$\mathrm{Risk}_n(\widehat{\theta}_{0:(n-1)}) \lesssim \min \left\{ \frac{B^2 d_0 \log d \log n}{\alpha n}, UB\sqrt{\frac{\log d}{n}} \right\}$$

where $B \geqslant \max_{\|\theta\|_1 \leqslant 2U} \big\|\nabla \ell_t(\theta)\big\|_\infty$.

Fast rate : better for large n, α

Slow rate : better for small n, α

COMPARISON WITH OTHER SPARSE ALGORITHMS

Procedure	Sequential	Rate	Polynomial
Lasso[1]	✗	$\frac{d_0 \log d}{\alpha n}$	✓
EWA + sparsity patern[2]	✗	$\frac{d_0 \log d}{n}$	✗
SeqSEW[3]	✓	$\frac{d_0 \log d}{n}$	✗
ℓ_1-RDA method[4]	✓	$\frac{d}{n}$	✓
SAEW	✓	$\frac{d_0 \log d}{\alpha n}$	✓

1 Bunea, Tsybakov, and Wegkamp, "Aggregation for Gaussian regression", 2007.
2 Rigollet and Tsybakov, "Exponential screening and optimal rates of sparse estimation", 2011.
3 Gerchinovitz, "Sparsity regret bounds for individual sequences in online linear regression", 2013.
4 Xiao, "Dual averaging methods for regularized stochastic learning and online optimization", 2010.

CONVEX OPTIMIZATION IN THE ℓ_1-BALL WITH SLOW RATE

Goal: perform online optimization in

$$\mathcal{B}_1\big(\theta_{\mathrm{center}}, \varepsilon\big) := \big\{\theta \in \mathbb{R}^d : \|\theta - \theta_{\mathrm{center}}\|_1 \leqslant \varepsilon\big\}$$

We define the $2d$ corners of the ℓ_1-ball $e_k = \theta_{\mathrm{center}} \pm \varepsilon(0, \ldots, 0, 1, 0, \ldots, 0)$

The exponentially gradient forecaster ($EG^{\pm}$)[5]

At each forecasting instance $t \geqslant 1$,

- assign to each corner e_k the weight

$$\widehat{p}_{k,t-1} = \frac{\exp\Big(-\eta \sum_{s=1}^{t-1} \nabla \ell_s(\widehat{\theta}_{s-1})^\top e_k\Big)}{\sum_{j=1}^{2d} \exp\Big(-\eta \sum_{s=1}^{t-1} \nabla \ell_s(\widehat{\theta}_{s-1})^\top e_j\Big)}$$

- form parameter $\widehat{\theta}_{t-1} = \sum_{k=1}^{2d} \widehat{p}_{k,t-1} e_k$

Performance: bound on the average regret, if $\theta^* \in \mathcal{B}_1(\theta_{\mathrm{center}}, \varepsilon)$ for η well-tuned

$$\sum_{t=1}^{n} \ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta^*) \lesssim \varepsilon B \sqrt{\frac{\log d}{n}}\,.$$

The learning rate η can be tuned online (doubling trick, η_t).

[5] Kivinen and Warmuth, “Exponentiated Gradient Versus Gradient Descent for Linear Predictors”, 1997.

PROOF

Lemma (Hoeffding)

If X is a random variable with $|X| \leqslant B$. Then,

$$\forall \eta \in \mathbb{R}, \qquad \mathbb{E}[X] \leqslant -\frac{1}{\eta}\log\left(\mathbb{E}\left[e^{-\eta X}\right]\right) + \frac{\eta B}{4}\,.$$

1. Upper bound the instantaneous gradient

$$\begin{aligned}
\nabla \ell_t(\widehat\theta_{t-1})^\top \widehat\theta_{t-1} \quad &\overset{\text{def. of } \widehat\theta_{t-1}}{=} \quad \sum_{k=1}^{2d} \widehat p_{k,t-1}\big(\nabla \ell_t(\widehat\theta_{t-1})^\top e_k\big) \\
&\overset{\text{Hoeffding}}{\leqslant} \quad -\frac{1}{\eta}\log\left(\sum_{k=1}^{K} \widehat p_{k,t} e^{-\eta \nabla \ell_t(\widehat\theta_{t-1})^\top e_k}\right) + \frac{\eta B}{4} \\
&\overset{\text{def. of } \widehat p_{k,t+1}}{=} \quad -\frac{1}{\eta}\log\left(\frac{\widehat p_{k,t}}{\widehat p_{k,t+1}} e^{-\eta \nabla \ell_t(\widehat\theta_{t-1})^\top e_k}\right) + \frac{\eta B}{4} \\
&= \quad \nabla \ell_t(\widehat\theta_{t-1})^\top e_k + \frac{1}{\eta}\log\frac{\widehat p_{k,t+1}}{\widehat p_{k,t}} + \frac{\eta B}{4}\,.
\end{aligned}$$

2. Sum over all t, the sum telescopes

$$\begin{aligned}
\sum_{t=1}^{n} \ell_t(\widehat\theta_{t-1}) - \ell_t(\widehat\theta^*) \overset{\text{Jensen}}{\leqslant} \sum_{1}^{n} \nabla \ell_t(\widehat\theta_{t-1})^\top\left(\widehat\theta_{t-1} - \theta^*\right) &\leqslant \max_{k=1,\dots,2d}\left\{\sum_{1}^{n} \nabla \ell_t(\widehat\theta_{t-1})^\top\left(\widehat\theta_{t-1} - e_k\right)\right\} \\
&\leqslant \max_{k=1,\dots,2d}\left\{\frac{1}{\eta}\log\frac{\cancel{\widehat p_{k,n}}}{\widehat p_{k,0}} + \frac{\eta B n}{4}\right\} \leqslant \frac{\log(2d)}{\eta} + \frac{\eta B n}{4}
\end{aligned}$$

Theorem

Let $0 < \delta < 1$, *then if* $\theta^* \in \mathcal{B}_1(\theta^*, \varepsilon)$, *EG*$^{\pm}$ *satisfies*

$$\alpha \left\| \bar{\theta}_{n-1} - \theta^* \right\|_2^2 \overset{\text{strong convexity}}{\leqslant} \mathbb{E}[\ell_n](\bar{\theta}_{n-1}) - \mathbb{E}[\ell_n](\theta^*) \lesssim \frac{\varepsilon B\sqrt{\log(d/\delta)}}{\sqrt{n}}$$

where $\bar{\theta}_{n-1} = \frac{1}{n} \sum_{t=1}^{n} \widehat{\theta}_{t-1}$

We observe the slow rate on the risk of order $UB\sqrt{log(d)/n}$.

Proof:

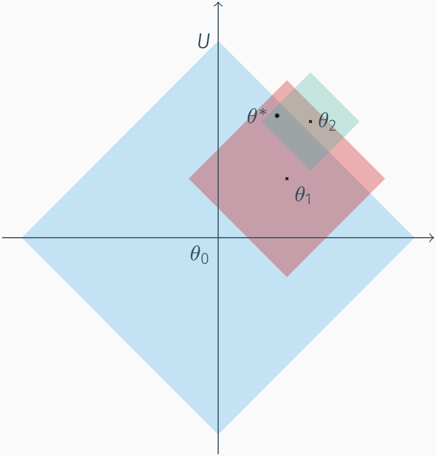
- Hoeffding inequality for martingal : with high probability

$$\sum_{t=1}^{n} \mathbb{E}[\ell_t](\widehat{\theta}_{t-1}) - \mathbb{E}[\ell_t](\theta^*) \lesssim \sum_{t=1}^{n} \ell_t(\widehat{\theta}_{t-1}) - \mathbb{E}[\ell_t](\theta^*) + \sqrt{n \log(1/\delta)}$$

- Jensen's inequality (convex i.i.d losses):

$$\mathbb{E}[\ell_n](\bar{\theta}_{n-1}) - \mathbb{E}[\ell_n](\theta^*) \leqslant \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[\ell_t](\widehat{\theta}_{t-1}) - \mathbb{E}[\ell_t](\theta^*)$$

ACCELERATION : FROM SLOW RATE TO FAST RATE



U
θ^*
θ_2
θ_1
θ_0

Algorithm

Parameters: $U, B, \alpha, \delta > 0$

Initialization: $\theta_{\text{center}} = 0 \in \mathbb{R}^d$, $t_1 = 1$

For sessions $i \geqslant 1$,

- Start a new $EG^{\pm}$ in $\mathcal{B}_1(\theta_{\text{center}}, U2^{-i})$ for $t \geqslant t_i$
- Get the high probability ℓ_1-ball for θ^*

$$\|\bar{\theta}_{t-1} - \theta^*\|_1^2 \leqslant d\|\bar{\theta}_{t-1} - \theta^*\|_2^2 \lesssim \frac{dU2^{-i}B\sqrt{\log(d/\delta)}}{\alpha\sqrt{t-t_i}} =: C(U,t)^2$$

- Define $t_{i+1} \geqslant t_i$ as the first time such that $C(U, t_{t_{i+1}}) \leqslant U2^{-(i+1)}$, i.e., for

$$\sqrt{t_{i+1} - t_i} \approx \frac{4dB\sqrt{\log(d/\delta)}}{U2^{-i}\alpha}$$

- $\theta_{\text{center}} \leftarrow \bar{\theta}_{t_{i+1}-1}$

After n time steps, we will have the slow rate high probability bound

$$\mathbb{E}[\ell_n](\bar{\theta}_{n-1}) - \mathbb{E}[\ell_n](\widehat{\theta}_{t-1}) \lesssim \frac{U2^{-i}B\sqrt{\log(d/\delta)}}{\sqrt{n}}$$

but with $U2^{-i} \approx \frac{dB\sqrt{\log(d/\delta)}}{\alpha\sqrt{n}}$.

Algorithm

Parameters: $U, B, \alpha, \delta > 0$

Initialization: $\theta_{\text{center}} = 0 \in \mathbb{R}^d$, $t_1 = 1$

For sessions $i \geqslant 1$,

- Start a new $EG^{\pm}$ in $\mathcal{B}_1(\theta_{\text{center}}, U2^{-i})$ for $t \geqslant t_i$
- Get the high probability ℓ_1-ball for θ^*

$$\|\bar{\theta}_{t-1} - \theta^*\|_1^2 \leqslant d\|\bar{\theta}_{t-1} - \theta^*\|_2^2 \lesssim \frac{dU2^{-i}B\sqrt{\log(d/\delta)}}{\alpha\sqrt{t - t_i}} =: C(U, t)^2$$

- Define $t_{i+1} \geqslant t_i$ as the first time such that $C(U, t_{t_{i+1}}) \leqslant U2^{-(i+1)}$, i.e., for

$$\sqrt{t_{i+1} - t_i} \approx \frac{4dB\sqrt{\log(d/\delta)}}{U2^{-i}\alpha}$$

- $\theta_{\text{center}} \leftarrow \bar{\theta}_{t_{i+1}-1}$

After n time steps, we will have the slow rate high probability bound

$$\mathbb{E}[\ell_n](\bar{\theta}_{n-1}) - \mathbb{E}[\ell_n](\widehat{\theta}_{t-1}) \lesssim \frac{U2^{-i}B\sqrt{\log(d/\delta)}}{\sqrt{n}} \lesssim \frac{dB^2\log(d/\delta)}{\alpha n}$$

but with $U2^{-i} \approx \frac{dB\sqrt{\log(d/\delta)}}{\alpha\sqrt{n}}$.

ACCELERATION: FROM SLOW RATE $1/\sqrt{n}$ TO FAST RATE $1/n$

Algorithm

Parameters: $U, B, \alpha, \delta > 0$ Initialization: $\theta_{\text{center}} = 0 \in \mathbb{R}^d$, $t_1 = 1$

For sessions $i \geqslant 1$,

- Start a new $EG^{\pm}$ in $\mathcal{B}_1(\theta_{\text{center}}, U2^{-i})$ for $t \geqslant t_i$
- Get the high probability ℓ_1-ball for θ^*

$$\|\bar{\theta}_{t-1} - \theta^*\|_1^2 \leqslant d\|\bar{\theta}_{t-1} - \theta^*\|_2^2 \lesssim \frac{dU2^{-i}B\sqrt{\log(d/\delta)}}{\alpha\sqrt{t-t_i}} =: C(U,t)^2$$

- Define $t_{i+1} \geqslant t_i$ as the first time such that $C(U, t_{t_{i+1}}) \leqslant U2^{-(i+1)}$, i.e., for

$$\sqrt{t_{i+1} - t_i} \approx \frac{4dB\sqrt{\log(d/\delta)}}{U2^{-i}\alpha}$$

- $\theta_{\text{center}} \leftarrow \bar{\theta}_{t_{i+1}-1}$

After n time steps, we will have the slow rate high probability bound

$$\mathbb{E}[\ell_n](\bar{\theta}_{n-1}) - \mathbb{E}[\ell_n](\widehat{\theta}_{t-1}) \lesssim \frac{U2^{-i}B\sqrt{\log(d/\delta)}}{\sqrt{n}} \lesssim \frac{dB^2\log(d/\delta)}{\alpha n}$$

but with $U2^{-i} \approx \frac{dB\sqrt{\log(d/\delta)}}{\alpha\sqrt{n}}$.

Algorithm (SAEW)

Parameters: $U, B, \alpha, \delta > 0$, $d_0 \geqslant 1$ Initialization: $\theta_{\text{center}} = 0 \in \mathbb{R}^d$, $t_1 = 1$

For sessions $i \geqslant 1$,

- Start a new $EG^{\pm}$ in $\mathcal{B}_1(\theta_{\text{center}}, U2^{-i})$ for $t \geqslant t_i$
- Get the high probability ℓ_1-ball for θ^*

$$\big\|[\bar{\theta}_{t-1}]_{d_0} - \theta^*\|_1^2 \leqslant d_0 \big\|[\bar{\theta}_{t-1}]_{d_0} - \theta^*\|_2^2 \lesssim \frac{d_0 U2^{-i} B \sqrt{\log(d/\delta)}}{\alpha\sqrt{t - t_i}} =: C(U, t)^2$$

  where $[\bar{\theta}_{t-1}]_{d_0}$ is the d_0-truncation of $\bar{\theta}_{t-1}$
- Define $t_{i+1} \geqslant t_i$ as the first time such that $C(U, t_{t_{i+1}}) \leqslant U2^{-(i+1)}$, i.e., for

$$\sqrt{t_{i+1} - t_i} \approx \frac{4 d_0 B \sqrt{\log(d/\delta)}}{U2^{-i}\alpha}$$

- $\theta_{\text{center}} \leftarrow \bar{\theta}_{t_{i+1}-1}$

After n time steps, we will have the slow rate high probability bound

$$\mathbb{E}[\ell_n](\bar{\theta}_{n-1}) - \mathbb{E}[\ell_n](\widehat{\theta}_{t-1}) \lesssim \frac{U2^{-i} B \sqrt{\log(d/\delta)}}{\sqrt{n}} \lesssim \frac{d_0 B^2 \log(d/\delta)}{\alpha n}$$

but with $U2^{-i} \approx \frac{d_0 B \sqrt{\log(d/\delta)}}{\alpha\sqrt{n}}$.

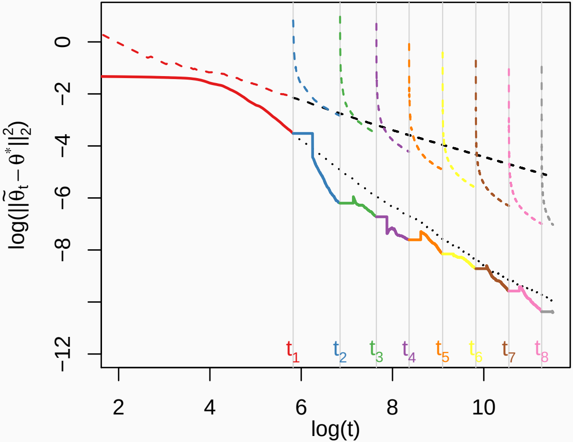


Figure 1: Logarithm of the ℓ_2-error of the averaged estimator.

Theorem

The average risk of SAEW is upper-bounded as

$$\mathrm{Risk}_{1:n}(\widehat{\theta}_{0:(n-1)}) \lesssim \min\left\{ UB\sqrt{\frac{\log(d/\delta)}{n}}, \frac{d_0 B^2}{\alpha n}\log(d/\delta)\log n + \frac{\alpha U^2}{d_0 n} \right\}.$$

Remarks:

- Both rates are optimal (in some sense)
- From the strong convexity assumption, this also ensures

$$\|\bar{\theta}_{n-1} - \theta^*\|_2^2 \lesssim \min\left\{ \frac{UB\sqrt{\log(d/\delta)}}{\alpha\sqrt{n}}, \frac{d_0 B^2}{n\alpha^2}\log(d/\delta)\log n + \frac{d_0 U^2}{\alpha n} \right\}$$

- the boundness of the gradients $B \geqslant \max_{\theta\in\mathcal{B}_1(0,2U)} \|\nabla \ell_t(\theta)\|_\infty$ can be weakened to unkown B under the subgaussian condition

Theorem

The average risk of SAEW is upper-bounded as

$$\mathrm{Risk}_{1:n}(\widehat{\theta}_{0:(n-1)}) \lesssim \min\left\{ UB\sqrt{\frac{\log(d/\delta)}{n}}, \frac{d_0 B^2}{\alpha n}\log(d/\delta)\log n + \frac{\alpha U^2}{d_0 n} \right\}.$$

Remarks:

- Both rates are optimal (in some sense)
- From the strong convexity assumption, this also ensures

$$\|\tilde{\theta}_{n-1} - \theta^*\|_2^2 \lesssim \min\left\{ \frac{UB\sqrt{\log(d/\delta)}}{\alpha\sqrt{n}}, \frac{d_0 B^2}{n\alpha^2}\log(d/\delta)\cancel{\log n} + \frac{d_0 U^2}{\alpha n^2} \right\}$$

  where $\tilde{\theta}_{n-1} = (t_i - t_{i-1})^{-1}\sum_{t=t_{i-1}}^{t_i-1}\widehat{\theta}_{t-1}$ for $t_i \leqslant n \leqslant t_{i+1}$
- the boundness of the gradients $B \geqslant \max_{\theta\in\mathcal{B}_1(0,2U)}\|\nabla\ell_t(\theta)\|_\infty$ can be weakened to unkown B under the subgaussian condition

We compare three online optimization procedures:

- RDA[6]: ℓ_1-regularized dual averaging method

$$\widehat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \underbrace{\frac{1}{t}\sum_{s=1}^{t} \nabla \ell_t(\widehat{\theta}_{s-1})^\top \theta}_{\text{Linearized loss}} + \underbrace{\lambda \|\theta\|_1}_{\ell_1 \text{regularization}} + \underbrace{\frac{\gamma}{\sqrt{t}}\|\theta\|_2^2}_{\text{force strong-convexity}} \right\}$$

Good performance for hand-written digits classification
Produces sparse estimators: but slow rate, or fast rate with
No sparse guarantees

- BOA[7]: exponential weights with second order regularization ($\approx EG^{\pm}$ with good tuning and high probability properties).
achieves fast rate for expert selection
no fast rate in the ℓ_2-ball

- SAEW: our acceleration of BOA

All methods are tuned in hindsight with the best parameters on a grid.

6 Xiao, "Dual averaging methods for regularized stochastic learning and online optimization", 2010
7 Wintenberger, "Optimal learning with Bernstein Online Aggregation", 2014

Let $(X_t, Y_t) \in [-X, X]^d \times [-Y, Y]$ be i.i.d. random pairs ($X, Y > 0$).

Goal: estimate linearly Y_t by approaching

$$\theta^* \in \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathbb{E}\left[(Y_t - X_t^\top \theta)^2\right]$$

The strong convexity assumption is achieved with $\alpha \leqslant \lambda_{\min}\left(\mathbb{E}\left[X_t X_t^\top\right]\right)$.

Experiment: $X_t \sim \mathcal{N}(0,1)$ for $d = 500$, $n = 2\,000$

$$Y_t = X_t^\top \theta^* + 0.1\,\varepsilon_t \qquad \text{with} \quad \varepsilon_t \sim \mathcal{N}(0,1) \quad \text{i.i.d.}$$

where $d_0 = \|\theta^*\|_0 = 5$, $U = \|\theta^*\|_1 = 1$ with non-zero coordinates i.i.d. $\propto \mathcal{N}(0,1)$, $\alpha = 1$.

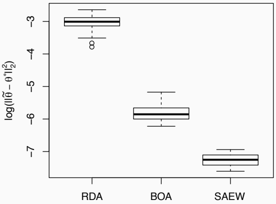


Figure 2: Boxplot (30 simulations)

least square regression with $d_0 = 5$, $d = 500$, $\sigma = 0.1$

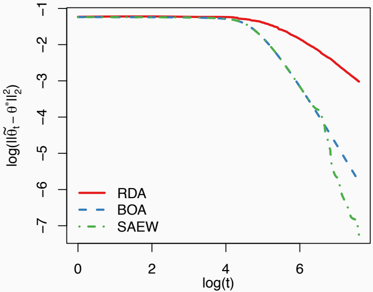


Figure 3: Log of the ℓ_2 error.

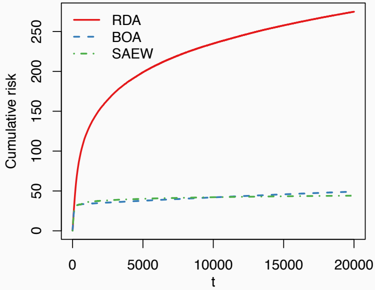


Figure 4: Cumulative risk.

Remarks: the cumulative risks are at most of order:

RDA: $\sigma^2 d \log n$ BOA: $\sigma^2 \sqrt{n \log d} + \log d$ SAEW: $\sigma^2 d_0 \log d \log n + \log d$

In practice: much better performance if we allow multiple pass on the training set.

Simulation: least square regression with $d_0 = 2$, $d = 2$, $\sigma = 0.3$

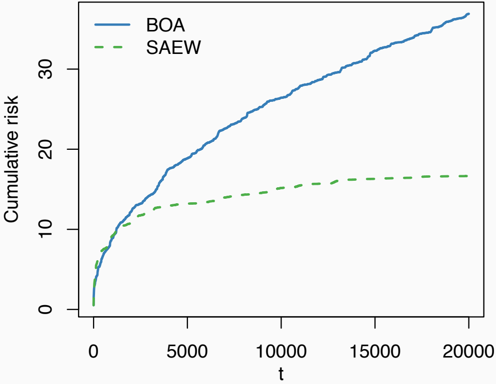


Figure 5: Cumulative risk for square linear regression with $d = d_0 = 2$.

Let $X, Y > 0$. Let $(X_t, Y_t) \in [-X, X]^d \times [-Y, Y]$ be i.i.d. random pairs.

Theorem

SAEW applied with $B = 2X(Y + 2XU)$ satisfies

$$\mathrm{Risk}_{1:n}\left(\widehat{\theta}_{0:(n-1)}\right) \lesssim \min\left\{ UX(Y + XU)\sqrt{\frac{\log(d/\delta)}{n}}, \right. \\ \left. \frac{d_0 X^2 (Y^2 + X^2 Y^2)}{\alpha n} \log\frac{d}{\delta} \log n + \frac{\alpha U^2}{d_0 n} \right\}.$$

A better tunning of $EG^{\pm}$ with $\eta_t \approx 1/\sqrt{\sum_{s=1}^{t} \|\nabla \ell_s(\widehat{\theta}_{s-1})\|_\infty^2}$ alows to substituted B^2 with $X^2\sigma^2$ with

$$\sigma^2 = \mathbb{E}\left[(Y_t - X_t^\top \theta^*)^2\right].$$

in the instantaneous risk of $\tilde{\theta}_{n-1}$.

Let $X, Y > 0$. Let $(X_t, Y_t) \in [-X, X]^d \times [-Y, Y]$ be i.i.d. random pairs.

Theorem

SAEW applied with $B = 2X(Y + 2XU)$ satisfies

$$\text{Risk}\left(\tilde{\theta}_{n-1}\right) \lesssim \min\left\{\sigma\sqrt{\frac{\log(d/\delta)}{n}}, \frac{d_0 X^2 \sigma^2}{n\alpha} \log\frac{d}{\delta} + \frac{\alpha U^2}{d_0 n^2}\right\}.$$

A better tunning of $EG^{\pm}$ with $\eta_t \approx 1/\sqrt{\sum_{s=1}^{t} \|\nabla \ell_s(\widehat{\theta}_{s-1})\|_\infty^2}$ alows to substituted B^2 with $X^2\sigma^2$ with

$$\sigma^2 = \mathbb{E}\left[(Y_t - X_t^\top \theta^*)^2\right].$$

in the instantaneous risk of $\tilde{\theta}_{n-1}$.

👍 Optimal rate for sparse least square regression.

The algorithm needs to know : $d_0, U, B, \alpha, \delta$

Run a meta-algorithm (BOA[8]) with parameters in a growing grid.

We leave the initial setting since we need

- to observe the gradients of all sub-algorithms.
- to clip the predictions $\widehat{\theta}_{t-1}^\top X_t \to [\widehat{\theta}_{t-1}^\top X_t]_{[-Y,Y]}$ otherwise we pay the maximal value of U considered in the final bound.
- we need strongly convex ℓ_t (instead of $\mathbb{E}[\ell_t]$ only)

This works to build an estimator for least square regression.

Theorem (Calibrated SAEW for Least Square Linear Regression)

The excess risk of the estimator produced by the meta-algorithm is of order

$$\mathcal{O}_n\Bigg(\underbrace{\frac{Y^2}{n}\log\left(\frac{(\log d)(\log n + \log Y)}{\delta}\right)}_{\text{Price of calibration}} + \underbrace{\frac{d_0 X^2\sigma^2}{\alpha^* n}\log\big(d/\delta\big)}_{\text{Fast rate with best } \alpha^*,\, d_0}\Bigg),$$

where α^* is the largest stong-convexity parameter.

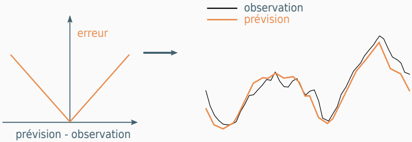
Can we substitute α^* with local strong convexity (cf. Lasso)?

8 Wintenberger, "Optimal learning with Bernstein Online Aggregation", 2014.

QUANTILE LINEAR REGRESSION

Let $\tau \in (0,1)$. Let $(X_t, Y_t) \in \mathbb{R}^d \times \mathbb{R}$ be i.i.d. random pairs.

Goal: estimate the conditional τ-quantile of Y_t given X_t.



Popular solution: linear regression with the pinball loss by $\rho_\tau : u \in \mathbb{R} \to u(\tau - \mathbb{1}_{u<0})$
The conditional quantile $q_\tau(Y_t|X_t)$ is the solution of

$$q_\tau(Y_t|X_t) \in \arg\min_g \mathbb{E}\left[\rho_\tau\left(Y_t - g(X_t)\right)\middle|X_t\right].$$

→ minimize the pinball loss q_τ = τ-quantile prediction

non-strongly convex loss

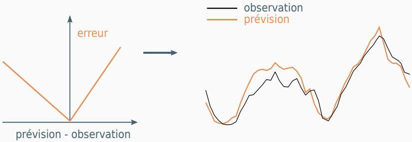
strongly convex risk[9] → ok for fixed parameter

[9] Steinwart and Christmann, "Estimating conditional quantiles with the help of the pinball loss", 2011.

QUANTILE LINEAR REGRESSION

Let $\tau \in (0,1)$. Let $(X_t, Y_t) \in \mathbb{R}^d \times \mathbb{R}$ be i.i.d. random pairs.

Goal: estimate the conditional τ-quantile of Y_t given X_t.



Popular solution: linear regression with the pinball loss by $\rho_\tau : u \in \mathbb{R} \to u(\tau - \mathbb{1}_{u<0})$
The conditional quantile $q_\tau(Y_t|X_t)$ is the solution of

$$q_\tau(Y_t|X_t) \in \arg\min_g \mathbb{E}\left[\rho_\tau\left(Y_t - g(X_t)\right)\middle|X_t\right] .$$

→ minimize the pinball loss q_τ = τ-quantile prediction

non-strongly convex loss

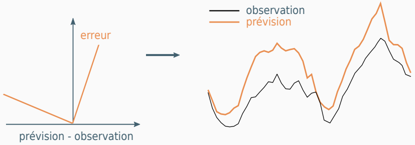
strongly convex risk[9] → ok for fixed parameter

[9] Steinwart and Christmann, "Estimating conditional quantiles with the help of the pinball loss", 2011.

QUANTILE LINEAR REGRESSION

Let $\tau \in (0,1)$. Let $(X_t, Y_t) \in \mathbb{R}^d \times \mathbb{R}$ be i.i.d. random pairs.

Goal: estimate the conditional τ-quantile of Y_t given X_t.



Popular solution: linear regression with the pinball loss by $\rho_\tau : u \in \mathbb{R} \to u(\tau - \mathbb{1}_{u<0})$
The conditional quantile $q_\tau(Y_t|X_t)$ is the solution of

$$q_\tau(Y_t|X_t) \in \arg\min_g \mathbb{E}\left[\rho_\tau\left(Y_t - g(X_t)\right)\middle|X_t\right] .$$

→ minimize the pinball loss q_τ = τ-quantile prediction

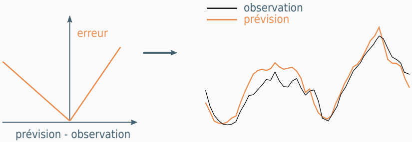
non-strongly convex loss　　strongly convex risk[9] → ok for fixed parameter

[9] Steinwart and Christmann, “Estimating conditional quantiles with the help of the pinball loss”, 2011.

QUANTILE LINEAR REGRESSION

Let $\tau \in (0,1)$. Let $(X_t, Y_t) \in \mathbb{R}^d \times \mathbb{R}$ be i.i.d. random pairs.

Goal: estimate the conditional τ-quantile of Y_t given X_t.



Popular solution: linear regression with the pinball loss by $\rho_\tau : u \in \mathbb{R} \to u(\tau - \mathbb{1}_{u<0})$
The conditional quantile $q_\tau(Y_t|X_t)$ is the solution of

$$q_\tau(Y_t|X_t) \in \arg\min_g \mathbb{E}\left[\rho_\tau\left(Y_t - g(X_t)\right)\middle|X_t\right] .$$

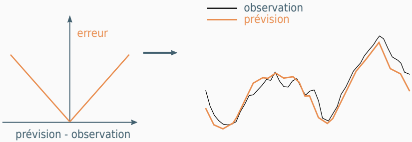
→ minimize the pinball loss q_τ = τ-quantile prediction

non-strongly convex loss strongly convex risk[9] → ok for fixed parameter

[9] Steinwart and Christmann, “Estimating conditional quantiles with the help of the pinball loss”, 2011.

Let $\tau \in (0,1)$. Let $(X_t, Y_t) \in \mathbb{R}^d \times \mathbb{R}$ be i.i.d. random pairs.

Goal: estimate the conditional τ-quantile of Y_t given X_t.



Popular solution: linear regression with the pinball loss by $\rho_\tau : u \in \mathbb{R} \to u(\tau - \mathbb{1}_{u<0})$
The conditional quantile $q_\tau(Y_t|X_t)$ is the solution of

$$q_\tau(Y_t|X_t) \in \arg\min_g \mathbb{E}\left[\rho_\tau\left(Y_t - g(X_t)\right)\middle|X_t\right] .$$

→ minimize the pinball loss q_τ = τ-quantile prediction

non-strongly convex loss

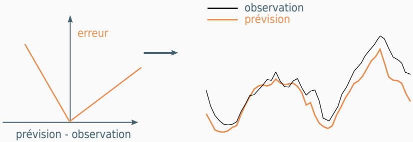
strongly convex risk[9] → ok for fixed parameter

9 Steinwart and Christmann, “Estimating conditional quantiles with the help of the pinball loss”, 2011.

QUANTILE LINEAR REGRESSION

Let $\tau \in (0,1)$. Let $(X_t, Y_t) \in \mathbb{R}^d \times \mathbb{R}$ be i.i.d. random pairs.

Goal: estimate the conditional τ-quantile of Y_t given X_t.



Popular solution: linear regression with the pinball loss by $\rho_\tau : u \in \mathbb{R} \to u(\tau - \mathbb{1}_{u<0})$
The conditional quantile $q_\tau(Y_t|X_t)$ is the solution of

$$q_\tau(Y_t|X_t) \in \arg\min_g \mathbb{E}\left[\rho_\tau\left(Y_t - g(X_t)\right)\middle|X_t\right] .$$

→ minimize the pinball loss q_τ = τ-quantile prediction

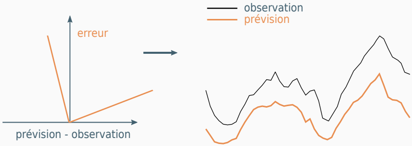
non-strongly convex loss strongly convex risk[9] → ok for fixed parameter

[9] Steinwart and Christmann, “Estimating conditional quantiles with the help of the pinball loss”, 2011.

QUANTILE LINEAR REGRESSION

Let $\tau \in (0,1)$. Let $(X_t, Y_t) \in \mathbb{R}^d \times \mathbb{R}$ be i.i.d. random pairs.

Goal: estimate the conditional τ-quantile of Y_t given X_t.



Popular solution: linear regression with the pinball loss by $\rho_\tau : u \in \mathbb{R} \to u(\tau - \mathbb{1}_{u<0})$
The conditional quantile $q_\tau(Y_t|X_t)$ is the solution of

$$q_\tau(Y_t|X_t) \in \arg\min_g \mathbb{E}\left[\rho_\tau\left(Y_t - g(X_t)\right)\middle|X_t\right] .$$

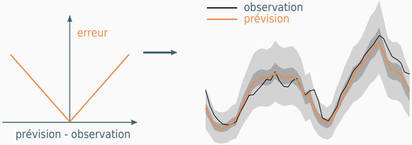
→ minimize the pinball loss q_τ = τ-quantile prediction

non-strongly convex loss   strongly convex risk[9] → ok for fixed parameter

[9] Steinwart and Christmann, "Estimating conditional quantiles with the help of the pinball loss", 2011.

Let $\tau \in (0,1)$. Let $(X_t, Y_t) \in \mathbb{R}^d \times \mathbb{R}$ be i.i.d. random pairs.

Goal: estimate the conditional τ-quantile of Y_t given X_t.



Popular solution: linear regression with the pinball loss by $\rho_\tau : u \in \mathbb{R} \to u(\tau - \mathbb{1}_{u<0})$
The conditional quantile $q_\tau(Y_t|X_t)$ is the solution of

$$q_\tau(Y_t|X_t) \in \arg\min_g \mathbb{E}\left[\rho_\tau\left(Y_t - g(X_t)\right)\middle|X_t\right] .$$

→ minimize the pinball loss q_τ = τ-quantile prediction

non-strongly convex loss strongly convex risk[9] → ok for fixed parameter

[9] Steinwart and Christmann, “Estimating conditional quantiles with the help of the pinball loss”, 2011.

Experiment: $d_0 = 5$, $d = 100$

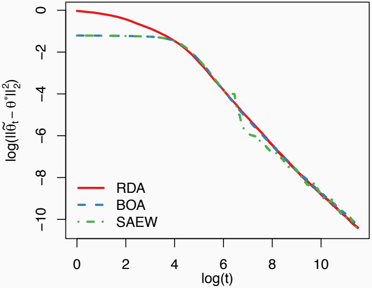


Figure 6: Log. of the ℓ_2-error of $\bar{\theta}_t$

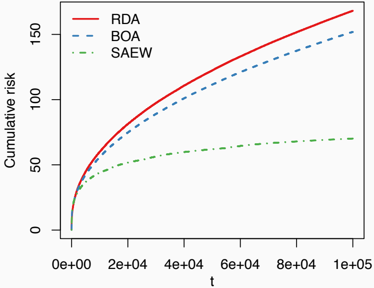


Figure 7: Cumulative risk

All the methods empirically get the fast rate $1/n$ for the ℓ_2-error of the estimator...
But only SAEW

- has the theoretical guarantee
- has $\mathcal{O}(\log n)$ cumulative risk

Some future work:

- Is averaging an efficient acceleration procedure for $EG^{\pm}$?
- Calibration of the parameters in the original online optimization setting
- Produce sparse estimators $\widehat{\theta}_{t-1}$
  → improve the dependency on the strong convexity parameter (only local)
- Oracle bound: no assumption on the sparsity of θ^*

THANK YOU !

REFERENCES

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

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

Gaillard, P. and O. Wintenberger. "Sparse Accelerated Exponential Weights". Accepted at AISTAT'17. 2017.

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

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

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

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

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

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