Efficient online algorithms for fast-rate regret bounds under sparsity

Pierre Gaillard (pierre.gaillard@inria.fr) Olivier Wintenberger (olivier.wintenberger@upmc.fr)

Problem: online optimization of convex functions

Setting Sequence of loss functions $\ell_1, \dots, \ell_n$ to be optimized sequentially. Loss functions are convex with bounded gradient $\|\nabla \ell_t(\theta)\|_\infty \leqslant G$ and can be random (i.i.d.) or adversarial.

For each iteration $t = 1, \dots, n$,
- Forecaster chooses $\widehat{\theta}_{t-1} \in \Theta \subset \mathbb{R}^d$
- Environment chooses and reveals $\ell_t : \Theta \to \mathbb{R}$.
- Forecaster incurs loss $\ell_t(\theta)$

The forecaster aims at minimizing the average excess risk

$$R_n(\theta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\big[\ell_t(\widehat{\theta}_{t-1})|\mathcal{F}_{t-1}\big] - \mathbb{E}\big[\ell_t(\theta)|\mathcal{F}_{t-1}\big]\,,$$

where $\mathcal{F}_{t-1} = \sigma(\{\ell_1, \dots, \ell_{t-1}\})$ denotes past information.

Goal Designing efficient algorithms which satisfy upper-bounds on $R_n(\theta)$ depending on the sparsity of θ rather than d with a faster rate than $O(1/\sqrt{n})$. Ideal bound would be

$$R_n(\theta) \lesssim \frac{\|\theta\|_0 \log d}{n}\,, \qquad \theta \in \Theta\,. \tag{$*$}$$

We will consider two cases $\Theta = \{\theta_1, \dots, \theta_K\}$ and $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leqslant 1\}$.

Contribution 1: finite $\Theta = \{\theta_1, \dots, \theta_K\}$, adversarial data

Finite Θ corresponds to online prediction with expert advice.

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

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

- General convexity: $\beta = 0$.
- Exp-concavity or strong-convexity: $\beta = 1$.

Classical worst-case bound (Hedge)

$$R_n(\theta) \lesssim \sqrt{\frac{\log(K)}{n}}$$

Bound under weak exp-concavity

$$R_n(\theta) \lesssim \left(\frac{\log(K)}{n}\right)^{\frac{1}{2-\beta}}$$

Quantile bound Prior $\widehat{\pi}_0$ on Θ. Smaller whenever many parameters θ_k perform well or when a good prior knowledge $\widehat{\pi}_0$ is available.

Algorithm BOA [11] or Squint [8]

$$\widehat{\pi}_t(k) \propto \sum_{\text{Grid of } \eta} \eta e^{\eta\sum_{s=1}^t r_s(k) - \eta^2 \sum_{s=1}^t r_s(k)^2}\pi_0(k) \quad \text{where } r_t(k) = \nabla \ell_t(\widehat{\theta}_{t-1})^\top(\widehat{\theta}_{t-1} - \theta_k)\,.$$

and predict $\widehat{\theta}_t = \sum_{k=1}^K \widehat{\pi}_t(k)\theta_k$.

Theorem BOA (and Squint) achieve with high-probability a quantile bound of the form: for all distribution π on Θ

$$\mathbb{E}_\pi\big[R_n(\theta_k)\big] \lesssim \left(\frac{\mathcal{K}(\pi, \widehat{\pi}_0)}{n}\right)^{\frac{1}{2-\beta}}$$

where $\lesssim$ denotes an approximate inequality and $\mathcal{K}(\pi, \widehat{\pi}_0) = \sum_{k=1}^K \pi(k)\log\frac{\pi(k)}{\widehat{\pi}_0(k)}$ is the Kullback-Leibler divergence.

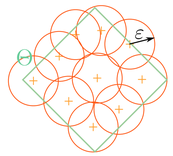
Warmup: obtaining fast-rate bound by discretizing the space

The aim is to extend the preceding result to $\Theta = \mathcal{B}_1$ instead of finite $\theta_1, \dots, \theta_K$ to obtain a sparse inequality similar to $(*)$.

Discretize $\mathcal{B}_1$ with a finite approximation and apply BOA.

Theorem Discretizing $\mathcal{B}_1$ with a finite $\frac{1}{n}$-covering in ℓ_1-norm and carefully choosing the prior $\widehat{\pi}_0$ to favor sparse parameters, BOA satisfies with high probability

$$R_n(\theta) \lesssim \left(\frac{\|\theta\|_0 \log\frac{dn}{\|\theta\|_0}}{\alpha n}\right)^{\frac{1}{2-\beta}}.$$



Large grid $\Rightarrow$ Prohibitive complexity $O(n^d)$.

We analyse the performance of the algorithm with arbitrary discretization grid. The improvement in performance is measured by a pseudo-metric called averaging accelerability.

Contribution 2: $\Theta = \mathcal{B}_1$, adversarial data

We provide an efficient algorithm (BOA+) with a sparse regret bound. The algorithm is based on the construction of an adaptive discretization grid of small size.

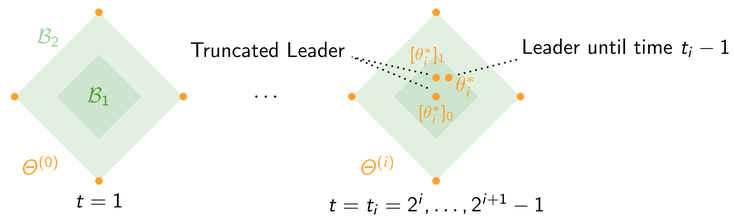
Assumption (strong-convexity)

$$\forall \theta_1, \theta_2 \in \mathcal{B}_2, \qquad \ell_t(\theta_1) - \ell_t(\theta_2) \leqslant \nabla \ell_t(\theta_1)^\top(\theta_1 - \theta_2) - \frac{\mu}{2}\|\theta_1 - \theta_2\|^2\,.$$

Algorithm (BOA+) Restart BOA on doubling sessions of length 2^i on adaptive discretization grid containing truncations of the leader

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

where $\theta_i^* \in \arg\min_{\theta\in\mathcal{B}_1}\sum_{t=1}^{t_i-1}\ell_t(\theta)$ is the leader until time $t_i - 1$.



Theorem BOA+ achieves w.h.p. for all $\theta \in \mathcal{B}_1$

$$R_n(\theta) \lesssim \frac{\sqrt{\|\theta\|_0 d}}{n} \wedge \frac{\|\theta\|_0}{n^{3/4}}\,.$$

Efficient: per-round time-complexity $\simeq d$.

Significant gain $\sqrt{\frac{\|\theta\|_0}{d}} \wedge \sqrt{\frac{\|\theta\|_0}{n}}$ for sparse θ with respect to classical guarantees.

Valid for all $\theta \to$ estimation-approximation trade-off. Seems suboptimal.

Comparison with existing procedures in sparse adversarial environment

Procedure	Rate	Polynomial	Assumption	Sparsity setting
Kale et al. [4, 7]	$\text{Poly}(d)/\sqrt{T}$	Yes	Conv.	Sparse observed gradients
[3, 9, 12]	$\sqrt{\frac{\log d}{T}}$ or $\frac{d}{T}$	Yes	(Strong) Conv.	Sparse estimators
SeqSEW [6]	$\frac{d_0 \log d}{T}$	No	Strong Conv.	Sparse bound
SABOA	$\sqrt{\frac{\log d}{T}} \wedge \frac{\sqrt{d_0 d}}{T}\log d$	Yes	Strong Conv.	Sparse bound

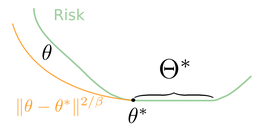
Contribution 3: $\theta = \mathcal{B}_1$, i.i.d. data

Stochastic setting $\ell_1, \dots, \ell_n$ are i.i.d. $\Theta^* := \arg\min_{\theta\in\mathcal{B}_1}\mathbb{E}[\ell_t(\theta)]$ set of minimizers.

Assumption (Łojasiewicz) There exist $\beta \in (0, 1)$ and $\mu > 0$ such that

$$\forall \theta \in \mathcal{B}_1, \quad \exists \theta^* \in \Theta^* \quad \text{s.t.} \qquad \mu\big\|\theta - \theta^*\big\|_2^2 \leqslant \mathbb{E}[\ell_t(\theta) - \ell_t(\theta^*)]^\beta$$

Properties of Łojasiewicz's Assumption:
- Convexity: $\beta = 0$.
- Strong convexity: $\beta = 1$.
- Allows multiple optima $\to$ important when dealing with degenerated collinear design (smallest non-zero eigenvalue).



Algorithm (SABOA) Restart BOA on doubling sessions of length 2^i on adaptive discretization grids containing hard-truncated and dilated soft-thresholded versions of the current estimator $\bar{\theta}_t = (1/t)\sum_{s\leqslant t}\widehat{\theta}_s$.

Theorem SABOA achieves under weak exp-concavity and w.h.p. for all $\theta^* \in \Theta^*$

$$R_n(\theta^*) \lesssim \left(\frac{G^2}{\gamma^2\mu}\frac{d_0 \log d}{n}\right)^{\frac{1}{2-\beta}} \qquad \text{if } \Theta^* \subset \mathcal{B}_{1-\gamma}\,,$$

where $d_0 = \max_{\theta^*\in\Theta^*}\|\theta^*\|_0$ and in any case

$$R_n(\theta^*) \lesssim \left(\frac{G^2}{\mu}\frac{d_0^2 \log d}{n}\right)^{\frac{1}{2-\beta}}.$$

Tuning of the parameters: SABOA adapts automatically to unknown parameters β, α, μ and d_0 and thus to the regularity of the risk.

Constrained set of minimizers: $\theta^* \in \Theta^*$ are likely to be sparse.

Efficient: per-round time complexity $\approx d$.

Additional factor d_0 when the minimizers θ^* lie on the border of the ball ($\|\theta^*\|_1 = 1$). Open question: can it be removed while keeping an efficient procedure?

On the radius of the ℓ_1-ball: analysis done for radius 1 but extension to any radius may be done by rescaling the loss function.

Comparison with sparse existing procedures in i.i.d. environment

Procedure	Setting	Rate	Assumptions / Setting	Optimum over
Lasso [2]	B	$d_0 \log d/T$	Mutual Coherence	$\mathbb{R}^d$
Kale et al. [4, 7]	S	$d_0^2 \log d/T$	Strong Conv. + Sparse Gradients	$\mathcal{B}_1$
[1, 5, 10]+SABOA	S	$d_0 \log d/T$	Strong conv. or Łojasiewicz ($\beta = 1$)	$\mathbb{R}^d$
SABOA	S	$d_0^2 \log d/T$	Łojasiewicz ($\beta = 1$)	$\mathcal{B}_1$

References

[1] A. Agarwal et al. "Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions". NIPS. (2012).
[2] F. Bunea et al. "Sparsity oracle inequalities for the Lasso". *EJS* (2007).
[3] J. C. Duchi et al. "Composite objective mirror descent". COLT. (2010).
[4] D. J. Foster et al. "Online sparse linear regression". COLT. (2016).
[5] P. Gaillard and O. Wintenberger. "Sparse accelerated exponential weights". AISTATS. (2017).
[6] S. Gerchinovitz. "Sparsity regret bounds for individual sequences in online linear regression". *JMLR* (2013).
[7] S. Kale et al. "Adaptive feature selection: computationally efficient online sparse linear regression under RIP". ICML. (2017).
[8] W. M. Koolen and T. Van Erven. "Second-order Quantile Methods for Experts and Combinatorial Games." COLT. (2015).
[9] J. Langford et al. "Sparse online learning via truncated gradient". *JMLR* (2009).
[10] J. Steinhardt et al. "The statistics of streaming sparse regression". *arXiv:1412.4182* (2014).
[11] O. Wintenberger. "Optimal learning with Bernstein online aggregation". *Machine Learning* (2017).
[12] L. Xiao. "Dual averaging methods for regularized stochastic learning and online optimization". *JMLR* (2010).