Sparsity regret bounds for individual sequences in online linear regression

Sébastien Gerchinovitz

DMA, École Normale Supérieure
Paris, France

Introduction: sparsity in the stochastic setting

Consider the linear regression model with fixed or random design: we observe independent pairs $(X_1, Y_1), \dots, (X_T, Y_T) \in \mathcal{X} \times \mathbb{R}$ such that

$$Y_t = \sum_{j=1}^{d} u_j^* \varphi_j(X_t) + \varepsilon_t \ , \quad \mathbb{E}[\varepsilon_t | X_t] = 0 \ , \quad 1 \leqslant t \leqslant T \ .$$

Prediction problem: estimate $\sum_{j=1}^{d} u_j^* \varphi_j$ knowing $(\varphi_j)_j$ but not $\boldsymbol{u}^* \in \mathbb{R}^d$.

- Small dimensions $d \leqslant T$: optimal rate $= \Theta\big(d/T\big)$.
- Higher dimensions $d > T$: no accurate prediction in general.

But if $\boldsymbol{u}^*$ is sparse, i.e., if $\|\boldsymbol{u}^*\|_0 \triangleq \big|\{j : u_j^* \neq 0\}\big|$ is small, then the faster rate $\|\boldsymbol{u}^*\|_0 \, (\ln d)/T$ is achievable.

Algos: ℓ^0-regularization [BM01, BTW07], ℓ^1-regularization [CT07, vdG08, BRT09], exponential weighting [DT08, AL11, RT11].

Outline

In this talk, we consider a deterministic online setting known as online linear regression on individual sequences.

We introduce the notion of sparsity regret bound, which is a deterministic bound expressed in terms of $\|\boldsymbol{u}^*\|_0$.

Outline of the talk:

1. Introduction of the notion of sparsity regret bound
2. Main results with individual sequences
3. Adaptativity results with i.i.d. data

Setting: online linear regression on individual sequences

Deterministic online framework studied, e.g., by [CBLW96, AW01, Vov01].

Prediction task: at each time t, predict the observation $y_t \in \mathbb{R}$ from the input data $x_t \in \mathcal{X}$ and some base forecasters $\varphi_j : \mathcal{X} \to \mathbb{R}$, $j = 1, \ldots, d$.

Initial step: the environment chooses arbitrary deterministic sequences $(y_t)_{t \geqslant 1}$ in $\mathbb{R}$ and $(x_t)_{t \geqslant 1}$ in $\mathcal{X}$ but the forecaster has not access to them.

At each time round $t \in \mathbb{N}^*$,

1. The environment reveals the input data $x_t \in \mathcal{X}$.
2. The forecaster chooses a prediction $\widehat{y}_t \in \mathbb{R}$.
3. The environment reveals the observation $y_t \in \mathbb{R}$, and the forecaster incurs the loss $(y_t - \widehat{y}_t)^2$.

Prediction goal

Goal: to predict almost as well as the best linear forecaster $\boldsymbol{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^{d} u_j \varphi_j$, $\boldsymbol{u} \in \mathbb{R}^d$, i.e., to satisfy a regret bound of the form

$$\sum_{t=1}^{T} \bigl(y_t - \widehat{y}_t\bigr)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^{\boldsymbol{d}}} \left\{ \sum_{t=1}^{T} \bigl(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\bigr)^2 + \Delta_{T,d}(\boldsymbol{u}) \right\} ,$$

where the regret term $\Delta_{T,d}(\boldsymbol{u})$ should be small (sublinear in T).

Example: the sequential Ridge regression forecaster of [AW01, Vov01] satisfies $\Delta_{T,d}(\boldsymbol{u}) \lesssim d \ln\bigl(T \|\boldsymbol{u}\|_2^2\bigr)$.

The bound $d \ln(T)$ is sublinear in T in small dimension $d \ll T / \ln(T)$.

What about the high dimensions?

In high dimension $d \gg T/\ln(T)$, the bound $d\ln(T)$ of the Ridge regression forecaster is no longer sublinear in T.

But, as in the stochastic setting, low regret is possible under a sparsity scenario: assume that some $\boldsymbol{u}^* \in \mathbb{R}^d$ has a low cumulative loss and that $\|\boldsymbol{u}^*\|_0 \ll T/\ln(T)$.

Then the sequential Ridge forecaster applied to the support of $\boldsymbol{u}^*$ would have a regret $\Delta_{T,d}(\boldsymbol{u}^*) \lesssim \|\boldsymbol{u}^*\|_0 \ln\big(T\,\|\boldsymbol{u}^*\|_2^2\big) \ll T$.

(This is a benchmark bound since $\mathrm{supp}(\boldsymbol{u}^*)$ is unknown.)

Sparsity regret bounds

In this talk, we prove that bounds proportional to $\|\boldsymbol{u}^*\|_0$ are achievable (up to logarithmic factors), i.e., we derive bounds of the form:

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^{\boldsymbol{d}}} \left\{ \sum_{t=1}^{T}\left(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\right)^2 + \left(\|\boldsymbol{u}\|_0 + 1\right) g_{T,d}\left(\|\boldsymbol{u}\|_1\right) \right\} ,$$

where g grows at most logarithmically in T, d, and $\|\boldsymbol{u}\|_1 \triangleq \sum_{j=1}^{d} |u_j|$.
We call these bounds sparsity regret bounds.

These bounds are deterministic online counterparts of the so-called sparsity oracle inequalities in the stochastic setting, which are of the form:

$$R\left(\widehat{\boldsymbol{u}}_T\right) \leqslant (1+a)\inf_{\boldsymbol{u}} \left\{ R(\boldsymbol{u}) + C(a)\,\frac{\|\boldsymbol{u}\|_0 \ln d}{T} \right\} .$$

Related works in online convex optimization

Recent papers [LLZ09, SST09, Xia10, DSSST10] addressed the sparsity issue in the online deterministic setting, but from a quite different angle:

- they focus on algorithms that output sparse linear combinations,
- while we are interested in algorithms with sparsity regret bounds:

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 + \big(\|\boldsymbol{u}\|_0 + 1\big)\, g_{T,d}\big(\|\boldsymbol{u}\|_1\big) \right\} .$$

Our sparsity regret bounds are optimal on ℓ^0-balls of small radii (up to logarithmic factors).

On the contrary, the regret bounds mentioned above have a worse dependence in $T, d, \|\boldsymbol{u}\|_1$, since they are of order $\sqrt{dT}$ or $\|\boldsymbol{u}\|_1 \sqrt{T \ln d}$.

1. Introduction of the notion of sparsity regret bound
   - Setting: online linear regression
   - Regret bounds in high dimension

2. Main results with individual sequences
   - Known bound B_y on the observations
   - Unknown bound B_y on the observations

3. Adaptivity results with i.i.d. data
   - Application to i.i.d. data
   - Adaptivity to the unknown variance

Algorithm: Sequential Sparse Exponential Weighting (SeqSEW)

Parameters: threshold B, inverse temperature η, and prior scale τ.

At each time round $t \geqslant 1$, the algorithm $\mathrm{SeqSEW}_{\tau}^{B,\eta}$ predicts as

$$\widehat{y}_t \triangleq \int_{\mathbb{R}^d} \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big]_B \, p_t(\mathrm{d}\boldsymbol{u}) \ ,$$

where $[z]_B = \max\big\{-B, \min\{B, z\}\big\}$ is the standard clipping, and where the probability distribution p_t over $\mathbb{R}^d$ is defined by

$$p_t(\mathrm{d}\boldsymbol{u}) \triangleq \frac{1}{W_t} \exp\left(-\eta \sum_{s=1}^{t-1} \Big(y_s - \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x_s)\big]_B\Big)^2\right) \pi_\tau(\mathrm{d}\boldsymbol{u})$$

for some normalizing constant W_t. The sparsity-favoring prior π_τ on $\mathbb{R}^d$ was introduced by [DT07] in the stochastic setting and is defined by

$$\pi_\tau(\mathrm{d}\boldsymbol{u}) \triangleq \prod_{j=1}^{d} \frac{(3/\tau)\,\mathrm{d}u_j}{2\big(1 + |u_j|/\tau\big)^4} \ .$$

Known bound B_y on the observations

We first assume that two a priori bounds B_y and B_Φ are available to the forecaster:

$$y_1, \dots, y_T \in [-B_y, B_y] \qquad \text{and} \qquad \sum_{j=1}^{d} \sum_{t=1}^{T} \varphi_j^2(x_t) \leqslant B_\Phi \ .$$

Theorem (G.)

Under the above assumptions, the algorithm $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$ *tuned with* $B = B_y$, $\eta = 1/(8B_y^2)$, *and* $\tau = \sqrt{16B_y^2/B_\Phi}$ *satisfies*

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 + 32 \left\|\boldsymbol{u}\right\|_0 B_y^2 \ln\left(1 + \frac{\sqrt{B_\Phi}\left\|\boldsymbol{u}\right\|_1}{4B_y \left\|\boldsymbol{u}\right\|_0}\right) \right\} + 16B_y^2$$

This upper bound is a sparsity regret bound as defined earlier.

Proof: use of a deterministic online PAC-Bayesian bound of [Aud09] combined with the form of the heavy-tailed prior π_τ studied in [DT07].

Unknown bound B_y on the observations

Algorithm SeqSEW^*_τ: we automatically adapt to $B_y = \max_{1 \leqslant t \leqslant T} |y_t|$ by replacing the threshold B and the inverse temperature η with

$$B_t \triangleq \left(2^{\lceil \log_2 \max_{1 \leqslant s \leqslant t-1} y_s^2 \rceil}\right)^{1/2} \approx \max_{1 \leqslant s \leqslant t-1} |y_s| \qquad \text{and} \qquad \eta_t \triangleq \frac{1}{8 B_t^2} \ .$$

Theorem (G.)

For a known bound $B_\Phi \geqslant \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t)$, but for an unknown $B_y = \max_{1 \leqslant t \leqslant T} |y_t|$, the algorithm SeqSEW^*_τ tuned with $\tau = 1/\sqrt{B_\Phi}$ satisfies

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 + 64 \, \|\boldsymbol{u}\|_0 \, B_y^2 \ln\left(1 + \frac{\sqrt{B_\Phi} \, \|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right) \right\} + 32 B_y^2 + 1 \ .$$

Proof via an adaptive PAC-Bayesian lemma.

Adaptation to B_Φ can be carried out via a standard doubling trick.

Application to i.i.d. data

Batch setting: we observe $(X_1, Y_1), \dots, (X_T, Y_T) \overset{\text{i.i.d.}}{\sim} (X, Y) \in \mathcal{X} \times \mathbb{R}$
The goal is to estimate the regression function $x \mapsto f(x) \triangleq \mathbb{E}[Y|X = x]$.

The sample $(X_1, Y_1), \dots, (X_T, Y_T)$ is treated in a sequential fashion. We run the algorithm SeqSEW^*_τ with $\tau = 1/\sqrt{dT}$ from time 1 to time T.
We estimate $f : \mathcal{X} \to \mathbb{R}$ with the Cesaro average $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ defined by

$$\widehat{f}_T(x) \triangleq \frac{1}{T} \sum_{t=1}^{T} \underbrace{\int_{\mathbb{R}^d} \left[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x)\right]_{B_t} p_t(\mathrm{d}\boldsymbol{u})}_{=\widehat{y}_t \text{ if } x = x_t} .$$

Though this online to batch conversion is standard, here $\widehat{f}_T$ does not depend on any prior knowledge such as

$$\mathbb{E}\Big[\big(Y - \mathbb{E}\big[Y\big|X\big]\big)^2\Big] , \qquad \|\varphi_j\|_\infty , \qquad \text{or} \qquad \|f - \varphi_j\|_\infty .$$

Adaptivity to the unknown variance

Theorem (A sparsity oracle inequality, G.)

$$\mathbb{E}\left\| f - \widehat{f}_T \right\|_{P_X}^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \|f - \boldsymbol{u}\cdot\boldsymbol{\varphi}\|_{P_X}^2 + 64\,\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right] \frac{\|\boldsymbol{u}\|_0}{T} \ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\right\}$$
$$+ \frac{1}{dT}\sum_{j=1}^{d} \|\varphi_j\|_{P_X}^2 + \frac{32}{T}\,\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right] .$$

$\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]$ can be upper bounded under various assumptions: e.g., if $\|f\|_\infty < +\infty$ and $\forall\lambda\in\mathbb{R},\ \mathbb{E}\left[\exp\left(\lambda\left[Y - \mathbb{E}[Y|X]\right]\right) \mid X\right] \leqslant e^{\lambda^2\sigma^2/2}$, then

$$\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right] \leqslant 2\left(\|f\|_\infty^2 + 2\sigma^2 \ln(2eT)\right)$$

This yields a risk bound similar to [DT10, Prop.1], but more adaptively:

- our bound is not restricted to ℓ^1-balls of finite radii;
- we do not require the knowledge of σ^2.

Adaptivity to the unknown variance

Theorem (A sparsity oracle inequality, G.)

$$\mathbb{E}\left\| f - \widehat{f}_T \right\|_{P_X}^2 \leqslant \inf_{u \in \mathbb{R}^d} \left\{ \|f - u \cdot \varphi\|_{P_X}^2 + 64\, \mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Y_t^2 \right] \frac{\|u\|_0}{T} \ln\left(1 + \frac{\sqrt{dT}\, \|u\|_1}{\|u\|_0} \right) \right\}$$
$$+ \frac{1}{dT} \sum_{j=1}^{d} \|\varphi_j\|_{P_X}^2 + \frac{32}{T}\, \mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Y_t^2 \right] .$$

$\mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Y_t^2\right]$ can be upper bounded under various assumptions: e.g., if $\|f\|_\infty < +\infty$ and $\forall \lambda \in \mathbb{R},\ \mathbb{E}\left[\exp\left(\lambda\left[Y - \mathbb{E}[Y|X]\right]\right) \mid X\right] \leqslant e^{\lambda^2 \sigma^2 / 2}$, then

$$\mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Y_t^2 \right] \leqslant 2\left(\|f\|_\infty^2 + 2\,\sigma^2 \ln(2\mathrm{e}\,T) \right) .$$

This yields a risk bound similar to [DT10, Prop.1], but more adaptively:

- our bound is not restricted to ℓ^1-balls of finite radii;
- we do not require the knowledge of σ^2.

Conclusion and ongoing work

Main contributions:

- introduction of the notion of sparsity regret bound;
- online adaptation to the unknown quantities B_y and B_Φ;
- application to i.i.d. data: risks bounds which are adaptive to the unknown variance of the noise.

Work in progress:

- derivation of sparsity regret bounds for algorithms which output sparse linear combinations (such as sequential Lasso).

Conclusion and ongoing work

Main contributions:

- introduction of the notion of sparsity regret bound;
- online adaptation to the unknown quantities B_y and B_Φ;
- application to i.i.d. data: risks bounds which are adaptive to the unknown variance of the noise.

Work in progress:

- derivation of sparsity regret bounds for algorithms which output sparse linear combinations (such as sequential Lasso).

These slides are available on my web page:
http://www.math.ens.fr/∼gerchinovitz

Thank you for your attention!

P. Alquier and K. Lounici.

PAC-Bayesian bounds for sparse regression estimation with exponential weights.

Electron. J. Stat., 5:127–145, 2011.

J.-Y. Audibert.

Fast learning rates in statistical inference through aggregation.

Ann. Statist., 37(4):1591–1646, 2009.

K. S. Azoury and M. K. Warmuth.

Relative loss bounds for on-line density estimation with the exponential family of distributions.

Mach. Learn., 43(3):211–246, 2001.

L. Birgé and P. Massart.

Gaussian model selection.

J. Eur. Math. Soc., 3:203–268, 2001.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov.

Simultaneous analysis of Lasso and Dantzig selector.

Ann. Statist., 37(4):1705–1732, 2009.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp.

Aggregation for Gaussian regression.

Ann. Statist., 35(4):1674–1697, 2007.

N. Cesa-Bianchi, P. M. Long, and M. K. Warmuth.

Worst-case quadratic loss bounds for prediction using linear functions and gradient descent.

IEEE Trans. Neural Networks, 7(3):604–619, 1996.

E. Candes and T. Tao.

The Dantzig selector: statistical estimation when p is much larger than n.

Ann. Statist., 35(6):2313–2351, 2007.

J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari.

Composite objective mirror descent.

In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT'10)*, pages 14–26, 2010.

A. Dalalyan and A. B. Tsybakov.

Aggregation by exponential weighting and sharp oracle inequalities.

In *Proceedings of the 20th Annual Conference on Learning Theory (COLT'07)*, pages 97–111, 2007.

A. Dalalyan and A. B. Tsybakov.

Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity.

72(1-2):39–61, 2008.

A. Dalalyan and A. B. Tsybakov.

Mirror averaging with sparsity priors.

Bernoulli, 2010.

To appear.

J. Langford, L. Li, and T. Zhang.

Sparse online learning via truncated gradient.

J. Mach. Learn. Res., 10:777–801, 2009.

P. Rigollet and A. B. Tsybakov.

Exponential Screening and optimal rates of sparse estimation.

Ann. Statist., 39(2):731–771, 2011.

S. Shalev-Shwartz and A. Tewari.

Stochastic methods for ℓ^1-regularized loss minimization.

In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, pages 929–936, 2009.

S. A. van de Geer.

High-dimensional generalized linear models and the Lasso.

Ann. Statist., 36(2):614–645, 2008.

V. Vovk.

Competitive on-line statistics.

Internat. Statist. Rev., 69:213–248, 2001.

L. Xiao.

Dual averaging methods for regularized stochastic learning and online optimization.

J. Mach. Learn. Res., 11:2543–2596, 2010.