Comment to "Generic chaining and the ℓ_1-penalty" by Sara van de Geer

Guillaume Lecué[1]

I would like to congratulate the author for the interesting ideas and intuition that are put forward in this contribution. I will focus my comment on the role of the generic chaining and on the geometry of ℓ_1-balls.

In [9], the oracle inequalities provide both a prediction result (a bound on the excess risk $\mathcal{E}(\hat{\theta}, \theta^0)$) and a coefficient estimation result (a bound on $\|\hat{\theta} - \theta^0\|_1$ when $\theta^0 \in \Theta$). There are two main steps in the author's proof. The first one follows from some tricky algebraic arguments and leads to Equation (3) of Theorem 2.1 (Equation (4) in Theorem 2.1 and Theorem 2.2 are similar in nature). Along the lines of this step, the role of the Margin assumption (Condition 2.1 in [9]): for all $\theta \in \Theta$,

$$\mathcal{E}(\theta; \theta^0) := P(\rho_\theta - \rho_{\theta^0}) \geq G(\tau(\theta - \theta^0)), \tag{1}$$

and the effective sparsity parameter for $S_0 = \{j : \theta_j^0 \neq 0\}$,

$$\Gamma^2(L, S_0) = \max\Big(\frac{\|\theta_{S_0}\|_1^2}{\tau(\theta)^2} : \big\|\theta_{S_0^c}\big\|_1 \leq L \|\theta_{S_0}\|_1\Big) \tag{2}$$

are highlighted. In particular, the norm τ characterizing the "local behavior" of the excess risk $\theta \longmapsto \mathcal{E}(\theta, \theta^0)$ around $\theta^0 \in \Theta$ in (1) appears to be the correct norm with respect to which the distortion with respect to the ℓ_1-norm has to be measured over the cone $\{\theta \in \mathbb{R}^p : \big\|\theta_{S_0^c}\big\|_1 \leq L \|\theta_{S_0}\|_1\}$ intersected with Θ (note that, in this cone, $\|\theta_{S_0}\|_1 \leq \|\theta\|_1 \leq (1 + L) \|\theta_{S_0}\|_1$). This first step does not require any probabilistic

[1]CNRS, LAMA, Université Paris-Est Marne-la-vallée, 77454 France. Supported by French Agence Nationale de la Recherche ANR Grant "Prognostic" ANR-09-JCJC-0101-01. Email: guillaume.lecue@univ-mlv.fr

tool and this bound (i.e. Equation (3) in Theorem 2.1) is obtained on the event

$$\mathcal{T}(\theta^0) = \Big\{ |(P - P_n)(\rho_\theta - \rho_{\theta^0})| \leq \lambda_0 \left\|\theta - \theta^0\right\|_1 \vee \lambda_0^2 : \forall \theta \in \Theta \Big\}. \tag{3}$$

The second step in the author's argument is to show that the event $\mathcal{T}(\theta^0)$ holds with high probability when λ_0 is of the order of $\sqrt{\log p/n}$. This is the step where empirical processes theory - and in particular, the tools developed in [8] - may be particularly useful. This is the place where Fernique-Slepian theorem can be use as a simple alternative to the generic chaining based argument in [9].

1 An alternative to the generic chaining argument in [9]

In [9], the first step to prove that the event $\mathcal{T}(\theta^0)$ holds with high probability is to use the peeling device. The second step is to study the empirical process $\theta \mapsto (P - P_n)(\rho_\theta - \rho_{\theta^0})$ indexed by $\Theta_M(\theta^0) = \{\theta \in \Theta : \|\theta - \theta^0\|_1 \leq M\}$ by means of symmetrization and concentration of suprema of Rademacher processes. At that point of the argument, everything boils down to upper bound the expectation (conditionally to $\mathbf{X} = (X_1, \ldots, X_n)$)

$$\mathbb{E}_\varepsilon \sup_{\theta \in \Theta_M(\theta^0)} \left| Y^\varepsilon(\theta, \theta^0) \right| \text{ where } Y^\varepsilon(\theta, \theta^0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (\rho_\theta^c - \rho_{\theta^0}^c)(X_i).$$

For this issue, we suggest an alternative proof to the generic chaining based argument in [9] (involving Talagrand majorizing measure theorem [6]).

This alternative proof is based upon two ingredients: a comparison theorem (cf. Lemma 4.5 in [2]),

$$\mathbb{E}_\varepsilon \sup_{\theta \in \Theta} \Big| \sum_{i=1}^n \varepsilon_i a_i(\theta) \Big| \lesssim \mathbb{E}_g \sup_{\theta \in \Theta} \Big| \sum_{i=1}^n g_i a_i(\theta) \Big|, \tag{4}$$

where the ε_i's are iid Rademacher and the g_i's are iid Gaussian (and the a_i's are any real-valued functions); and Fernique-Slepian theorem (cf. Theorem 3.15 in [2]): if $(Z(\theta))_{\theta \in \Theta}$ and $(X(\theta))_{\theta \in \Theta}$ are two Gaussian processes such that for all $\theta, \tilde{\theta} \in \Theta$

$$\mathbb{E}\big(Z(\theta) - Z(\tilde{\theta})\big)^2 \leq \mathbb{E}\big(X(\theta) - X(\tilde{\theta})\big)^2, \tag{5}$$

then

$$\mathbb{E}\sup_{\theta,\tilde{\theta}\in\Theta}|Z(\theta)-Z(\tilde{\theta})|\leq\mathbb{E}\sup_{\theta,\tilde{\theta}\in\Theta}|X(\theta)-X(\tilde{\theta})|. \tag{6}$$

In order to use the comparison theorem (4), we introduce the following Gaussian process: for any $\theta\in\Theta_M(\theta^0)$,

$$Z^g(\theta,\theta^0)=\frac{1}{\sqrt{n}}\sum_{i=1}^n g_i(\rho_\theta^c-\rho_{\theta^0}^c)(X_i).$$

It follows from the comparison principle in (4) that (conditionaly on $\mathbf{X}$),

$$\mathbb{E}_\varepsilon\sup_{\theta\in\Theta_M(\theta^0)}\big|Y^\varepsilon(\theta,\theta^0)\big|\lesssim\mathbb{E}_g\sup_{\theta\in\Theta_M(\theta^0)}\big|Z^g(\theta,\theta^0)\big|. \tag{7}$$

Assume that the pseudo-metric associated to the Gaussian process $(Z^g(\theta,\theta^0))_{\theta\in\Theta_M(\theta^0)}$ is such that for any $\theta,\tilde{\theta}\in\Theta_M(\theta^0)$

$$\mathbb{E}_g\big(Z^g(\theta,\theta^0)-(Z^g(\tilde{\theta},\theta^0)\big)^2\lesssim d(\theta,\tilde{\theta})^2 \tag{8}$$

where d is the natural pseudo-metric associated to some other Gaussian process $(X(\theta,\theta^0))_{\theta\in\Theta_M(\theta^0)}$. Then by Fernique-Slepian theorem, we have

$$\mathbb{E}_g\sup_{\theta,\tilde{\theta}\in\Theta_M(\theta^0)}\big|Z^g(\theta,\theta^0)-Z^g(\tilde{\theta},\theta^0)\big|\lesssim\mathbb{E}_g\sup_{\theta,\tilde{\theta}\in\Theta_M(\theta^0)}\big|X(\theta,\theta^0)-X(\tilde{\theta},\theta^0)\big|. \tag{9}$$

For instance, if we assume that (8) holds for the pseudo-metric (we use the notation of [9])

$$(\theta,\tilde{\theta})\longmapsto d(\theta,\tilde{\theta})^2=\sum_{k=1}^r\Big\|\sum_{j=1}^{p_k}(\theta_{j,k}-\tilde{\theta}_{j,k})\psi_{j,k}\Big\|_n^2, \tag{10}$$

then (9) holds for the following Gaussian process introduced in [9],

$$X(\theta,\theta^0)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\sum_{k=1}^r\sum_{j=1}^{p_k}(\theta_{j,k}-\theta_{j,k}^0)\psi_{j,k}(X_i,i)g_{i,k}$$

where the $g_{i,k}$'s are iid standard Gaussian variables. Condition (8) for the pseudo-metric (10) is equivalent to Condition 4.1 in [9]. In particular, Condition 4.1 in [9] can be seen as a comparison assumption between the canonical pseudo-metrics associated to two Gaussian processes (the process Z^g coming naturally from the

study and the process X for which the supremum is easier to handle thanks to some linearity properties).

Therefore, under Condition 4.1, one has (conditionaly on $\mathbf{X}$),

$$\mathbb{E}_\varepsilon \sup_{\theta\in\Theta_M(\theta^0)} \left|Y^\varepsilon(\theta,\theta^0)\right| \lesssim \mathbb{E}_g \sup_{\theta,\tilde\theta\in\Theta_M(\theta^0)} \left|X(\theta,\theta^0)-X(\tilde\theta,\theta^0)\right|.$$

Then, we recover Theorem 4.2 in [9] thanks to a duality argument and a Gaussian maximal inequality:

$$\mathbb{E}_g \sup_{\theta,\tilde\theta\in\Theta_M(\theta^0)} \left|X(\theta,\theta^0)-X(\tilde\theta,\theta^0)\right| \lesssim \sqrt{\log p}\max_{j,k}\|\psi_{j,k}\|_n\,.$$

2 Some comments on the geometry of ℓ_1-balls

It follows from the analysis of [9] that the regularization parameter λ is of the order of $\sqrt{(\log p)/n}$. The complexity term $\sqrt{\log p}$ comes from the Gaussian mean width $\ell^*(B_1^p)\sim\sqrt{\log p}$ where for any set $T\subset\mathbb{R}^p$ and for iid Standard Gaussian variables $g_1,\dots,g_p$,

$$\ell^*(T)=\mathbb{E}\sup_{t\in T}\sum_{j=1}^p g_i t_i.$$

This improves upon the result of [5], where the regularization parameter is taken of the order of $\sqrt{(\log n)^3(\log(p\vee n))/n}$. This last result follows from some entropy bound (cf. proof of Lemma 2 in [5]) and somehow the regularization parameter cannot be taken smaller than $Dudley(B_1^p,\ell_2^p)/\sqrt{n}$ where we denote by $Dudley(B_1^p,\ell_2^p)$ the Dudley entropy integral of B_1^p with respect to the ℓ_2^p-metric defined for any set $T\subset\mathbb{R}^p$ by

$$Dudley(T,\ell_2^p)=\int_0^\infty \sqrt{\log N(T,\varepsilon B_2^p)}d\varepsilon.$$

Thanks to [4], we have $Dudley(B_1^p,\ell_2^p)\sim(\log p)^{3/2}$. Note that in [5], the authors were able to "replace" some $\log p$ factors in $Dudley(B_1^p,\ell_2^p)$ by some $\log n$ factors in λ because, in fact, this is the expected complexity of a random n-dimensional section of B_1^p that measures the complexity of the problem.

The interesting point is that there is a gap between the two different complexity measures of the unit ℓ_1-ball B_1^p:

$$\ell^*(B_1^p)\sim\sqrt{\log p}\quad\text{ and }\quad Dudley(B_1^p,\ell_2^p)\sim(\log p)^{3/2}. \tag{11}$$

In particular, as mentioned in [5], trying to obtain an optimal value for λ through a Dudley's entropy integral will result inevitably in a logarithmic loss.

The gap between the Gaussian mean width and the Dudley entropy integral observed on the set B_1^p in (11) is somehow extremal in $\mathbb{R}^p$ since for any set $T \subset \mathbb{R}^p$,

$$\ell^*(T) \lesssim Dudley(B_1^p, \ell_2^p) \lesssim (\log p)\ell^*(T).$$

Indeed, the left-hand side is the classical chaining argument. For the right-hand side, if $\varepsilon_0 := \max_{\varepsilon>0} \big(N(T, \varepsilon B_2^p) \geq p\big)$, then by a volumetric argument (cf. Lemma 4.16 in [3]) and Sudakov inequality (cf. Theorem 3.18 in [2]),

$$\begin{aligned}
&\int_0^{\varepsilon_0} \sqrt{\log N(T, \varepsilon B_2^p)} d\varepsilon \lesssim \int_0^{\varepsilon_0} \sqrt{\log N(T, \varepsilon_0 B_2^p) + \log N(\varepsilon_0 B_2^p, \varepsilon B_2^p)} d\varepsilon \\
&\lesssim \varepsilon_0 \sqrt{\log N(T, \varepsilon_0 B_2^p)} + \int_0^{\varepsilon_0} \sqrt{p \log\big(5\varepsilon_0/\varepsilon\big)} d\varepsilon \lesssim \varepsilon_0 \sqrt{\log N(T, \varepsilon_0 B_2^p)} \lesssim \ell^*(T).
\end{aligned}$$

Then, if for any $s \in \mathbb{N}$, T_s denotes a maximal ε_s-net of T with respect to ℓ_2^p where $\epsilon_s := \inf_{\varepsilon>0} \big(N(T, \varepsilon B_2^p) \leq 2^{2^s}\big)$, it follows from Sudakov inequality that

$$\begin{aligned}
&\int_{\varepsilon_0}^{\infty} \sqrt{\log N(T, \varepsilon B_2^p)} d\varepsilon \lesssim \sum_{s=0}^{\lceil \log p \rceil} 2^{s/2} \sup_{t \in T} d_{\ell_2^p}(t, T_s) \\
&\lesssim (\log p) \max_{\varepsilon>0} \varepsilon \sqrt{\log N(T, \varepsilon B_2^p)} \lesssim (\log p)\ell^*(T).
\end{aligned}$$

It is interesting to note that, in the case of ellipsoids in $\mathbb{R}^p$, the gap between the Dudley entropy integral (w.r.t. ℓ_2^p) and the Gaussian mean width is at most $\sqrt{\log p}$ (cf. Chapter 2 in [7]). In the case of B_q^p-balls, it can be seen, thanks to the entropy estimates of [4], that there is no gap between the Dudley entropy integral and the Gaussian mean width: for any $1 < q \leq 2$,

$$Dudley(B_q^p, \ell_2^p) \sim \ell^*(B_q^p) \sim p^{1-1/q}.$$

Other examples of such sets having no gap between the Dudley entropy integral and the Gaussian mean width is the purpose of a result of Fernique that can be found, for instance, in Theorem 2.7.4 and Corollary 2.7.5 in [1].

As a conclusion, the study of the empirical processes naturally associated to the study of ℓ_1-based algorithms should avoid any "Dudley entropy integral based

approach" in order to avoid logarithmic losses. This is to me, one of the most important message behind [9].

Acknowledgement: I would like to thank Ramon van Handel and Shahar Mendelson for the different discussions we had on the generic chaining method which helped to improve this discussion paper.

References

[1] R. M. Dudley. *Uniform central limit theorems*, volume 63 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999.

[2] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[3] Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.

[4] Carsten Schütt. Entropy numbers of diagonal operators between symmetric Banach spaces. *J. Approx. Theory*, 40(2):121–128, 1984.

[5] Nicolas Städler, Peter Bühlmann, and Sara van de Geer. ℓ_1-penalization for mixture regression models. *TEST*, 19(2):209–256, 2010.

[6] Michel Talagrand. Regularity of Gaussian processes. *Acta Math.*, 159(1-2):99–149, 1987.

[7] Michel Talagrand. *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005. Upper and lower bounds of stochastic processes.

[8] Sara A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

[9] Sara A. van de Geer. Generic chaining and the ℓ_1-penalty. Technical report, ETH Zurich, 2012.