Sankhyā : The Indian Journal of Statistics
2017, Volume 79-A, Part 2, pp. 208-211

Discussion of the Paper on Concentration for (Regularized) Empirical Risk Minimization by Sara van de Geer and Martin Wainwright

Sourav Chatterjee
Stanford University, Stanford, USA

This paper is an important step forward in understanding empirical risk minimization in high dimensional problems. Broadly speaking, empirical risk minimization is a problem of the following type: Suppose that $X_1, \ldots, X_n$ are independent observations taking values in some space $\mathcal{X}$. Let $\mathcal{F}$ be a collection of real-valued functions on $\mathcal{X}$, and let $\text{pen}(f)$ be a penalty attached to each $f \in \mathcal{F}$. The empirical risk associated with the function f is

$$P_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i),$$

and the empirical risk minimization problem with the given penalty is the problem of minimizing

$$P_n f + \text{pen}(f)$$

over $f \in \mathcal{F}$. The minimizer, when unique, is usually denoted by $\hat{f}$.

As an example, consider linear regression with ℓ^1 penalty. Here the observations are i.i.d. vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$, where each X_i is a p-dimensional random vector of covariates and Y_i is a real-valued response. For each $\beta \in \mathbb{R}^p$, let

$$f_\beta(y) := (y - \beta^T x)^2,$$

where β^T is the transpose of β. Then the problem of linear regression with ℓ^1 penalty can be put into the above general framework by taking $\mathcal{F} = \{f_\beta\}_{\beta \in \mathbb{R}^p}$ and $\text{pen}(f_\beta) = \lambda|\beta|_1$, where $|\beta|_1$ is the ℓ^1 norm of β and λ is the penalty parameter.

Let us now return to the general framework. For $f \in \mathcal{F}$, let $Pf := \mathbb{E}(P_n f)$. Let f_0 be a minimizer of Pf. The risk of $\hat{f}$ is defined as

$$R(\hat{f}) := P\hat{f} - Pf_0 + \text{pen}(\hat{f}),$$

and its square-root is denoted by

$$\tau(\hat{f}) := \sqrt{R(\hat{f})}.$$

For example, in the linear regression problem described above, let β_0 be a minimizer of $\mathbb{E}(Y - \beta^T X)^2$, where (X, Y) is a random vector with the same distribution as $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let $\hat{\beta}$ be the minimizer of the least squares problem with ℓ^1 penalty, with penalty parameter λ. The risk of $\hat{\beta}$, according to the above definition, is

$$g(\hat{\beta}) - g(\beta_0) + \lambda|\hat{\beta}|_1,$$

where

$$g(\beta) := \mathbb{E}(Y - \beta^T X)^2.$$

Note that $g(\beta)$ is the prediction error when the parameter value is taken to be β.

The authors show that in a large class of problems, the risk (or rather, the square-root of the risk) of an empirical risk minimizer concentrates in a small neighborhood of its expected value. Specifically, they show that under suitable conditions, there is some deterministic real number s_0, depending on the problem, such that

$$|\tau(\hat{f}) - s_0| = o_P(s_0)$$

with high probability, where $o_P(s_0)$ denotes a random quantity whose magnitude is very small compared to s_0.

The value of such concentration inequalities lies in the fact that the standard methods for analyzing empirical risk minimizers usually give only upper bounds. A concentration inequality gives an upper bound as well as a lower bound. In my 2014 paper (Chatterjee, 2014), a concentration inequality of this flavor was proved for least squares under convex constraint in a Gaussian setting. The authors have taken this program far ahead of what was contained in my paper; in particular, they have been able to extend the theory to cover empirical processes, density estimation, log-linear models, etc.

The authors have also discovered a more "direct" approach to proving the relevant concentration inequalities, which, in my opinion, is more satisfactory than the approach I took. This is Theorem 2.1 in the paper, which proves the following remarkable claim. Suppose that Y is an n-dimensional Gaussian random vector with independent components, each with variance one but unknown mean. Let $\text{pen} : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex penalty function. Let $\hat{g}$ be the minimizer of the penalized least squares risk

$$\|Y - g\|_n^2 + \text{pen}(g)$$

as g ranges over $\mathbb{R}^n$, where

$$\|x\|_n^2 := \frac{1}{n}\sum x_i^2$$

for a vector $x = (x_1, \ldots, x_n)$. Let $g^0 := \mathbb{E}(Y)$ and $m_0 := \mathbb{E}\|\hat{g} - g^0\|_n$. Then Theorem 2.1 in the paper says that for any $t > 0$,

$$\mathbb{P}\big(\big|\|\hat{g} - g^0\|_n - m_0\big| \geq \sqrt{2t/n}\big) \leq e^{-t}.$$

The problem of computing m_0 is discussed in a later part of the paper. For the problem I considered in my paper, m_0 has a formula in terms of maxima of Gaussian processes. The authors give similar formulas in more general settings, but the formulas are not always easy to evaluate.

The paper is so well-written and comprehensive that it's hard to think of anything to add. There is one example which the authors may have overlooked (or may be intentionally omitted for the sake of brevity?) — it is the problem of matrix estimation. In recent years matrix estimation problems under a variety of sparsity assumptions have become popular. For example, Davenport et al. (2014) have proposed matrix estimation under nuclear norm penalty. This fits perfectly into the framework of Theorem 2.1 of this paper. It would be interesting to see if m_0 in such problems can be explicitly evaluated, at least up to leading order, using the myriad results available from random matrix theory.

Another possible extension is to non-convex penalties. I have the following heuristic in mind. Consider the same setting as above. For each $t \in \mathbb{R}$, let

$$S_t := \{g \in \mathbb{R}^n : \|g - g^0\|_n^2 + \mathrm{pen}(g) = t\}.$$

For each $g \in \mathbb{R}$, let

$$X_g := \frac{2}{n}(Y - g^0)^T(g^0 - g).$$

Note that $(X_g)_{g\in\mathbb{R}^n}$ is a centered Gaussian field. For each t, let

$$M_t := \min\{X_g : g \in S_t\}.$$

Being the minimum of a Gaussian field, M_t has some degree of concentration around its mean m_t. On the other hand, note that $\hat{t} := \|\hat{g} - g^0\|_n^2 + \mathrm{pen}(\hat{g})$ minimizes $M_t + t$ as t ranges over $\mathbb{R}$, since

$$\|Y - g\|_n^2 + \mathrm{pen}(g) = \|Y - g^0\|_n^2 + X_g + \|g^0 - g\|_n^2 + \mathrm{pen}(g).$$

Thus, under appropriate conditions, $\hat{t}$ should be close to the deterministic minimizer of $m_t + t$. Note that in the above heuristic, the convexity of pen is not an important factor.

If the above heuristic can be made into a theory, it may be possible to apply it to more discrete types of high-dimensional problems, such as those arising in models of networks. In such models the parameter space, instead of being $\mathbb{R}^n$, is the space of graphs on n vertices.

References

chatterjee, s. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.*, **42** no. 6, 2340–2381.

davenport, m. a., plan, y., van den berg, e. and wootters, m. (2014). 1-bit matrix completion. *Inf. Inference*, **3** no. 3, 189–223.

Sourav Chatterjee
Department of Statistics,
Stanford University,
Stanford, USA
E-mail: souravc@stanford.edu

Paper received: 20 May 2017.