M-estimation and complexity regularization

Sara van de Geer
Seminar für Statistik, ETH Zürich

1 Empirical processes

Consider a sample $Z_1, \ldots, Z_n$ of independent random variables, in some space $\mathcal{Z}$, and let $\gamma : \mathcal{Z} \to \mathbf{R}$ be a (measurable) function. We write the empirical average as

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

and the theoretical mean as

$$P\gamma := \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\gamma(Z_i).$$

Let Γ be a collection of functions on $\mathcal{Z}$. Empirical process theory is about the study of quantities of the type

$$\mathbf{Z} := \sup_{\gamma\in\Gamma}|(P_n - P)\gamma|,$$

in particular the study of probability and moment inequalities for $\mathbf{Z}$. Of further interest is the empirical process

$$\nu_n := \{\nu_n(\gamma) := \sqrt{n}(P_n - P)\gamma : \ \gamma \in \Gamma\}.$$

Here, asymptotic continuity (tightness) is a key concept. This is the following property:

$$\sup_{\sigma(\gamma-\gamma_0)\le\epsilon_n}|\nu_n(\gamma - \gamma_0)| \to^{\mathbf{P}} 0,$$

as $n \to \infty$, with $\{\epsilon_n\}$ a sequence of positive numbers decreasing to zero. Moreover,

$$\sigma^2(\gamma) := \frac{1}{n}\sum_{i=1}^{n}\mathrm{var}(\gamma(Z_i)).$$

In fact, we will examine the *increments* or *modulus of continuity* of the empirical process, which is the behavior of, for instance, the moments

$$\psi(\epsilon) := \mathbf{E}\sup_{\sigma(\gamma-\gamma_0)\le\epsilon}|\nu_n(\gamma - \gamma_0)|,$$

as function of ϵ.

2 Application to M-estimation

Suppose $\Gamma \subset \mathbf{\Gamma}_0$ is a given collection of loss functions. The M-estimator is

$$\hat{\gamma} := \arg\min_{\gamma \in \Gamma} P_n \gamma.$$

It is to be understood as an estimator of the target

$$\gamma_0 := \arg\min_{\gamma \in \mathbf{\Gamma}_0} P\gamma.$$

Note since $\mathbf{\Gamma}_0 \supset \Gamma$, the target γ_0 may not be an element of the class Γ over which we perform empirical risk minimization. The best approximation within Γ of the target is defined as

$$\gamma^* := \arg\min_{\gamma \in \Gamma} P\gamma.$$

The excess risk is defined as

$$\mathcal{E}(\gamma) := P(\gamma - \gamma_0), \ \gamma \in \Gamma.$$

We moreover call $\mathcal{E}^* := \mathcal{E}(\gamma^*)$ the approximation error. The behavior of the excess risk $\hat{\mathcal{E}} := \mathcal{E}(\hat{\gamma})$ of the estimator $\hat{\gamma}$ will be our topic of interest. The following simple inequality is our starting point.

Lemma 2.1 *It holds that*

$$\hat{\mathcal{E}} \leq -\nu_n(\hat{\gamma} - \gamma^*)/\sqrt{n} + \mathcal{E}^*.$$

Thus, the excess risk $\hat{\mathcal{E}}$ is bounded by two terms. The second term is the approximation error, and the first term can be thought of as the estimation error. This first term can be handled using empirical process theory.

For example, suppose we can show that

$$\mathbf{V}_n := \sup_{\gamma \in \Gamma} \frac{|\nu_n(\gamma - \gamma^*)}{\psi(\sigma_\gamma \vee \sigma^*)}$$

is a tight sequence of random variables (for example that moments exist and do not explode as $n \to \infty$). Here, we define

$$\sigma_\gamma := \sigma(\gamma - \gamma_0),$$

and $\sigma^* := \sigma_{\gamma^*}$. Moreover, ψ is some (concave) strictly increasing function.

Lemma 2.2 *Suppose that the margin condition*

$$\mathcal{E}(\gamma) \geq G(\sigma_\gamma), \forall \ \gamma \in \Gamma$$

holds. Here, G is some (convex) increasing function. Assume that $G_\psi := G \circ \psi^{-1}$ is strictly convex. Let H be the convex conjugate of G_ψ. Then for all $0 < \delta < 1$,

$$(1-\delta)\hat{\mathcal{E}} \leq \delta H\left(\frac{\mathbf{V}_n}{\delta\sqrt{n}}\right) + (1+\delta)\mathcal{E}^*.$$

Thus, Lemma 2.2 gives a bound for the estimation error in terms of the modulus of continuity ψ of the empirical process.

3 Modulus of continuity and entropy

Definition *Let (Λ, d) be a subset of a metric space. The* δ-covering number *$N(\delta, \Lambda, d)$ of Λ is the smallest value of N such that there exist $\{\lambda_j\}_{j=1}^N$ with*

$$\min_{1 \leq j \leq N} d(\lambda, \lambda_j) \leq \delta, \ \forall \ \lambda \in \Lambda.$$

The entropy *$\mathcal{H}(\cdot, \Lambda, d)$ is then defined as*

$$\mathcal{H}(\cdots, \Lambda, d) := \log(1 + N(\cdot, \Lambda, d)).$$

For a probability measure Q on $\mathcal{Z}$, let $\|\cdot\|_Q$ denote the $L_2(Q)$-norm. Suppose the entropy condition

$$\sup_{\text{probability measures } Q} \mathcal{H}(\cdot, \Lambda, \|\cdot\|_Q) \leq \mathcal{H}(\cdot),$$

where $\mathcal{H}$ is a continuous function for which the integral

$$\psi(\cdot) := 24 \int_0^{\cdot} \sqrt{\mathcal{H}(u)} du,$$

exists.

We will prove the following theorem.

Theorem 3.1 *Assume that the functions γ in Γ are bounded in sup-norm by some constant K:*

$$\sup_{z \in \mathcal{Z}} |\gamma(z)| \leq K, \ \forall \ \gamma \in \Gamma.$$

Let H be the convex conjugate of $v \mapsto (\psi^{-1}(v))^2$. Then for $\epsilon^2 \geq 2H(4K/\sqrt{n})$, we have

$$\mathbf{E}\left(\sup_{\sigma(\gamma - \gamma_0) \leq \epsilon} |\nu_n(\gamma - \gamma_0)| \right) \leq \psi(4\epsilon)$$

4 Further themes

The technical tools we shall use involve Hoeffding's and Bernstein's inequalities, and contraction inequalities (Ledoux and Talagrand [1991]).

Note that Theorem 3.1 is a statement about the mean of the empirical process. In the part on empirical process theory, we will also discuss concentration inequalities, which say that the empirical process is concentrated around its mean with large probability (Bousquet [2002], Massart [2000]).

Lemma 2.2, shows that a good choice of the model class Γ involves a trade-off between estimation error and approximation error. We will discuss penalized empirical risk minimization and the so-called oracle inequalities (del Barrio

et al. [2007]). In particular, we will look at high-dimensional (generalized) linear models (van de Geer [2006], van de Geer [2007]),and ℓ_q penalties ($0 \leq q \leq 1$), and at additive models involving many components.

For most of the results, we will provide a complete proof. And of course, we will discuss many examples.

References

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes rendus-Mathématique*, 334(6):495–500, 2002.

E. del Barrio, P. Deheuvels, and S. van de Geer. *Lectures on empirical processes: theory and statistical applications*. European Mathematical Society, Zürich, 2007.

M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Springer-Verlag New York, 1991.

P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000.

S.A. van de Geer. High dimensional generalized linear models and the Lasso. *Research report, to appear in Annals of Statistics*, 2006.

S.A. van de Geer. The deterministic Lasso. *JSM proceedings*, 2007.