M-estimators with ℓ_1-penalty

Sara van de Geer

May, 2012

Let $X_1, \ldots, X_n$ be independent observations with values in some observation space $\mathcal{X}$, and let for θ in a parameter space $\Theta \subset \mathbb{R}^p$ be given a loss function $\rho_\theta : \mathcal{X} \to \mathbb{R}$. The parameter θ is potentially high-dimensional, i.e. possibly $p \gg n$. We study the ℓ_1-regularized M-estimator

$$\hat{\theta} := \arg\min_{\theta \in \Theta} \left\{ P_n \rho_\theta + \lambda \|\theta\|_1 \right\}.$$

Here, we use the notation $P_n \rho_\theta := \sum_{i=1}^n \rho_\theta(X_i)/n$, i.e., it is the empirical measure of the function ρ_θ, often referred to as the empirical risk. Moreover, $\lambda > 0$ is a tuning parameter and $\|\theta\|_1 := \sum_{j=1}^p |\theta_j|$ is the ℓ_1-norm of θ.

A special case is the Lasso (Tibshirani [1996])), which has quadratic loss:

$$\rho_\theta(X) := (Y - \theta^T Z)^2, \; X = (Y, Z),$$

where $Y \in \mathbb{R}$ is the response variable and $Z \in \mathbb{R}^p$ are covariables.

It is known that generally the choice $\lambda \asymp \sqrt{\log p/n}$ is appropriate.

We address the following question: is the choice $\lambda \asymp \sqrt{\log p/n}$ also appropriate for nonlinear situations?

The example described above is a linear situation. More generally, we call the situation linear if for some $\psi : \mathcal{X} \to \mathbb{R}^p$

$$(P_n - P)(\rho_\theta - \rho_{\tilde{\theta}}) = (\theta - \tilde{\theta})^T (P_n - P)\psi, \ \forall \ \theta, \ \tilde{\theta},$$

where $P\rho_\theta := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\rho_\theta(X_i)$ is the theoretical risk.

For example any generalized linear model (GLM) loss function with canonical link function is a linear situation. Also density estimation using an exponential family is a linear situation. A non-linear situation occurs for instance in linear least squares regression with random design. Our focus is more on other examples, such as mixture models (Staedler et al. [2010]) or mixed effect models (Schelldorfer et al [2011]), etc.

Let us define the “true” parameter

$$\theta^0 := \arg\min_{\theta \in \bar{\Theta}} P\rho_\theta, \ \bar{\Theta} \supset \Theta.$$

Let $\theta^* \in \Theta$ be some “approximation” of θ^0.
The quantity that governs our choice for the tuning parameter is the behavior over ℓ_1-balls $\Theta_M(\theta^*) := \{\theta \in \Theta : \ \|\theta - \theta^*\|_1 \leq M\}$ of the empirical process $(P_n - P)(\rho_\theta - \rho_{\theta_*})$.

In the linear case, the supremum of the empirical process can be easily bounded using the dual norm inequality

$$\sup_{\theta \in \Theta_M(\theta^*)} |(P_n - P)(\rho_\theta - \rho_{\tilde{\theta}})| \leq \|(P_n - P)\psi\|_\infty M, \tag{1}$$

where for a vector $v \in \mathbb{R}^p$, $\|v\|_\infty := \max_{1 \leq j \leq p} |v_j|$ is the uniform norm. Moreover, for example for $\mathcal{N}(0, 1/n)$-random variables $\{V_j\}_{j=1}^p$ (say), it holds that

$$\max_{1 \leq j \leq p} |V_j| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right).$$

We show that in many non-linear cases, one still has

$$\sup_{\theta \in \Theta_M(\theta^*)} |(P_n - P)(\rho_\theta - \rho_{\theta^*})| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right). \tag{2}$$

This follows rather easily from a generic chaining (Talagrand [1996, 2005]) and Sudakov minoration argument.

Example:

The Gaussian mixture model

$$\rho_\theta(Y, Z) = \log\left(\sum_{k=1}^{r} \pi_k \phi_{\sigma_k}(Y - \beta_k^T Z_k)\right).$$

We call such a model an *extended* GLM.
In Staedler et al. [2010] the tuning parameter is taken of order $\lambda \asymp \sqrt{\log^3 n \log(p \vee n)/n}$ (and the penalty is $\lambda\|\beta\|_1 = \lambda \sum_{k=1}^{r} \|\beta_k\|_1$, i.e., it does not include the parameters π and σ).
The generic chaining argument gives a multivariate version of the contraction theorem. This leads to the reduced choice $\lambda \asymp \sqrt{\log p/n}$.

Our results rely on the following condition.

Condition (Componentwise Lipschitz condition) *There exist functions* $\{\psi_j\}$ *such that or all* θ *and* $\tilde{\theta}$ *in* Θ

$$\left|[\rho_\theta(X_i) - \mathbb{E}\rho_\theta(X_i)] - [\rho_{\tilde{\theta}}(X_i) - \mathbb{E}\rho_{\tilde{\theta}}(X_i)]\right| \leq \sum_{j=1}^{p} |\theta_j - \tilde{\theta}_j|\psi_j(X_i), \ \forall \ i.$$

We let for θ and θ^* in Θ,

$$Y(\theta, \theta^*) := (P_n - P)(\rho_\theta - \rho_{\theta^*}).$$

Define the excess risk

$$\mathcal{E}(\theta; \theta_0) := P(\rho_\theta - \rho_{\theta_0}).$$

For sets S, and vectors $\theta \in \mathbb{R}^p$, we let

$$\theta_{j,S} := \theta_j 1\{j \in S\}, \ j = 1, \ldots, p.$$

Definition (Compatibility of norms) *Let*

$$\delta(L, S) := \min\{\tau(\theta) : \ \|\theta_S\|_1 = 1, \|\theta_{S^c}\|_1 \leq L\}.$$

Then $\Gamma^2(L, S) := 1/\delta^2(L, S)$ *is called the effective sparsity (of the set* S*) and* $\phi^2(L, S) := |S|\delta^2(L, S)$ *is the compatibility constant.*

The following condition quantifies the curvature of $\mathcal{E}(\theta;\theta_0)$ around its minimizer θ_0.

Condition (Margin condition) *We say that the margin condition holds for all $\theta \in \Theta_M(\theta^*)$ if for some norm τ on Θ, and some strictly convex non-negative function G, satisfying $G(0) = 0$,*

$$\mathcal{E}(\theta;\theta^0) \geq G(\tau(\theta - \theta_0)),\ \forall\ \theta \in \Theta_M(\theta^*).$$

Definition (Convex conjugate) *Let G be a strictly convex non-negative function with $G(0) = 0$. The convex conjugate of G is*

$$H(v) := \sup_{u \geq 0}\left\{uv - G(u)\right\},\ v \geq 0.$$

We define the set

$$\mathcal{T}_M(\theta^*) := \{|Y(\theta, \theta^*)| \leq \lambda_0 \|\theta - \theta^*\|_1, \ \forall\, \theta \in \Theta_M(\theta^*)\},$$

and let $\mathcal{T}(\theta^*) := \mathcal{T}_\infty(\theta^*)$ (and $\Theta_\infty(\theta^*) = \Theta$).
Our task is to show that with $\lambda_0 \asymp \sqrt{\log p/n}$, the set $\mathcal{T}_M(\theta^*)$ has large probability (for any θ^* and suitable M).

Theorem

Let $\lambda > \lambda_0$. Assume the margin condition) for all $\theta \in \Theta$. If $\theta^0 \in \Theta$, we have on $\mathcal{T}(\theta_0)$, for all $0 < \delta < 1$,

$$(1-\delta)\mathcal{E}(\hat{\theta};\theta_0) + (\lambda - \lambda_0)\|\hat{\theta} - \theta^0\|_1 \leq \delta H\left(\frac{2\lambda\Gamma(L, S_0)}{\delta}\right), \tag{3}$$

with $L = (\lambda + \lambda_0)/(\lambda - \lambda_0)$. Moreover, for all $0 < \delta < 1$ and all $\theta^* \in \Theta$, on $\mathcal{T}(\theta^*)$,

$$(1-\delta)\mathcal{E}(\hat{\theta};\theta_0) + (\lambda - \lambda_0)\|\hat{\theta} - \theta^*\|_1$$

$$\leq 2\delta H\left(\frac{4(1+\delta)\lambda\Gamma(L_\delta, S_*)}{\delta^2}\right) + (1+\delta)\mathcal{E}(\theta^*;\theta_0), \tag{4}$$

with $L_\delta = 2\left((1+\delta)/\delta\right)\left((\lambda + \lambda_0)/(\lambda - \lambda_0)\right)$. Here $S_* := \{j : \theta^*_j \neq 0\}$ is the support set of θ^*.

The next theorem assumes convexity and then needs the margin condition only in a neighborhood of θ^*.

Theorem

Let $\lambda > \lambda_0$. Suppose Θ is a subset of a convex set $\bar{\Theta}$ and that the map $\theta \mapsto \rho_\theta$, $\theta \in \bar{\Theta}$, is convex. Let $(\lambda - \lambda_0)M_0$ and $(\lambda - \lambda_0)M_*$ be the first bound given in the previous theorem, i.e.,

$$M_0 := \frac{\delta}{\lambda - \lambda_0} H\left(\frac{2\lambda\Gamma(L, S_0)}{\delta}\right).$$

If $\theta^0 \in \Theta$ and the margin condition holds for all $\theta \in \Theta_{2M_0}(\theta_0)$ then on $\mathcal{T}_{2M_0}(\theta_0)$,

$$(1 - \delta)\mathcal{E}(\hat{\theta}; \theta_0) + (\lambda - \lambda_0)\|\hat{\theta} - \theta^0\|_1 \leq (\lambda - \lambda_0)M_0.$$

Theorem

(continued) Let $(\lambda - \lambda_0)M_*$ *be the second bound given in the previous theorem, i.e.,*

$$M_* := \frac{2\delta}{\lambda - \lambda_0}\left\{H\left(\frac{4(1+\delta)\lambda\Gamma(L_\delta, S_*)}{\delta^2}\right) + (1+\delta)\mathcal{E}(\theta^*; \theta_0)\right\}.$$

If the margin condition holds for all $\theta \in \Theta_{2M_*}(\theta^*)$ *then on* $\mathcal{T}_{2M^*}(\theta^*)$*,*

$$(1-2\delta)\mathcal{E}(\hat{\theta}; \theta_0) + (\lambda - \lambda_0)\|\hat{\theta} - \theta^*\|_1 \leq (\lambda - \lambda_0)M_*.$$

Symmetrization, contraction and deviation inequalities, and the peeling device

Let $g : \mathcal{X} \to \mathbb{R}$ satisfy $\mathbb{E}g^2(X_i) < \infty$ for all i. We use the notation

$$\|g\|_n^2 := \frac{1}{n}\sum_{i=1}^n (g(X_i) - \mathbb{E}g(X_i))^2, \ \|g\|^2 := \frac{1}{n}\sum_{i=1}^n \mathrm{var}(g(X_i)).$$

The sample is $\mathbf{X} := (X_1, \ldots, X_n)$. Let $\varepsilon_1, \cdots, \varepsilon_n$ be a Rademacher sequence independent of $\mathbf{X}$. We define the symmetrized empirical process

$$P_n^\varepsilon \rho_\theta := \frac{1}{n}\sum_{i=1}^n [\rho_\theta(X_i) - \mathbb{E}\rho_\theta(X_i)]\varepsilon_i,$$

and we let

$$Y^\varepsilon(\theta, \theta^*) := P_n^\varepsilon(\rho_\theta - \rho_{\theta^*}).$$

We reduce the problem to studying the conditional expectation

$$E_n := \mathbb{E}\left(\left[\sup_{\theta \in \Theta_M(\theta^*)} |Y^\varepsilon(\theta, \theta^*)| \right] \middle| \mathbf{X} \right).$$

Symmetrization

We cite the following result (Pollard[1984]).

Lemma

Let $R := \sup_{\theta \in \Theta_M(\theta^*)} \|\rho_\theta - \rho_{\theta^*}\|$ *and let* $t \geq 4$. *Then*

$$\mathbb{P}\left(\sup_{\theta \in \Theta_M(\theta^*)} |Y(\theta, \theta^*)| > 4R\sqrt{\frac{2t}{n}} \right) \leq 4\mathbb{P}\left(\sup_{\theta \in \Theta_M(\theta^*)} |Y^\varepsilon(\theta, \theta^*)| > R\sqrt{\frac{2t}{n}} \right).$$

Contraction

Suppose that for all $\theta, \tilde{\theta} \in \Theta$,

$$\left|[\rho_\theta(X_i) - \mathbb{E}\rho_\theta(X_i)] - [\rho_{\tilde{\theta}}(X_i) - \mathbb{E}\rho_{\tilde{\theta}}(X_i)]\right| \leq |f_\theta(X_i) - f_{\tilde{\theta}}(X_i)|, \ \forall\ i,$$

for some functions $f_\theta : \mathcal{X} \to \mathbb{R}$, $\theta \in \Theta$.
By the contraction inequality of Ledoux and Talagrand [1991]

$$E_n := \mathbb{E}\left(\left[\sup_{\theta \in \Theta_M(\theta^*)} |Y^\varepsilon(\theta, \theta^*)|\right] \middle| \mathbf{X}\right) \leq 2\mathbb{E}\left(\left[\sup_{\theta \in \Theta_M(\theta^*)} |X^\varepsilon(\theta, \theta^*)|\right] \middle| \mathbf{X}\right),$$

with

$$X^\varepsilon(\theta, \theta^*) := P_n^\varepsilon(f_\theta - f_{\theta^*}).$$

A deviation inequality

Write

$$R_n := \sup_{\theta \in \Theta_M(\theta^*)} \|\rho_\theta - \rho_{\theta^*}\|_n.$$

We have for all $t > 0$ (see Massart [2000]),

$$\mathbb{P}\left(\left[\sup_{\theta \in \Theta_M(\theta^*)} |Y^\varepsilon(\theta, \theta^*)| \right] \geq E_n + R_n \sqrt{\frac{2t}{n}} \right) \leq \exp[-t].$$

Combining this with the symmetrization result we obtain the following corollary.

Corollary

Let for some $\bar{R}$, $\sup_{\theta \in \Theta_M(\theta^*)} \|\rho_\theta - \rho_{\theta^*}\| \leq \bar{R}$, and let $t \geq 4$. Then for any $\bar{E}$,

$$\mathbb{P}\left(\left[\sup_{\theta \in \Theta_M} |Y(\theta, \theta^*)| \right] \geq 8\bar{E} + 4\bar{R}\sqrt{\frac{2t}{n}} \right)$$

$$\leq 4 \exp[-t] + 4\mathbb{P}(R_n > \bar{R} \vee E_n > \bar{E}).$$

The component wise Lipschitz condition yields

$$\sup_{\theta \in \Theta_M(\theta^*)} \|\rho_\theta - \rho_{\theta^*}\|_n \leq MK_n,$$

where

$$K_n := \max_{1 \leq j \leq p} \|\psi_j\|_n.$$

Thus, on the set

$$\mathcal{T}_0 := \left\{ \max_{1 \leq j \leq p} \|\psi_j\|_n \leq \bar{K} \right\}, \tag{5}$$

we can bound the random radii R_n by $M\bar{K}$. We will see that also

$$E_n \leq \lambda_0 MK_n,$$

for some constant $\lambda_0 \asymp \sqrt{\log p/n}$.

In some cases (regression with fixed design) the set $\mathcal{T}_0$ is not random, and the assumption $\max_{1\leq j\leq p}\|\psi_j\|_n \leq \bar{K}$ is a matter of normalization. In other situations, one can for example apply Bernstein's inequality:

Lemma

Suppose that for all j,

$$\frac{2L^2}{n}\sum_{i=1}^{n}\left[\mathbb{E}\exp[\psi_j^2(X_i)/L^2]-1\right]\leq\tau^2.$$

Then for all $t>0$,

$$\mathbb{P}\left(\max_{1\leq j\leq p}\left|\|\psi_j\|_n^2-\mathbb{E}\|\psi_j\|_n^2\right|\geq 2\tau L\sqrt{\frac{2(t+\log p)}{n}}+\frac{2L(t+\log p)}{n}\right)$$

$$\leq 2\exp[-t].$$

Proof. The sub-Gaussianity implies that for all $m \in \{1, 2, 3, \ldots\}$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}|\psi_j^2(X_i)|^m/n \leq \frac{L^{2m}m!}{n}\sum_{i=1}^{n}\left[\mathbb{E}\exp[\psi_j^2(X_i)/L^2]-1\right] \leq \frac{m!}{2}L^{2(m-1)}\tau^2.$$

But then

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}|\psi_j^2(X_i) - \mathbb{E}\psi_j^2(X_i)|^m \leq \frac{m!}{2}2^m L^{2(m-1)}\tau^2.$$

By Bernstein's inequality, for all $t > 0$,

$$\mathbb{P}\left(|(P_n - P)\psi_j^2| \geq 2\tau L\sqrt{\frac{2t}{n}} + \frac{2Lt}{n}\right) \leq 2\exp[-t],$$

and hence, by the union bound, for all $t > 0$,

$$\mathbb{P}\left(\max_{1\leq j\leq p}|(P_n - P)\psi_j^2| \geq 2\tau L\sqrt{\frac{2(t+\log p)}{n}} + \frac{2L(t+\log p)}{n}\right)$$
$$\leq 2\exp[-t].$$

The peeling device

The peeling device goes back to Alexander [1985], the terminology being introduced in vdG [2000].
Suppose all $M \leq \bar{M}$, and for all $t > 0$,

$$\mathbb{P}\left(\sup_{\theta \in \Theta_M(\theta^*)} |Y(\theta, \theta^*)| \geq \frac{\lambda_* M}{\mathrm{e}} \left(1 + K_* \left[\sqrt{\frac{t}{\log p}} + \frac{t}{n} \right] \right) \right) \leq 6 \exp[-t]. \quad (6)$$

For $M_j := \mathrm{e}^{-j} \bar{M}$, $j = 0, \ldots, p-1$, and all $t > 0$,

$$\mathbb{P}\left(\sup_{\theta \in \Theta_{\bar{M}}(\theta^*)} \frac{|Y(\theta, \theta^*)|}{\|\theta - \theta^*\|_1 \vee \mathrm{e}^{-(p-1)}} \geq \lambda_* \left(1 + K_* \left[\sqrt{\frac{t + \log p}{\log p}} + \frac{t + \log p}{n} \right] \right) \right)$$

$$\leq \sum_{j=1}^{p} \mathbb{P}\left(\sup_{\theta \in \Theta_{M_{j-1}}(\theta^*)} |Y(\theta, \theta^*)| > \lambda_* M_j \left(1 + K_* \left[\sqrt{\frac{t + \log p}{\log p}} + \frac{t + \log p}{n} \right] \right) \right)$$

$$\leq 6 \exp[\log p - (\log p + t)] \leq 6 \exp[-t].$$

Bounds for the symmetrized process

In the previous section we argued that the main task is to establish bounds for the expectation of the symmetrized process, i.e., for

$$\mathbb{E}\left(\left[\sup_{\theta \in \Theta_M(\theta^*)} |Y^{\varepsilon}(\theta, \theta^*)| \right] \middle| \mathbf{X} \right).$$

One can then derive deviation inequalities , and hence theoretical bounds for the tuning parameter of the ℓ_1-regularized M-estimator.

Linear functions Lets us briefly recall the linear case. Let ρ_θ be linear:

$$\rho_\theta(x) := \sum_{j=1}^{p} \theta_j \psi_j(x).$$

One then clearly has

$$\mathbb{E}\left(\left[\sup_{\theta \in \Theta_M(\theta^*)} |Y^\varepsilon(\theta, \theta^*)| \right] \middle| \mathbf{X} \right) \leq M \|\varepsilon^T \psi / n\|_\infty.$$

Moreover, by Hoeffding's inequality,

$$\mathbb{E}\left(\left[\max_{1 \leq j \leq p} |\varepsilon^T \psi_j / n \right] \middle| \mathbf{X} \right) \leq \sqrt{\frac{2 \log(2p)}{n}} K_n,$$

where $K_n := \max_{1 \leq j \leq p} \|\psi_j\|_n$.

Generalized linear functions

Suppose that for all $\theta, \tilde{\theta} \in \Theta$, Suppose that for all $\theta, \tilde{\theta} \in \Theta$,

$$\left| [\rho_\theta(X_i) - \mathbb{E}\rho_\theta(X_i)] - [\rho_{\tilde{\theta}}(X_i) - \mathbb{E}\rho_{\tilde{\theta}}(X_i)] \right| \leq |f_\theta(X_i) - f_{\tilde{\theta}}(X_i)|, \ \forall \ i,$$

where $f_\theta = \sum_{j=1}^p \theta_j \psi_j$, $\theta \in \Theta$. Then by the contraction inequality, and the arguments for the linear case

$$\mathbb{E}\left(\left[\sup_{\theta \in \Theta_M(\theta^*)} |Y^\varepsilon(\theta, \theta^*)| \middle| \mathbf{X} \right] \right) \leq 2M \sqrt{\frac{2\log(2p)}{n}} K_n,$$

with $K_n := \max_{1 \leq j \leq p} \|\psi_j\|_n$.

Extended generalized linear functions

Condition (Extended GLM condition) *The exist non-negative functions* $\{\psi_{j,k} : j = 1, \ldots p_k,\ k = 1, \ldots, r\}$ *(with* $\sum_{k=1}^{r} p_k = p$*) such that for all* θ *and* $\tilde{\theta}$*, it holds that*

$$|\rho_\theta - \rho_{\tilde{\theta}}| \leq \sum_{k=1}^{r} |\sum_{j=1}^{p_k} (\theta_{j,k} - \theta_{j,k})\psi_{j,k}|.$$

Theorem

(Multivariate contraction theorem) Assume the extended GLM condition. Let $\xi_{1,k}, \ldots, \xi_{n,k}$, $k = 1, \ldots, r$, *be independent* $\mathcal{N}(0,1)$*-distributed random variables, independent of* $X_1, \ldots, X_n$. *Let*

$$X_k(\theta, \theta^*) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{p_k} (\theta_{j,k} - \theta^*_{j,k}) \psi_{j,k}(X_i) \xi_{i,k},$$

and

$$X(\theta, \theta^*) := \sum_{k=1}^{r} X_k(\theta, \theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{k=1}^{r} \sum_{j=1}^{p_k} (\theta_{j,k} - \theta^*_{j,k}) \psi_{j,k}.$$

Then for a universal constant C,

$$\mathbb{E}\left(\left[\sup_{\theta \in \Theta_M(\theta^*)} |Y^{\varepsilon}(\theta, \theta^*)| \right] \middle| \mathbf{X} \right) \leq 2^{r-1} C \mathbb{E}\left(\left[\sup_{\theta} X(\theta, \theta^*) \right] \middle| \mathbf{X} \right).$$

Proof. Note first that

$$\mathbb{E}\left(|X_\theta - X_{\tilde{\theta}}|^2|\mathbf{X}\right) = \sum_{k=1}^{r} \| \sum_{j=1}^{p_k} (\theta_{j,k} - \tilde{\theta}_{j,k})\psi_{j,k}\|_n^2.$$

For all θ and $\tilde{\theta}$ we have

$$\|\rho_\theta - \rho_{\tilde{\theta}}\|_n^2 \leq \| \sum_{k=1}^{r} | \sum_{j=1}^{p_k} (\theta_{j,k} - \tilde{\theta}_{j,k})\psi_{j,k}|\|_n^2 \leq 2^{r-1} \sum_{k=1}^{r} \| \sum_{j} \theta_{j,k}\psi_{j,k}\|_n^2.$$

Let $d^2(\theta, \tilde{\theta}) := \sum_{k=1}^{r} \| \sum_j \theta_{j,k}\psi_{j,k}\|_n^2$. By Hoeffding's inequality

$$\mathbb{P}(|Y^\varepsilon(\theta, \theta^*) - Y^\varepsilon(\tilde{\theta}, \theta^*)| \geq 2^{r-1} d(\theta, \tilde{\theta})\sqrt{2t}) \leq 2\exp[-t].$$

Hence, using Theorem 2.1.5 in Talagrand's book we get for a universal constant C,

$$\mathbb{E}\left(\left[\sup_{\theta \in \Theta_M(\theta^*)} |Y^\varepsilon(\theta, \theta^*)|\right] \middle| \mathbf{X}\right) \leq 2^{r-1} C \mathbb{E}\left(\left[\sup_{\theta} X(\theta, \theta^*)\right] \middle| \mathbf{X}\right).$$

As a direct consequence we obtain the bounds of interest for our problem.

Theorem

Assume the extended GLM condition and let $K_n := \max_{j,k} \|\psi_{j,k}\|_n$. *We have for a universal constant C,*

$$\mathbb{E}\left(\left[\sup_{\theta\in\Theta_M(\theta^*)} |Y^{\varepsilon}(\theta,\theta^*)|\right]\middle|\mathbf{X}\right) \leq C\sqrt{\frac{2\log(2p)}{n}}K_n.$$

Proof. We have

$$\mathbb{E}\left(\left[\sup_{\theta\in\Theta_M(\theta^*)} X(\theta,\theta^*)\right]\middle|\mathbf{X}\right) \leq M\sqrt{\frac{2\log(2p)}{n}}K_n.$$

□

Non-linear functions We assume ρ_θ is components-wise Lipschitz in θ. Define for $\psi = (\psi_1, \ldots, \psi_p)^T$,

$$\Sigma_n := \frac{1}{n} \sum_{i=1}^{n} \psi(X_i)\psi^T(X_i).$$

Let $\underline{\Lambda}_n^2$ be the smallest eigenvalue of Σ_n and $\bar{\Lambda}_n^2$ be its largest eigenvalue. We assume that $\underline{\Lambda}_n > 0$, thus excluding the case $p > n$.

Theorem

For a universal constant C, it holds that

$$\mathbb{E}\left(\left[\sup_{\theta \in \Theta_M(\theta^*)} |Y^\varepsilon(\theta, \theta^*)| \right] \middle| \mathbf{X} \right) \leq CM\sqrt{\frac{2\log(2p)}{n}} \left(\bar{\Lambda}_n / \underline{\Lambda}_n \right).$$

Proof. Use that

$$\| \sum_{j=1}^{k} \theta_j \psi_j \|_n^2 \geq \underline{\Lambda}_n^2 \|\theta\|_2^2,$$

and

$$k$$

The geometry of ℓ_1-balls

We describe here the generic chaining bound, specialized to our context and notation adjusted to our setting. Let $\xi_1, \ldots, \xi_n$ be independent $\mathcal{N}(0,1)$-distributed random variables, and $\mathcal{V}$ be a subset of $\mathbb{R}^n$. Define

$$X_v := \frac{1}{n}\sum_{i=1}^{n} v_i \xi_i, \ \|v\|_n^2 := \frac{1}{n}\sum_{i=1}^{n} v_i^2.$$

A sequence of partitions $\{\mathcal{A}_s\}_{s=0}^{\infty}$ of $\mathcal{V}$ is *admissible* if it is an increasing sequence. For each $v \in \mathcal{V}$ and each s, the set $A_s(v)$ is defined as the unique element of $\mathcal{A}_s$ that contains v, and $\Delta(A_s(v))$ as the diameter of $A_s(v)$. Write

$$\gamma_2(\mathcal{V}, \|\cdot\|_n) := \inf \sup_{v \in \mathcal{V}} \sum_{s \geq 0} 2^{s/2} \Delta(\mathcal{A}_s(v)),$$

where the infimum is taken over all admissible partitions.

Theorem

(The majorizing measure theorem, see Talagrand [2005]) For some universal constant C, we have

$$\frac{1}{C}\gamma_2(\mathcal{V}, \|\cdot\|_n) \leq \mathbb{E}\left[\sup_{v\in\mathcal{V}} X_v\right] \leq C\gamma_2(\mathcal{V}, \|\cdot\|_n).$$

Talagrand derives the lower bound in the above theorem from Sudakov's minoration argument. As a consequence, Talagrand presents the following result.

Theorem

(Talagrand [2005]) Let $\{Y_v : v \in \mathcal{V}\}$ *be a stochastic process that satisfies for all* $t > 0$

$$\mathbb{P}\left(|Y_v - Y_{\tilde{v}}| \geq \|v - \tilde{v}\|_n \sqrt{t} \right) \leq 2\exp[-t], \ \forall \ v, \ \tilde{v}.$$

Then for a universal constant C, we have

$$\mathbb{E}\left[\sup_{v,\tilde{v} \in \mathcal{V}} |Y_v - Y_{\tilde{v}}| \right] \leq C\mathbb{E}\left[\sup_{v \in \mathcal{V}} X_v \right].$$

Let us compare here the situation with Dudley's entropy bound. We formulate it using chaining along a tree, as in BvdG[2011]. Define $R_n := \sup_{v \in \mathcal{V}} \|v\|_n$. Let for each $s \in \{0, 1, \ldots, S\}$, $\{v_j^s\}_{j=1}^{N_s} \subset \mathcal{V}$ be a minimal $2^{-s}R_n$-covering set of $\mathcal{V}$, that is, for all $v \in \mathcal{V}$ and all s there is a v_j^s such that $\|v - v_j^s\|_n \leq 2^{-s}R_n$. Then for all v, we can find a end node $v^S \in \{v_j^S\}$ such that $\|v - v^S\|_n \leq 2^S R_n$, and for each end node $v^S \in \{v_j^S\}$ one can find a branch $\{v^0, \ldots, v^S\}$ such that $\|v^s - v^{s-1}\|_n \leq 2^{-s-1}R_n$ for all $s = 1, \ldots, S$.

Moreover, we can write

$$X_v = \sum_{s=0}^{S} (X_{v^s} - X_{v^{s-1}}) + X_v - X_{v^S}.$$

Since

$$|X_v - X_{v^S}| \leq 2^{-S} R_n \sqrt{\sum_{i=1}^{n} \xi_i^2 / n},$$

one now arrives at Dudley's bound

$$\mathbb{E}\left[\sup_{v \in \mathcal{V}} X_v\right] \leq \sum_{s=0}^{S} 2^{-(s-1)} R_n \sqrt{\frac{2 \log(2N_s)}{n}} + 2^{-S} R_n. \tag{7}$$

Consider now a special case. We let $\{\psi_j\}_{j=1}^p$ be p vectors in $\mathbb{R}^n$, and let

$$\mathcal{V} := \{\sum_{j=1}^p \theta_j \psi_j : \ \|\theta\|_1 \leq 1\}.$$

Let $K_n := \max_{1 \leq j \leq p} \|\psi_j\|_n$.
The following lemma rephrases the first part of Theorem 2.1.6 in Talagrand [2005]

Lemma

It holds for some universal constant C that

$$\gamma_2(\mathcal{V}, \|\cdot\|_n) \leq C\sqrt{\frac{2\log(2p)}{n}} K_n.$$

Indirect Proof.

Clearly, by the dual norm inequality

$$\sup_{v\in\mathcal{V}} X_v = \sup_{\|\theta\|_1\leq 1} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\theta_j\psi_{i,j}\xi_i = \max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\psi_{i,j}\xi_i\right|.$$

Hence,

$$\mathbb{E}\left[\sup_{v\in\mathcal{V}} X_v\right] \leq \mathbb{E}\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\psi_{i,j}\xi_i\right| \leq \sqrt{\frac{2\log(2p)}{n}}K_n.$$

The result now follows from Theorem 10. □

In his book, Talagrand now poses the research question to prove the above lemma directly. We claim that this cannot be done by applying Dudley's bound. It holds that

$$\log(2N_s) \leq 2^{2s} \log(4p), \ \forall \ s. \tag{8}$$

Insert this in (7) with the bound $R_n \leq K_n$, to find that

$$\mathbb{E}\left[\sup_{v \in \mathcal{V}} X_v\right] \leq 2(S+1)K_n \sqrt{\frac{2\log(4p)}{n}} + 2^{-S}K_n.$$

Minimizing this over S gives a bound of order $(\log n)\sqrt{\log p/n}K_n$.

Concluding remarks

We summarized results concerning symmetrization, contraction, deviation inequalities and chaining.Their application in statistical theory has been highlighted by Massart [2000]. We have added now a new application, where generic chaining allows one to remove additional $\log n$ factors. For example, we have improved the choice $\lambda \asymp \sqrt{\log^3 n \log(p \vee n)/n}$ in Staedler et al. [2010] to $\lambda \asymp \sqrt{\log p/n}$. The geometric arguments to bound γ_2 in the case of convex hulls are still to be developed. Somehow, the generic chaining bound γ_2 better exploits the impossibility to play cat and mouse.

THANK YOU!