M-estimation

and Complexity Regularization

Sara van de Geer
Seminar für Statistik, ETH Zürich

Toulouse, June, 2008

Program

1. Notation and motivation
2. Probability and moment inequalities
3. Empirical risk minimization over a finite class
4. Empirical risk minimization over an infinite class
5. Symmetrization
6. Entropy
7. The expectation of the empirical process

1. Notation and motivation

Let $Z_1, \ldots, Z_n$ be independent random variables in $\mathcal{Z}$.

Notation: for $\gamma : \mathcal{Z} \rightarrow \mathbf{R}$,

the theoretical measure is

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

and the empirical measure is

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

We moreover write

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

Consider a class Γ of functions γ on $\mathcal{Z}$.

EPT (Empirical Process Theory) is about the study of

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

as process indexed by Γ.

In particular, the study of probability and moment inequalities for

$$\mathbf{V}_n := \sup_{\gamma \in \Gamma} |\nu_n(\gamma)|.$$

Statistical motivation:

We will consider empirical risk minimization (M-estimation).

Let Γ be a class of loss functions, indexed by a parameter.

Parametric:

$$\Gamma = \{\gamma_\theta : \theta \in \Theta\}, \ \Theta \subset \mathbf{R}^r.$$

Nonparametric:

$$\Gamma = \{\gamma_f : f \in \mathcal{F}\},$$

with $\mathcal{F}$ some collection of functions.

Empirical risk minimizer

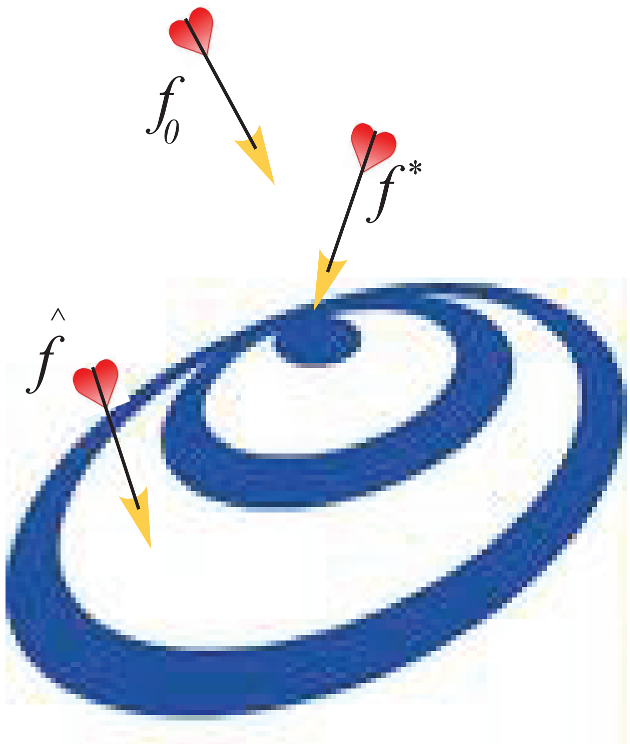
$$\hat{f} := \arg\min_{f \in \mathcal{F}} P_n \gamma_f, \ \hat{\gamma} := \gamma_{\hat{f}}.$$

Let $\mathbf{F} \supset \mathcal{F}$. The target is

$$f^0 := \arg\min_{f \in \mathbf{F}} P \gamma_f, \ \gamma^0 := \gamma_{f^0}.$$

Best approximation in the class

$$f^* := \arg\min_{f \in \mathcal{F}} P \gamma_f, \ \gamma^* := \gamma_{f^*}.$$


f_0
f^*
$\hat{f}$

Excess risk

$$\mathcal{E}(\gamma) := P(\gamma - \gamma^0).$$

Approximation error

$$\mathcal{E}^* := \mathcal{E}(\gamma^*).$$

Basic inequality: For $\hat{\mathcal{E}} = \mathcal{E}(\hat{\gamma})$,

$$\hat{\mathcal{E}} \leq -(P_n - P)(\hat{\gamma} - \gamma^*) + \mathcal{E}^*.$$

Proof. $\ldots$

Define

$$\sigma^2(\gamma) := \frac{1}{n}\sum_{i=1}^{n} E[\gamma(Z_i) - E\gamma(Z_i)]^2.$$

Let $\hat{\sigma} := \sigma(\hat{\gamma} - \gamma^0)$, $\sigma^* := \sigma(\gamma^* - \gamma^0)$.
More generally, let d be some metric on $\mathbf{F}$, and $\hat{d} := d(\hat{f}, f^0)$, $d^* := d(f^*, f^0)$.

Let $\psi(\cdot)$ (some concave function) be the "modulus of continuity" of the empirical process, that is, ψ is such that

$$\mathbf{V}_n := \sup_{\gamma \in \Gamma} \frac{|\sqrt{n}(P_n - P)(\gamma - \gamma^*)|}{\psi(d(f, f^0) \vee d^*)}$$

is a "bounded" random variable.

Let $G(\cdot)$ (some convex function) be the margin, i.e.,

$$\mathcal{E}(\gamma_f) \geq G(d(f, f^0)), \; \forall \; f \in \mathcal{F}.$$

Example(Classification)
Suppose that $Z_i = (X_i, Y_i)$, with $Y_i \in \mathcal{Y} := \{0, 1\}$ a label, $i = 1, \ldots, n$. Let **F** be a class of functions $f : \mathcal{X} \to [0, 1]$. We consider 0/1-loss

$$\gamma_f(x, y) = \gamma(f(x), y) := (1-y)f(x)+y(1-f(x)).$$

For $a \in [0, 1]$, write

$$l(a, \cdot) := \mathbb{E}(\gamma(a, Y_i)|X_i = \cdot)$$

$$= (1 - \eta)a + \eta(1 - a) = a(1 - 2\eta) + \eta \ ,$$

where $\eta = \mathbb{E}(Y_i|X_i = \cdot)$.

The target is the overall minimizer

$$f^0 := \arg \min_{a \in [0,1]} l(a, \cdot) \ .$$

It is clear that f^0 is the Bayes rule

$$f^0 = \mathrm{l}\{1 - 2\eta < 0\} + q\{1 - 2\eta = 0\} \ ,$$

with q an arbitrary value in $[0, 1]$.
We moreover have

$$P(\gamma_f - \gamma_{f^0}) = P|(f - f^0)(1 - 2\eta)| \ .$$

Consider the functions

$$H_1(v) = vP\mathrm{l}\{|1-2\eta| < v\},\ v \in [0,1]\ ,$$

and

$$G_1(u) = \max_v\{uv - H_1(v)\},\ u \in [0,1]$$

(assuming the maximum exists).

Lemma *The inequality*

$$P(\gamma_f - \gamma_{f^0}) \geq G(\sigma(\gamma_f - \gamma_{f^0}))$$

holds with $G(u) = G_1(u^2)$, $u \in [0,1]$.

Suppose that $G_\psi := G \circ \psi^{-1}$ is strictly convex.

Definition The convex conjugate of G_ψ is

$$H_\psi(v) := \sup_u [uv - G_\psi(u)].$$

Lemma For all $0 < \lambda_n^2 < 1$,

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

Proof. $\ldots$

Bernstein's inequality

Let $\gamma_j : \mathcal{Z} \to \mathbf{R}$, $j = 1, \ldots, p$.

Assume that for all j, i *and* $m \geq 2$,

$$E\gamma_j(Z_i) = 0, \; P|\gamma_j - \gamma^*|^m \leq \frac{m!}{2}(2K)^{m-2} d^2(f_j, f^*).$$

Then for all $m \leq 1 + \log p$,

$$\mathbb{E}^{1/m}\left(\max_{1 \leq j \leq p} \frac{|P_n(\gamma_j - \gamma^*)|}{d(f_j, f^*) \vee \tau} \right)^m \leq \lambda_n + \frac{K\lambda_n^2}{\tau},$$

where $\lambda_n^2 := \frac{2\log(2p)}{n}$.

Moreover, for all $t > 0$,

$$\mathbf{P}\left[\max_{1\le j\le p}\frac{|P_n(\gamma_j-\gamma^*)|}{d(f_j,f^*)\vee\tau}\ge\sqrt{\lambda_n^2+2t}+\frac{K(\lambda_n^2+2t)}{\tau}\right]$$

$$\le\exp[-nt].$$

Bousquet's inequality. *Let* $\gamma : \mathcal{Z} \to \mathbf{R}$, $\gamma \in \Gamma$. *Assume that for all* γ, i,

$$E\gamma(Z_i) = 0,\ |\gamma - \gamma^*| \leq 2K.$$

Let

$$\mathbf{Z} := \sup_{\gamma \in \Gamma} \frac{|P_n(\gamma - \gamma^*)|}{\sigma(\gamma - \gamma^*) \vee \tau}.$$

Then $\forall\, t > 0$,

$$\mathbb{P}\left(\mathbf{Z} \geq \mathbb{E}\mathbf{Z} + \sqrt{2t}\sqrt{1 + 4K\mathbb{E}\mathbf{Z}} + \frac{2tK}{3} \right) \leq \mathrm{e}^{-nt}.$$

Hoeffding's inequality

Suppose for $1 \leq j \leq p$,

$$E\gamma_j(Z_i) = 0,\ |(\gamma_j - \gamma^*)(Z_i)| \leq c_{i,j}\ \forall\ i.$$

Let

$$d^2(\gamma_j, \gamma^*) := \frac{1}{n}\sum_{i=1}^{n} c_{i,j}^2.$$

Then

$$\mathbb{E}\left(\max_{1\leq j\leq p} \frac{|P_n(\gamma_j - \gamma^*)|}{d(\gamma_j, \gamma^*)}\right) \leq \lambda_n.$$

Moreover, for all $t > 0$*,*

$$\mathbf{P}\left(\max_{1\leq j\leq p} \frac{|P_n(\gamma_j - \gamma^*)|}{d(\gamma_j, \gamma^*)} \geq \ \lambda_n + \sqrt{2t}\right)$$

$$\leq \exp\left[-nt\right].$$

3. Empirical risk minimization over a finite class

Let $\gamma_j : \mathcal{Z} \rightarrow \mathbf{R}$, $j = 1, \ldots, p$ be given loss functions in a class $\Gamma \subset \mathbf{\Gamma}$. We define the model selection estimator

$$P_n \hat{\gamma} := \min_{1 \leq j \leq p} P_n \gamma_j .$$

The target is

$$P \gamma^0 := \arg \min_{\gamma \in \mathbf{\Gamma}} P \gamma .$$

The best approximation is

$$P \gamma^* := \min_{1 \leq j \leq p} P \gamma_j .$$

We define the excess risks

$$\hat{\mathcal{E}} := P(\hat{\gamma} - \gamma^0),$$

and

$$\mathcal{E}^* := P(\gamma^* - \gamma^0).$$

Moreover, we let

$$\sigma^2(\gamma) := \frac{1}{n} \sum_{i=1}^{n} E[\gamma(Z_i) - E\gamma(Z_i)]^2.$$

3.1 Bounded loss, standard margin condition

Lemma. *Suppose*

$$\mathbf{P}|\gamma_j - \gamma^*|^m \leq \frac{m!}{2} K^{m-2} d^2(f_j, f^*),$$

and the standard margin condition

$$\mathcal{E}(\gamma_j) \geq d^2(f_j, f^*)/C.$$

Then for $\mathcal{E}^* \geq \lambda_n^2$*,*

$$\mathbb{E}^{1/m} \left(\sqrt{\frac{\hat{\mathcal{E}}}{\mathcal{E}^*}} \right)^m \leq 1 + \sqrt{\frac{C\lambda_n^2}{\mathcal{E}^*}} + \frac{K\lambda_n^2}{\mathcal{E}^*}.$$

Remark. The result with $m = 2$ reads

$$\mathbb{E}\left[\frac{\hat{\mathcal{E}}}{\mathcal{E}^*}\right] \leq \left[1 + \sqrt{\frac{C\lambda_n^2}{\mathcal{E}^*}} + \frac{K\lambda_n^2}{\mathcal{E}^*}\right]^2 .$$

Corollary. *When*

$$\mathcal{E}^* \gg (K + C)\,\lambda_n^2,$$

it holds that

$$\mathbb{E}\left(\sqrt{\frac{\hat{\mathcal{E}}}{\mathcal{E}^*}}\right)^m \to 1.$$

Example: density estimation.

Define

$$\hat{\mathcal{K}} := P(\gamma_{(\hat{f}+f^*)/2} - \gamma_{f^0}),$$

and

$$\mathcal{K}^* := P(\gamma_{f^*} - \gamma_{f^0}).$$

Lemma *Suppose that*

$$\sqrt{\frac{f^0}{f^*}} \le \frac{C}{8} .$$

Then

$$\mathbb{E}^{1/m}\left(\frac{\hat{\mathcal{K}}}{\mathcal{K}^*}\right)^{m/2} \le 1 + C\sqrt{\frac{\lambda_n^2}{\mathcal{K}^*}} + \frac{\lambda_n^2}{\mathcal{K}^*}.$$

3.2 Bounded loss, general margin condition

Lemma *Suppose that the margin condition holds, with strictly convex margin function* G. *Let* H *be the convex conjugate of* G. *Assume that for some* $r \leq 1 + \log p$, *the function* $H(v^{\frac{1}{r}}),\ v > 0$, *is concave. Assume moreover that the exponential moment condition holds for some* $K > 0$.

Then for all $0 < \delta < 1$*, and* $\varepsilon > 0$*, we have*

$$(1-\delta)\mathbb{E}\hat{\mathcal{E}} \leq 2\delta H \left(\sqrt{\frac{\lambda_n^2}{\delta}} + \frac{K\lambda_n^2}{2\delta G^{-1}(\mathcal{E}^* \vee \varepsilon)} \right) + (1+\delta)\mathcal{E}^* \, .$$

Lemma.

Suppose

$$\mathbf{P}|\gamma_j - \gamma^*|^m \leq \frac{m!}{2} K^{m-2} d^2(f_j, f^*),$$

and the general margin condition

$$\mathcal{E}(\gamma_j) \geq d^{2\kappa}(f_j, f^0)/C.$$

Then,

$$\mathbb{E}^{1/m}\left(\left(\frac{\hat{\mathcal{E}}}{\mathcal{E}^*}\right)^{\frac{1}{2\kappa}}\right)^m \leq 1$$

$$+\bar{c}_\kappa \left(\sqrt{\frac{C^{\frac{1}{\kappa}} \lambda_n^2}{(\mathcal{E}^*)^{\frac{2\kappa-1}{\kappa}}}} + \frac{K\lambda_n^2}{\mathcal{E}^*} \right)^{\frac{1}{2\kappa-1}},$$

where

$$\bar{c}_\kappa := \left(\frac{1 + (2\kappa - 1)^{\frac{1}{2\kappa-1}}}{(2\kappa)^{\frac{1}{2\kappa-1}}} \right).$$

3.3 Unbounded loss, under standard margin condition

Lemma. *Suppose*

$$|\gamma_j(\cdot) - \gamma_l(\cdot)| \le \mathbf{K}(\cdot),\ \forall (j, l),$$

and that for some $s > 1$,

$$\|\mathbf{K}\|_s^s := P\mathbf{K}^s < \infty.$$

Assume the standard margin condition

$$\mathcal{E}(\gamma_j) \ge \sigma^2(\gamma_j - \gamma^0)/C,\ \forall j.$$

Then for $\log p \geq 1$,

$$\mathbb{E}^{1/2}\left(\frac{\hat{\mathcal{E}}}{\mathcal{E}_*}\right) \leq 1 + \sqrt{\frac{C\lambda_n^2}{\mathcal{E}^*}}$$

$$+c_s(2\|\mathbf{K}\|_s)^{\frac{s}{s+1}}\left(\lambda_n^2\right)^{\frac{s-1}{s+1}}(\mathcal{E}^*)^{-\frac{s}{s+1}},$$

where

$$c_{\mathcal{S}} := \left(\frac{s-1}{2}\right)^{\frac{2}{s+1}} + \left(\frac{s-1}{2}\right)^{-\frac{s-1}{s+1}}.$$

Corollary. *When*

$$\mathcal{E}^* \gg 2c_{\mathcal{S}}^{\frac{s+1}{s}} \|\mathbf{K}\|_s \left(\lambda_n^2\right)^{\frac{s-1}{s}} + C\lambda_n^2,$$

it holds that

$$\mathbb{E}\frac{\hat{\mathcal{E}}}{\mathcal{E}^*} \to 1.$$

4. Empirical risk minimization over an infinite class

Estimator

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

Target

$$\gamma^0 := \arg\min_{\gamma \in \mathbf{\Gamma}} P\gamma, \ \mathbf{\Gamma} \supset \Gamma.$$

Best approximation

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

We assume $\forall \ \gamma \in \Gamma$,

1. boundedness: $\|\gamma - \gamma^*\| \leq K$,

2. margin condition: $\mathcal{E}(\gamma) \geq G(\sigma(\gamma - \gamma^0))$.

Recall: we sketched a result involving the weighted empirical process

$$\mathbf{V}_n := \sup_{\gamma \in \Gamma} \frac{\sqrt{n}|(P_n - P)(\gamma - \gamma^*)|}{\psi(\sigma(\gamma - \gamma^0) \vee \sigma^*)}.$$

We will now show how to obtain this from the un-weighted process, using the peeling device.

Let for $\sigma > 0$,

$$\mathbf{Z}_n(\sigma) := \sup_{\sigma(\gamma - \gamma^0) \le \sigma} |(P_n - P)(\gamma - \gamma^*)|.$$

Assume that (for some σ_0 and) all $\sigma \geq \sigma_0$, we have the
3. increments of the empirical process:

$$\mathbb{E}\mathbf{Z}_n(\sigma) \leq \psi(\sigma),$$

where $\psi(\sigma) \geq \sigma$.

Let (as before)

$$G_\psi := G \circ \psi^{-1},$$

so that

$$G_\psi^{-1} = \psi \circ G^{-1}.$$

Assume the
4. decay condition: there is an $0 < \alpha < 1$ such that,

$$\frac{G_\psi^{-1}(\epsilon)}{\epsilon^\alpha} \downarrow \text{ in } \epsilon.$$

Peeling device. *Let* $\epsilon \geq G(\sigma_0)$. *Then*

$$\mathbb{E}\left(\sup_{\mathcal{E}(\gamma) > \epsilon} \frac{\sqrt{n}|(P_n - P)(\gamma - \gamma^*)|}{\mathcal{E}(\gamma)} \right) \leq C_\alpha \frac{G_\psi^{-1}(\epsilon)}{\epsilon},$$

where $C_\alpha := \alpha^{-\frac{\alpha}{1-\alpha}}/(1-\alpha)$.

Theorem *Let* $\epsilon_t \geq G(\sigma_0)$ *and*

$$(1-\delta)\epsilon_t > \delta H_\psi \left(\frac{4C_\alpha + 2\sqrt{2t}}{\delta\sqrt{n}} \right) + \frac{2Kt}{3n} + \mathcal{E}^*,$$

where H_ψ *is the convex conjugate of* G_ψ. *Then*

$$\mathbf{P}(\hat{\mathcal{E}} > \epsilon_t) \leq \exp[-t].$$

5. Symmetrization

Definition A Rademacher sequence is a sequence of independent random variables $\{\xi_i\}_{i=1}^n$ with

$$\mathbf{P}(\xi_i = +1) = \mathbf{P}(\xi_i = -1) = \frac{1}{2}.$$

Notation

Let $\{Z_i'\}$ be an independent copy of $\{Z_i\}$, and let $\{\xi_i\}$ be a Rademacher sequence, independent of $\{Z_i\}$ and $\{Z_i'\}$. Define

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

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

and write

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

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

$$\|P_n^{\xi}\|_{\Gamma} := \sup_{\gamma \in \Gamma} |P_n^{\xi}\gamma|.$$

Lemma *We have*

$$\mathbb{E}\|P_n - P\|_{\Gamma} \leq 2\mathbb{E}\|P_n^{\xi}\|_{\Gamma}.$$

6. Entropy

Let (Γ, d) be a subset of a metric space.

Definition

For $u > 0$, a u-covering of Γ is defined as a collection $\{\gamma_j\}_{j=1}^N$ such that
$\forall \gamma$ there is a γ_j with $d(\gamma, \gamma_j) \leq u$.

The covering number $N(\cdot, \Gamma, d)$ is defined for all $u > 0$ as
$N(u, \Gamma, d) := \min\{N :$ there is a u-covering $\{\gamma_j\}_{j=1}^N\}$.

The entropy is $\mathcal{H}(\cdot, \Gamma, d) := \log(1 + N(\cdot, \Gamma, d))$.

Example. Let

$$\Gamma = \{\gamma : [0,1] \to [0,1] :\ \|\gamma^{(m)}\|_\infty \leq 1\}.$$

Then

$$\mathcal{H}(u, \Gamma, \|\cdot\|_\infty) \leq A_m u^{-\frac{1}{m}},\ u > 0.$$

Example. Let

$$\Gamma = \{\gamma : \mathbf{R} \to [0,1] :\ \gamma \uparrow\}.$$

Let Q be some probability measure and $\|\cdot\|_Q$ be the $L_2(Q)$-norm. Then

$$\mathcal{H}(u, \Gamma, \|\cdot\|_Q) \leq A u^{-1},\ u > 0.$$

7. Moment inequalities for an infinite class

Let Γ be some class of functions.

We assume

$$\sup_{\text{prob. measures } Q} \mathcal{H}(\cdot, \Gamma, \|\cdot\|_Q) \le \mathcal{H}(\cdot),$$

and write

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

Let $\|\cdot\|$ be the $L_2(P)$-norm and $\|\cdot\|_n$ be the $L_2(P_n)$-norm.

Define

$$\sigma := \sup_{\gamma \in \Gamma} \|\gamma\|, \ \hat{\sigma} := \sup_{\gamma \in \Gamma} \|\gamma\|_n.$$

Lemma *We have*

$$\mathbb{E}\|P_n^\xi\|_\Gamma \leq \frac{\mathbb{E}\psi(\hat{\sigma})}{2\sqrt{n}}.$$

Proof. ...

Contraction principle *(Ledoux and Talagrand (1991)) (It holds more generally for Lipschitz functions.) Suppose* $\|\gamma\|_\infty \leq K$ *for all* $\gamma \in \Gamma$. *Then*

$$\mathbb{E}\left(\sup_{\gamma\in\Gamma} |P_n^\xi \gamma^2|\right) \leq 2K\mathbb{E}\left(\sup_{\gamma\in\Gamma} |P_n^\xi \gamma|\right).$$

Lemma. Suppose $\|\gamma\|_\infty \leq K$ for all $\gamma \in \Gamma$. Let H be the convex conjugate of $G_\psi = G \circ \psi^{-1}$, with $G(u) = u^2$, $u > 0$. Then for all

$$\sigma^2 \geq \frac{\delta}{1-\delta} H\left(\frac{2K}{\sqrt{n}\delta}\right),$$

we have

$$\mathbb{E}\psi(\hat{\sigma}) \leq \psi\left(\sigma\sqrt{2/1-\delta}\right).$$

Proof. $\cdots$

Some references

Bartlett, P.L., Bousquet, O. and Mendelson, S., *Local Rademacher complexities*, Ann. Statist. **33** (2005), 1497–1537.

Bousquet, O., A Bennett concentration inequality and its application to suprema of empirical processes, C.R. Acad. Sci. Paris **3**34 (2002), 495–500.

E. del Barrio, P. Deheuvels and S.A. van de Geer (2007). *Oracle i nequalities and regularization.* EMS Series of Lectures in Mathematics

Koltchinskii, V., 2004 IMS Medallion Lecture: Local Rademacher complexities and oracle inequalities in risk minimization, Ann. Statist. **3**4 (2006).

MASSART, P. (2000). Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse* **9** 245–303.

S. van de Geer (2000). *Empirical Process in M-Estimation*. Cambridge University Press.

S. van de Geer (2008). High-dimensional generalized linear models and the Lasso. Ann. Statist.

A. van der Vaart and J. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer.