Empirical Processes in M-Estimation

by

Sara van de Geer

Handout at **New Directions in General Equilibrium Analysis**

Cowles Workshop, Yale University

June 15-20, 2003

Version: June 13, 2003

Most of the first part can be found in van de Geer (2000) *Empirical Processes in M-Estimation*, Cambridge University Press (see also the references therein). The second part also contains more recent work.

Contents

Part 1
Empirical processes and asymptotic normality of M-estimators

Part 2
Empirical processes and regularization of M-estimators

Part 1
Empirical processes and asymptotic normality of M-estimators

1. Introduction. Let $X_1, \ldots, X_n, \ldots$ be i.i.d. copies of a random variable X with values in $\mathcal{X}$ and with distribution P.

1.1. Law of large numbers for real-valued random variables. Consider the case $\mathcal{X} = \mathbf{R}$. Suppose the mean

$$\mu = EX$$

exists. Define the average

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \ n \geq 1.$$

Then, by the law of large numbers, as $n \to \infty$,

$$\bar{X}_n \to \mu, \text{ a.s.}$$

Now, let

$$F(t) = P(X \leq t), \ t \in \mathbf{R},$$

be the theoretical distribution function, and

$$F_n(t) = \frac{1}{n}\#\{X_i \leq t, \ 1 \leq i \leq n\}, \ t \in \mathbf{R},$$

be the empirical distribution function. Then by the law of large numbers, as $n \to \infty$,

$$F_n(t) \to F(t), \text{ a.s. for all } t.$$

The Glivenko-Cantelli Theorem says that

$$\sup_t |F_n(t) - F(t)| \to 0, \text{ a.s.}$$

This is a **uniform** law of large numbers.

Application: *Kolmogorov's goodness-of-fit test.* We want to test

$H_0 : \ F = F_0$.

Test statistic:

$$D_n = \sup_t |F_n(t) - F_0(t)|.$$

Reject H_0 for large values of D_n.

1.2. $\mathbf{R}^d$-valued random variables. Questions:

(i) What is a natural extension of half-intervals in $\mathbf{R}$ to higher dimensions?

(ii) Does Glivenko-Cantelli hold for this extension?

1.3. Definition Glivenko-Cantelli classes of sets. Let for any (measurable[1]) $A \subset \mathcal{X}$,

$$P_n(A) = \frac{1}{n}\#\{X_i \in A,\ 1 \le i \le n\}.$$

We call P_n the empirical measure (based on $X_1, \ldots, X_n$).

Let $\mathcal{D}$ be a collection of subsets of $\mathcal{X}$.

Definition 1.3.1. The collection $\mathcal{D}$ is called a **Glivenko-Cantelli** (GC) class if

$$\sup_{D \in \mathcal{D}} |P_n(D) - P(D)| \to 0, \text{ a.s.}$$

Example. Let $\mathcal{X} = \mathbf{R}$. The class of half-intervals

$$\mathcal{D} = \{1_{(-\infty,t]} :\ t \in \mathbf{R}\}$$

is GC. But when e.g. $P =$ uniform distribution on $[0,1]$ (i.e., $F(t) = t$, $0 \le t \le 1$), the class

$$\mathcal{B} = \{\text{all (Borel) subsets of } [0,1]\}$$

is **not** GC.

2. Which classes are Glivenko-Cantelli classes?

2.1. General GC classes. Let $\mathcal{D}$ be a collection of subsets of $\mathcal{X}$, and let $\{\xi_1, \ldots, \xi_n\}$ be n points in $\mathcal{X}$.

Definition 2.1.1. We write

$$\triangle^{\mathcal{D}}(\xi_1, \ldots, \xi_n) = \text{card}(\{D \cap \{\xi_1, \ldots, \xi_n\} :\ D \in \mathcal{D}\}$$

$$= \text{the number of subsets of } \{\xi_1, \ldots, \xi_n\} \text{ that } \mathcal{D} \text{ can distinguish.}$$

That is, count the number of sets in $\mathcal{D}$, when two sets D_1 and D_2 are considered as equal if $D_1 \triangle D_2 \cap \{\xi_1, \ldots, \xi_n\} = \emptyset$. Here

$$D_1 \triangle D_2 = (D_1 \cap D_2^c) \cup (D_1^c \cap D_2)$$

[1] We will skip measurability issues, and most of the time do not mention explicitly the requirement of measurability of certain sets or functions. This means that everything has to be understood *modulo* measurability.

is the symmetric difference between D_1 and D_2.

Remark. For our purposes, we will not need to calculate $\triangle^{\mathcal{D}}(\xi_1, \ldots, \xi_n)$ **exactly**, but only a good enough upper bound.

Example. Let $\mathcal{X} = \mathbf{R}$ and

$$\mathcal{D} = \{\mathrm{l}_{(-\infty,t]} : \ t \in \mathbf{R}\}.$$

Then for all $\{\xi_1, \ldots, \xi_n\} \subset \mathbf{R}$

$$\triangle^{\mathcal{D}}(\xi_1, \ldots, \xi_n) \leq n + 1.$$

Example. Let $\mathcal{D}$ be the collection of all finite subsets of $\mathcal{X}$. Then, if the points $\xi_1, \ldots, \xi_n$ are distinct,

$$\triangle^{\mathcal{D}}(\xi_1, \ldots, \xi_n) = 2^n.$$

Notation. The simultaneous distribution of $(X_1, \ldots, X_n, \ldots)$ is denoted by

$$\mathbf{P} = P \times \ldots \times P \times \ldots.$$

Theorem 2.1.2. (Vapnik and Chervonenkis (1971)). *We have*

$$\sup_{D \in \mathcal{D}} |P_n(D) - P(D)| \to 0 \text{ a.s.}$$

if and only if

$$\frac{1}{n} \log \Delta^{\mathcal{D}}(X_1, \ldots, X_n) \to^{\mathbf{P}} 0.$$

□

2.2. Vapnik-Chervonenkis classes.

Definition 2.2.1. Let

$$m^{\mathcal{D}}(n) = \sup\{\triangle^{\mathcal{D}}(\xi_1, \ldots, \xi_n) : \ \xi_1, \ldots, \xi_n \in \mathcal{X}\}.$$

We say that $\mathcal{D}$ is a **Vapnik-Chervonenkis** (VC) class if for certain constants c and r, and for all n,

$$m^{\mathcal{D}}(n) \leq cn^r,$$

i.e., if $m^{\mathcal{D}}(n)$ does not grow faster than a polynomial in n.

Important conclusion: For sets, VC $\Rightarrow$ GC.

Examples.

a) $\mathcal{X} = \mathbf{R}$, $\mathcal{D} = \{1_{(-\infty,t]} : \ t \in \mathbf{R}\}$. Since $m^{\mathcal{D}}(n) \leq n+1$, $\mathcal{D}$ is VC.

b) $\mathcal{X} = \mathbf{R}^d$, $\mathcal{D} = \{1_{(-\infty,t]} : \ t \in \mathbf{R}^d\}$. Since $m^{\mathcal{D}}(n) \leq (n+1)^d$, $\mathcal{D}$ is VC.

c) $\mathcal{X} = \mathbf{R}^d$, $\mathcal{D} = \{\{x : \ \theta^T x > t\}, \begin{pmatrix} \theta \\ t \end{pmatrix} \in \mathbf{R}^{d+1}\}$. Since $m^{\mathcal{D}}(n) \leq 2^d \binom{n}{d}$, $\mathcal{D}$ is VC.

The VC property is closed under measure theoretic operations:

Lemma 2.2.2. *Let $\mathcal{D}$, $\mathcal{D}_1$ and $\mathcal{D}_2$ be VC. Then the following classes are also VC:*

(i) $\mathcal{D}^c = \{D^c : D \in \mathcal{D}\}$,

(ii) $\mathcal{D}_1 \cap \mathcal{D}_2 = \{D_1 \cap D_2 : \ D_1 \in \mathcal{D}_1, \ D_2 \in \mathcal{D}_2\}$,

(iii) $\mathcal{D}_1 \cup \mathcal{D}_2 = \{D_1 \cap D_2 : \ D_1 \in \mathcal{D}_1, \ D_2 \in \mathcal{D}_2\}$. □

Examples.

- the class of intersections of two halfspaces,

- all ellipsoids,

- all half-ellipsoids,

- in $\mathbf{R}$, the class $\{\{x : \ \theta_1 x + \ldots + \theta_r x^r \leq t\} : \ \begin{pmatrix} \theta \\ t \end{pmatrix} \in \mathbf{R}^{r+1}\}$.

There are classes that are GC, but not VC.

Example. Let $\mathcal{X} = [0,1]^2$, and let $\mathcal{D}$ be the collection of all convex subsets of $\mathcal{X}$. Then $\mathcal{D}$ is not VC, but when P is uniform, $\mathcal{D}$ **is** GC.

Finally, we will not deny you the following definition and a very nice lemma.

Definition 2.2.3. The VC dimension of $\mathcal{D}$ is

$$V(\mathcal{D}) = \inf\{n : \ m^{\mathcal{D}}(n) < 2^n\}.$$

Lemma 2.2.4. *We have that $\mathcal{D}$ is VC if and only if $V(\mathcal{D}) < \infty$.* □

3. Convergence of means to expectations.

Notation. For a function $g : \mathcal{X} \to \mathbf{R}$, we write

$$\int g dP = Eg(X),$$

and

$$\int g dP_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

3.1. Uniform law of large numbers for classes of functions. Let $\mathcal{G}$ be a collection of real-valued functions on $\mathcal{X}$.

Definition 3.1.1. The class $\mathcal{G}$ is called a **Glivenko-Cantelli** (GC) class if

$$\sup_{g \in \mathcal{G}} | \int g dP_n - \int g dP | \to 0, \text{ a.s.}$$

Example. $\mathcal{G} = \{1_D : D \in \mathcal{D}\}$ is GC if $\mathcal{D}$ is GC.

3.2. VC classes of functions.

Definition 3.2.1. The **subgraph** of a function $g : \mathcal{X} \to \mathbf{R}$ is

$$\text{subgraph}(g) = \{(x, t) \in \mathcal{X} \times \mathbf{R} : \ g(x) \geq t\}.$$

A collection of functions $\mathcal{G}$ is called a VC class if the subgraphs $\{\text{subgraph}(g) : \ g \in \mathcal{G}\}$ form a VC class.

Examples ($\mathcal{X} = \mathbf{R}^d$).

a) $\mathcal{G} = \{g(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_d x_d : \ \theta \in \mathbf{R}^{d+1}\}$,

b) $\mathcal{G} = \{g(x) = |\theta_0 + \theta_1 x_1 + \ldots + \theta_d x_d| : \ \theta \in \mathbf{R}^{d+1}\}$.

c) $d = 1$, $\mathcal{G} = \left\{ g(x) = \begin{cases} a + bx & \text{if } x \leq c \\ d + ex & \text{if } x > c \end{cases} , \ \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} \in \mathbf{R}^5 \right\}$,

d) $d = 1$, $\mathcal{G} = \{g(x) = \mathrm{e}^{\theta x} : \ \theta \in \mathbf{R}\}$.

Definition 3.2.2. The **envelope** G of a collection of functions $\mathcal{G}$ is defined by

$$G(x) = \sup_{g \in \mathcal{G}} |g(x)|, \ x \in \mathcal{X}.$$

Theorem 3.2.3. *Suppose $\mathcal{G}$ is VC and that $\int G dP < \infty$. Then $\mathcal{G}$ is GC.* □

3.3. Exercises.

Exercise 1

Are the following classes of sets (functions) VC? Why (not)? Which classes are GC for all P?

1) The class of all rectangles in $\mathbf{R}^d$.

2) The class of all monotone functions on $\mathbf{R}$.

3) The class of functions on $[0,1]$ given by

$$\mathcal{G} = \{g(x) = a\mathrm{e}^{bx} + c\mathrm{e}^{dx} : \ (a,b,c,d) \in [0,1]^4\}.$$

4) The class of all sections in $\mathbf{R}^2$ (a section is of the form $\{(x_1,x_2) : \ x_1 = a_1 + r\sin t, x_2 = a_2 + r\cos t, \ \theta_1 \leq t \leq \theta_2\}$, for some $(a_1,a_2) \in \mathbf{R}^2$ and some $0 \leq \theta_1 \leq \theta_2 \leq 2\pi$).

5) The class of all star-shaped sets in $\mathbf{R}^2$ (a set D is star-shaped if for some $a \in D$ and all $b \in D$ also all points on the line segment joining a and b are in D).

Exercise 2

Let $\mathcal{G}$ be the class of all functions g on $[0,1]$ with derivative $\dot{g}$ satisfying $|\dot{g}| \leq 1$. Check that $\mathcal{G}$ is not VC. Show that $\mathcal{G}$ is GC by using partial integration and the Glivenko-Cantelli Theorem for the empirical distribution function.

4. Uniform central limit theorems.

4.1. Real-valued random variables.

Let $\mathcal{X} = \mathbf{R}$.

Central limit theorem in $\mathbf{R}$. *Suppose $EX = \mu$, and $\mathrm{var}(X) = \sigma^2$ exist. Then*

$$\mathbf{P}\left(\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \leq z\right) \to \Phi(z), \ \textit{for all } z,$$

where Φ is the standard normal distribution function. □.

Notation.

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \to^{\mathcal{L}} \mathcal{N}(0,1),$$

or

$$\sqrt{n}(\bar{X}_n - \mu) \to^{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

4.2. $\mathbf{R}^d$-valued random variables. Let $X_1, X_2, \ldots$ be i.i.d. $\mathbf{R}^d$-valued random variables, copies of X ($X \in \mathcal{X} = \mathbf{R}^d$), with expectation $\mu = EX$, and covariance matrix $\Sigma = EXX^T - \mu\mu^T$.

Central limit theorem in $\mathbf{R}^d$. *We have*

$$\sqrt{n}(\bar{X}_n - \mu) \to^{\mathcal{L}} \mathcal{N}(0, \Sigma),$$

i.e.

$$\sqrt{n}\left[a^T(\bar{X}_n - \mu)\right] \to^{\mathcal{L}} \mathcal{N}(0, a^T\Sigma a), \ \textit{for all } a \in \mathbf{R}^d.$$

□.

4.3. Donsker's Theorem. Let $\mathcal{X} = \mathbf{R}$. Recall the definition of the distribution function F and the empirical distribution function F_n:

$$F(t) = P(X_1 \le t), \ t \in \mathbf{R},$$

$$F_n(t) = \frac{1}{n}\#\{X_i \le t, \ 1 \le i \le n\}, \ t \in \mathbf{R}.$$

By the central limit theorem in $\mathbf{R}$ (Section 4.1), for all t

$$\sqrt{n}(F_n(t) - F(t)) \to^{\mathcal{L}} \mathcal{N}\left(0, F(t)(1 - F(t))\right).$$

Also, by the central limit theorem in $\mathbf{R}^2$ (Section 4.2), for all $s < t$,

$$\sqrt{n}\begin{pmatrix} F_n(s) - F(s) \\ F_n(t) - F(t) \end{pmatrix} \to^{\mathcal{L}} \mathcal{N}(0, \Sigma(s,t)),$$

where

$$\Sigma(s,t) = \begin{pmatrix} F(s)(1-F(s)) & F(s)(1-F(t)) \\ F(s)(1-F(t)) & F(t)(1-F(t)) \end{pmatrix}.$$

We are now going to consider the **stochastic process** $W_n = \{W_n(t) : \ t \in \mathbf{R}\}$. The process W_n is called the (classical) empirical process.

Definition 4.3.1. Let $\mathcal{K}_0$ be the collection of bounded functions on $[0,1]$. The stochastic process $B(\cdot) \in \mathcal{K}_0$, is called the standard **Brownian bridge** if

- $B(0) = B(1) = 0$,

- for all $r \ge 1$ and all $t_1, \ldots, t_r \in (0,1)$, the vector $\begin{pmatrix} B(t_1) \\ \vdots \\ B(t_r) \end{pmatrix}$ is multivariate normal with mean zero,

- for all $s \leq t$, $\mathrm{cov}(B(s), B(t)) = s(1-t)$.

- the sample paths of B are a.s. continuous.

Donsker's theorem. *Consider W_n and $W_F = B \circ F$ as elements of the space $\mathcal{K}$ of bounded functions on $\mathcal{R}$. We have*

$$W_n \to^{\mathcal{L}} W_F,$$

that is,

$$\mathbf{E}f(W_n) \to \mathbf{E}f(W_F),$$

for all continuous and bounded functions f. □

Reflection. Suppose F is continuous. Then, since B is almost surely continuous, also $W_F = B \circ F$ is almost surely continuous. So W_n must be approximately continuous as well in some sense. Indeed, we have for any t and any sequence t_n converging to t,

$$|W_n(t_n) - W_n(t)| \to^{\mathbf{P}} 0.$$

This is called **asymptotic continuity**.

4.4. Donsker classes. Let $X_1, \ldots, X_n, \ldots$ be i.i.d. copies of a random variable X, with values in the space $\mathcal{X}$, and with distribution P. Consider a class $\mathcal{G}$ of functions $g : \mathcal{X} \to \mathbf{R}$. The (theoretical) mean of a function g is

$$\int g dP = Eg(X),$$

and the (empirical) average (based on the n observations $X_1, \ldots, X_n$) is

$$\int g dP_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

Here P_n is the empirical distribution (based on $X_1, \ldots, X_n$).

Definition 4.4.1. The **empirical process** indexed by $\mathcal{G}$ is

$$\nu_n(g) = \sqrt{n} \int g d(P_n - P), \ g \in \mathcal{G}.$$

Let us recall the central limit theorem for g fixed. Denote the variance of $g(X)$ by

$$\sigma^2(g) = \mathrm{var}(g(X)) = Eg^2(X) - (Eg(X))^2.$$

If $\sigma^2(g) < \infty$, we have

$$\nu_n(g) \to^{\mathcal{L}} \mathcal{N}(0, \sigma^2(g)).$$

The central limit theorem also holds for finitely many g simultaneously. Let g_k and g_l be two functions and denote the covariance between $g_k(X)$ and $g_l(X)$ by

$$\sigma(g_k, g_l) = \mathrm{cov}(g_k(X), g_l(X)) = Eg_k(X)g_l(X) - Eg_k(X)Eg_l(X).$$

Then, whenever $\sigma^2(g_k) < \infty$ for $k = 1, \ldots, r$,

$$\begin{pmatrix} \nu_n(g_1) \\ \vdots \\ \nu_n(g_r) \end{pmatrix} \to^{\mathcal{L}} \mathcal{N}(0, \Sigma_{g_1,\ldots,g_r}),$$

where $\Sigma_{g_1,\ldots,g_r}$ is the variance-covariance matrix

$$(*) \qquad \Sigma_{g_1,\ldots,g_r} = \begin{pmatrix} \sigma^2(g_1) & \ldots & \sigma(g_1, g_r) \\ \vdots & \ddots & \vdots \\ \sigma(g_1, g_r) & \ldots & \sigma^2(g_r) \end{pmatrix}.$$

Definition 4.4.2. Let ν be a Gaussian process indexed by $\mathcal{G}$. Assume that for each $r \in \mathbf{N}$ and for each finite collection $\{g_1, \ldots, g_r\} \subset \mathcal{G}$, the r-dimensional vector

$$\begin{pmatrix} \nu(g_1) \\ \vdots \\ \nu(g_r) \end{pmatrix}$$

has a $\mathcal{N}(0, \Sigma_{g_1,\ldots,g_r})$-distribution, with $\Sigma_{g_1,\ldots,g_r}$ defined in (*). We then call ν the **P-Brownian bridge** indexed by $\mathcal{G}$.

Definition 4.4.3. Consider ν_n and ν as bounded functions on $\mathcal{G}$. We call $\mathcal{G}$ a **P-Donsker class** if

$$\nu_n \to^{\mathcal{L}} \nu,$$

that is, if for all continuous and bounded functions f, we have

$$\mathbf{E}f(\nu_n) \to \mathbf{E}f(\nu).$$

Definition 4.4.4. The process ν_n on $\mathcal{G}$ is called **asymptotically continuous** if for all $g_0 \in \mathcal{G}$, and all (possibly random) sequences $\{g_n\} \subset \mathcal{G}$ with $\sigma(g_n - g_0) \to^{\mathbf{P}} 0$, we have

$$|\nu_n(g_n) - \nu_n(g_0)| \to^{\mathbf{P}} 0.$$

We will use the notation

$$\|g\|^2 = \int g^2 dP,$$

i.e., $\|\cdot\|$ is the $L_2(P)$-norm.

Remark. Note that $\sigma(g) \leq \|g\|$.

Definition 4.4.5. The class $\mathcal{G} \subset L_2(P)$ is called **totally bounded** if for all $\delta > 0$ the number of balls with radius δ necessary to cover $\mathcal{G}$ is finite.

Theorem 4.4.6. *Suppose that $\mathcal{G}$ is totally bounded. Then $\mathcal{G}$ is a P-Donsker class if and only if ν_n (as process on $\mathcal{G}$) is asymptotically continuous.*

□

Theorem 4.4.7. *Suppose that $\mathcal{G}$ is a VC-graph class with envelope*

$$G = \sup_{g \in \mathcal{G}} |g|$$

satisfying $\int G^2 dP < \infty$. Then $\mathcal{G}$ is P-Donsker.

□

Remark. Thus, a VC class $\mathcal{G}$ with square integrable envelope G is asymptotically continuous. In particular, suppose that such a class $\mathcal{G}$ is parametrized by θ in some parameter space $\Theta \subset \mathbf{R}^r$ (say), i.e. $\mathcal{G} = \{g_\theta : \theta \in \Theta\}$. Let $z_n(\theta) = \nu_n(g_\theta)$. Question: do we have that for a (random) sequence θ_n with $\theta_n \to \theta_0$ (in probability), also

$$|z_n(\theta_n) - z_n(\theta_0)| \to^{\mathbf{P}} 0 \ ?$$

Indeed, if $\sigma(g_\theta - g_{\theta_0})$ converges to zero as θ converges to θ_0, the answer is yes. And so $\|g_\theta - g_{\theta_0}\| \to 0$ (mean square convergence) suffices for a yes answer.

5. M-estimators.

5.1. What is an M-estimator? Let $X_1, \ldots, X_n, \ldots$ be i.i.d. copies of a random variable X with values in $\mathcal{X}$ and with distribution P.

Let Θ be a parameter space (a subset of some metric space) and let

$$\gamma_\theta : \mathcal{X} \to \mathbf{R},$$

be some loss function. We estimate the unknown parameter

$$\theta_0 = \arg\min_{\theta \in \Theta} \int \gamma_\theta dP,$$

by the M-estimator

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \int \gamma_\theta dP_n.$$

Here, we assume that θ_0 exists and is unique and that $\hat{\theta}_n$ exists.

Examples.

(i) Location estimators. $\mathcal{X} = \mathbf{R}$, $\Theta = \mathbf{R}$, and

(i.a) $\gamma_\theta(x) = (x-\theta)^2$ (estimating the mean),

(i.b) $\gamma_\theta(x) = |x-\theta|$ (estimating the median).

(ii) Maximum likelihood. Let $\{p_\theta : \theta \in \Theta\}$ be a family of densities w.r.t. a σ-finite dominating measure μ, and

$$\rho_\theta = -\log p_\theta.$$

If $dP/d\mu = p_{\theta_0}$, $\theta_0 \in \Theta$, then indeed θ_0 is a minimizer of $\int \rho_\theta dP$, $\theta \in \Theta$.

(ii.a) Poisson distribution:

$$p_\theta(x) = \mathrm{e}^\theta \frac{\theta^x}{x!},\ x \in \{1, 2, \ldots\},\ \theta > 0.$$

(ii.b) Logistic distribution:

$$p_\theta(x) = \frac{\mathrm{e}^{\theta - x}}{(1+\mathrm{e}^{\theta-x})^2},\ x \in \mathbf{R},\ \theta \in \mathbf{R}.$$

5.2. Consistency and the law of large numbers. Define for $\theta \in \Theta$,

$$\Gamma(\theta) = \int \gamma_\theta dP,$$

and

$$\Gamma_n(\theta) = \int \gamma_\theta dP_n.$$

We first present an easy proposition with a too stringent condition ($\bullet$).

Proposition 5.2.1. *Suppose that $\theta \mapsto \Gamma(\theta)$ is continuous. Assume moreover that*

$$(\bullet) \qquad \sup_{\theta \in \Theta} |\Gamma_n(\theta) - \Gamma(\theta)| \to 0,\ a.s.,$$

i.e., that $\{\gamma_\theta :\ \theta \in \Theta\}$ is a GC class. Then $\hat{\theta}_n \to \theta_0$ a.s. □

The assumption ($\bullet$) is hardly ever met, because it is close to requiring compactness of Θ. We give a lemma, which replaces ($\bullet$) by a convexity assumption. This works out well when Θ is finite dimensional.

Lemma 5.2.2. *Suppose that Θ is a convex subset of $\mathbf{R}^r$, and that $\theta \mapsto \gamma_\theta$, $\theta \in \Theta$ is convex. Then $\hat{\theta}_n \to \theta_0$, a.s.* □

5.3. Asymptotic normality of M-estimators. In this section, we assume that we already showed that $\hat{\theta}_n$ is consistent, and that θ_0 is an interior point of $\Theta \subset \mathbf{R}^r$.

Definition 5.3.1. The (sequence of) estimator(s) $\hat{\theta}_n$ of θ_0 is called **asymptotically linear** if we may write

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n} \int l dP_n + o_{\mathbf{P}}(1),$$

where

$$l = \begin{pmatrix} l_1 \\ \vdots \\ l_r \end{pmatrix} : \mathcal{X} \to \mathbf{R}^r,$$

satisfies $\int l dP = 0$ and $\int l_k^2 dP < \infty$, $k = 1, \ldots, r$. The function l is then called the **influence function**. For the case $r = 1$, we call $\sigma^2 = \int l^2 dP$ the **asymptotic variance**.

Definition 5.3.2. Let $\hat{\theta}_{n,1}$ and $\hat{\theta}_{n,2}$ be two asymptotically linear estimators of θ_0, with asymptotic variances σ_1^2 and σ_2^2 respectively. Then

$$e_{1,2} = \frac{\sigma_2^2}{\sigma_1^2}$$

is called the **asymptotic relative efficiency** (of $\hat{\theta}_{n,1}$ as compared to $\hat{\theta}_{n,2}$).

5.4. Conditions a, b and c for asymptotic normality. We start with three conditions a, b and c, which are easier to check but more stringent. We later relax them to conditions A, B and C.

Condition a. There exists an $\epsilon > 0$ such that $\theta \mapsto \gamma_\theta$ is differentiable for all $|\theta - \theta_0| < \epsilon$ and all x, with derivative

$$\psi_\theta(x) = \frac{\partial}{\partial \theta} \gamma_\theta(x), \ x \in \mathcal{X}.$$

Condition b. We have as $\theta \to \theta_0$,

$$\int (\psi_\theta - \psi_{\theta_0}) dP = V(\theta - \theta_0) + o(1)|\theta - \theta_0|,$$

where V is a positive definite matrix.

Condition c. There exists an $\epsilon > 0$ such that the class

$$\{\psi_\theta : \ |\theta - \theta_0| < \epsilon\}$$

has envelope $\Psi \in L_2(P)$ and is Donsker. Moreover,

$$\lim_{\theta \to \theta_0} \|\psi_\theta - \psi_{\theta_0}\| = 0.$$

Lemma 5.4.1. *Suppose conditions a, b and c. Then $\hat{\theta}_n$ is asymptotically linear with influence function*

$$l = -V^{-1}\psi_{\theta_0},$$

so

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to^{\mathcal{L}} \mathcal{N}(0, V^{-1}JV^{-1}),$$

where

$$J = \int \psi_{\theta_0}\psi_{\theta_0}^T dP.$$

□

Example: Huber estimator. Let $\mathcal{X} = \mathbf{R}$, $\Theta = \mathbf{R}$. The Huber estimator corresponds to the loss function

$$\gamma_\theta(x) = \gamma(x - \theta),$$

with

$$\gamma(x) = x^2 1\{|x - \theta| \leq k\} + (2k|x| - k^2)1\{|x| > k\}, \ x \in \mathbf{R}.$$

Here, $0 < k < \infty$ is some fixed constant, chosen by the statistician. We will now verify a, b and c.

a)

$$\psi_\theta(x) = \begin{cases} 2k & \text{if } x - \theta \leq = k \\ -2(x - \theta) & \text{if } |x - \theta| \leq k \\ -2k & \text{if } x - \theta \geq k \end{cases}.$$

b) We have

$$\frac{d}{d\theta}\int \psi_\theta dP = 2(F(k + \theta) - F(-k + \theta)),$$

where $F(t) = P(X \leq t)$, $t \in \mathbf{R}$ is the distribution function. So

$$V = 2(F(k + \theta_0) - F(-k + \theta_0)).$$

c) Clearly $\psi_\theta : \theta \in \mathbf{R}$ is a VC class, with envelope $\Psi \leq 2k$.

So the Huber estimator $\hat{\theta}_n$ is asymptotically linear, with influence function

$$l(x) = \begin{cases} -\frac{k}{F(k+\theta_0)-F(-k+\theta_0)} & \text{if } x - \theta_0 \leq -k \\ \frac{x-\theta_0}{F(k+\theta_0)-F(-k+\theta_0)} & \text{if } |x - \theta_0| \leq k \\ \frac{k}{F(k+\theta_0)-F(-k+\theta_0)} & \text{if } x - \theta_0 > k \end{cases}.$$

The asymptotic variance is

$$\sigma^2 = \frac{k^2 F(-k + \theta_0) = \int_{-k+\theta_0}^{k+\theta_0}(x - \theta_0)^2 dF(x) + k^2(1 - F(k + \theta_0))}{(F(k + \theta_0) - F(k + \theta_0))^2}.$$

5.5. Asymptotics for the median. The median (see Example (i.b)) can be regarded as the limiting case of a Huber estimator, with $k \downarrow 0$. However, the loss function $\gamma_\theta(x) = |x - \theta|$ is not

differentiable, i.e., does not satisfy condition a. We give here a direct evaluation of the median. This serves as a preparation to relaxing a, b and c, to A, B and C.

Let $X \in \mathbf{R}$ have distribution F, and let F_n be the empirical distribution. The population median θ_0 is a solution of the equation

$$F(\theta_0) = 0.$$

We assume this solution exists and also that F has positive density f in a neighborhood of θ_0. We consider now for simplicity only even sample sizes n and let the sample median $\hat{\theta}_n$ be any solution of

$$F_n(\hat{\theta}_n) = 0.$$

Then we get

$$\begin{aligned} 0 &= F_n(\hat{\theta}_n) - F(\theta_0) \\ &= \left[F_n(\hat{\theta}_n) - F(\hat{\theta}_n)\right] + \left[F(\hat{\theta}_n) - F(\theta_0)\right] \\ &= \frac{1}{\sqrt{n}} W_n(\hat{\theta}_n) + \left[F(\hat{\theta}_n) - F(\theta_0)\right], \end{aligned}$$

where $W_n = \sqrt{n}(F_n - F)$ is the empirical process. Since F is continuous at θ_0, and $\hat{\theta}_n \to \theta_0$, we have by the asymptotic continuity of the empirical process (Section 4.3), that $W_n(\hat{\theta}_n) = W_n(\theta_0) + o_{\mathbf{P}}(1)$. We thus arrive at

$$\begin{aligned} 0 &= W_n(\theta_0) + \sqrt{n}\left[F(\hat{\theta}_n) - F(\theta_0)\right] + o_{\mathbf{P}}(1) \\ &= W_n(\theta_0) + \sqrt{n}[f(\theta_0) + o(1)][\hat{\theta}_n - \theta_0]. \end{aligned}$$

In other words,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{W_n(\theta_0)}{f(\theta_0)} + o_{\mathbf{P}}(1).$$

So the influence function is

$$l(x) = \begin{cases} -\frac{1}{f(\theta_0)} & \text{if } x \leq \theta_0 \\ +\frac{1}{f(\theta_0)} & \text{if } x > \theta_0 \end{cases},$$

and the asymptotic variance is

$$\sigma^2 = \frac{1}{4f(\theta_0)^2}.$$

We can now compare median and mean. It is easily seen that the asymptotic relative efficiency of the mean as compared to the median is

$$\mathrm{e}_{1,2} = \frac{1}{4\sigma_0^2 f(\theta_0)^2},$$

where $\sigma_0^2 = \mathrm{var}(X)$. So $\mathrm{e}_{1,2} = \pi/2$ for the normal distribution, and $\mathrm{e}_{1,2} = 1/2$ for the double exponential (Laplace) distribution. The density of the double exponential distribution is

$$f(x) = \frac{1}{\sqrt{2\sigma_0^2}} \exp\left[-\frac{\sqrt{2}|x - \theta_0|}{\sigma_0}\right], \ x \in \mathbf{R}.$$

Exercise. Suppose X has the logistic distribution with location parameter θ (see Example (ii.b)). Show that the maximum likelihood estimator has asymptotic variance equal to 3, and the median has asymptotic variance equal to 4. Hence, the asymptotic relative efficiency of the maximum likelihood estimator as compared to the median is 4/3.

5.6. Conditions A, B and C for asymptotic normality. In this section, we again assume that we already showed that $\hat{\theta}_n$ is consistent, and that θ_0 is an interior point of $\Theta \subset \mathbf{R}^r$. We are now going to relax the condition of differentiability of γ_θ.

Condition A. (Differentiability in quadratic mean.) There exists a function $\psi_0 : \mathcal{X} \to \mathbf{R}^r$, with components in $L_2(P)$, such that

$$\lim_{\theta \to \theta_0} \frac{\left\| \gamma_\theta - \gamma_{\theta_0} - (\theta - \theta_0)^T \psi_0 \right\|}{|\theta - \theta_0|} = 0.$$

Condition B. We have as $\theta \to \theta_0$,

$$\Gamma(\theta) - \Gamma(\theta_0) = \frac{1}{2}(\theta - \theta_0)^T V (\theta - \theta_0) + o(1)|\theta - \theta_0|^2,$$

with V a positive definite matrix.

Condition C. Define for $\theta \neq \theta_0$,

$$g_\theta = \frac{(\gamma_\theta - \gamma_{\theta_0})}{|\theta - \theta_0|}.$$

Suppose that for some $\epsilon > 0$, the class $\{g_\theta : \ 0 < |\theta - \theta_0| < \epsilon\}$ has envelope $G \in L_2(P)$ and that it is a Donsker class.

Lemma 5.6.1. *Suppose conditions A, B and C are met. Then $\hat{\theta}_n$ has influence function*

$$l = -V^{-1}\psi_0,$$

and so

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to^{\mathcal{L}} \mathcal{N}(0, V^{-1} J V^{-1}),$$

where $J = \int \psi_0 \psi_0^T dP$. □

5.7. Exercise. Let (X_i, Y_i), $i = 1, \ldots, n, \ldots$ be i.i.d. copies of (X, Y), where $X \in \mathbf{R}^d$ and $Y \in \mathbf{R}$. Suppose that the conditional distribution of Y given $X = x$ has median $m(x) = \alpha_0 + \alpha_1 x_1 + \ldots \alpha_d x_d$, with

$$\alpha = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_d \end{pmatrix} \in \mathbf{R}^{d+1}.$$

Assume moreover that given $X = x$, the random variable $Y - m(x)$ has a density f not depending on x, with f positive in a neighborhood of zero. Suppose moreover that

$$\Sigma = E \begin{pmatrix} 1 & X \\ X & XX^T \end{pmatrix}$$

exists. Let

$$\hat{\alpha}_n = \arg \min_{a \in \mathbf{R}^{d+1}} \frac{1}{n} \sum_{i=1}^{n} |Y_i - a_0 - a_1 X_{i,1} - \ldots - a_d X_{i,d}| \,,$$

be the least absolute deviations (LAD) estimator. Show that

$$\sqrt{n}(\hat{\alpha}_n - \alpha) \to^{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{4f^2(0)} \Sigma^{-1}\right),$$

by verifying conditions A, B and C.

Part 2
Empirical processes and regularization of M-estimators

1. A classical approach.

$$Y_i = f_0(x_i) + \epsilon_i,\ i = 1, \ldots, n.$$

- $\epsilon_1, \ldots, \epsilon_n$ independent $\mathcal{N}(0, \sigma^2)$.
- $x_i = i/n$, $i = 1, \ldots, n$.
- $f_0 : [0, 1] \to \mathbf{R}$ unknown function.
- $Y_i \in \mathbf{R}$ observations.

"Classical" estimator:

$$\hat{f}_n = \arg\min_f \left\{ \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(x_i)|^2 + \lambda_n^2 \int_0^1 |f'(x)|^2 dx \right\}.$$

Continuous version:

$$\hat{f} = \arg\min_f \left\{ \int_0^1 |y(x) - f(x)|^2 dx + \lambda^2 \int_0^1 |f'(x)|^2 dx \right\}.$$

Lemma 1.1. *Solution:*

$$\hat{f}(x) = \frac{C}{\lambda} \cosh(\frac{x}{\lambda}) + \frac{1}{\lambda} \int_0^x y(u) \sinh(\frac{u - x}{\lambda}) du,$$

where

$$C = Y(1) - \left\{ \frac{1}{\lambda} \int_0^1 Y(u) \sinh(\frac{1 - u}{\lambda}) du \right\} / \sinh(\frac{1}{\lambda}),$$

with

$$Y(x) = \int_0^x y(u) du.$$

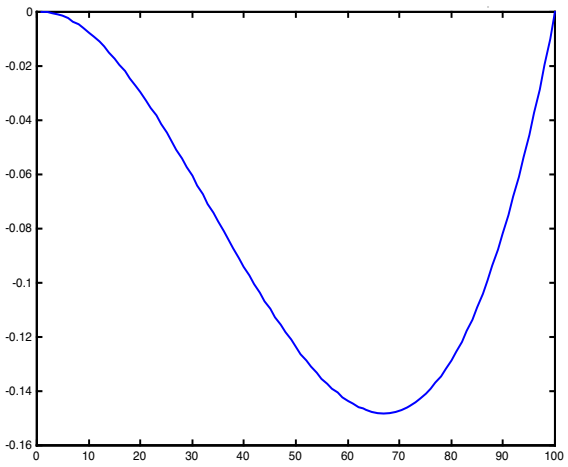
□

Choice of *regularization parameter* λ? E.g.,

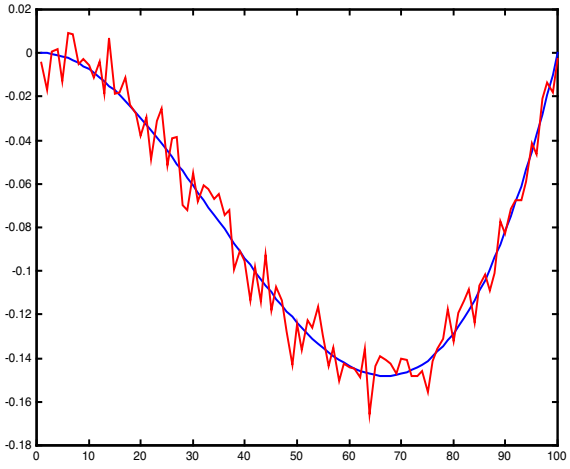
$$\hat{f}_n = \arg\min_f \min_\lambda \left\{ \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(x_i)|^2 + \left(\lambda^2 \int_0^1 |f'(x)|^2 dx + \frac{c}{n\lambda} \right) \right\}.$$

Here is the a MATLAB script for calculating the solution given in Lemma 1.1:

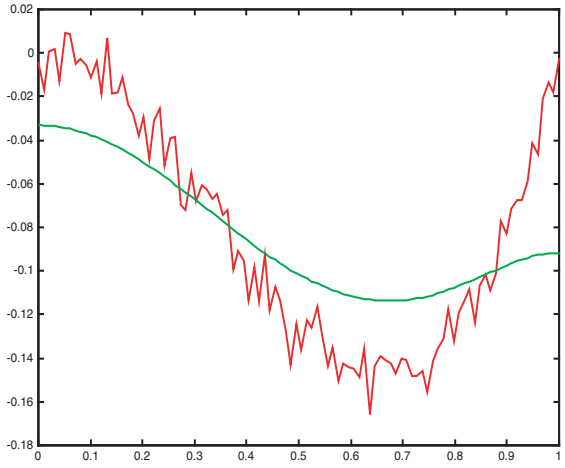
```
function smooth = smooth1(data,lambda)
% SMOOTH1 requires an input vector DATA and a regularization parameter LAMBDA
% SMOOTH1 calculates the smoothed version of the input vector DATA
% it uses a penalty on the squared L2 norm of the derivative of the function
% n=the number of observations
% the output FIT is the squared (normalized) L2 norm of the difference between
% DATA and fitted function

lambda
sigma=1/lambda;


n=size(data)*[1,1]'-1

dt = 1/(n-1);
x = [0:dt:1];


plot(x, data, 'r');
hold on;


size_x = size(x);
temp = zeros(size_x(2), 1);
temp2 = 0;

DATA = 0;
for j = 1:size_x(2)
   DATA = DATA + data(j) * dt;
end;

C = 0;
for j = 1:size_x(2)
   C = C + data(j) * (cosh(sigma * (-1 + j * dt)));
end;
C = (-DATA * (0) + C * dt) / sinh(sigma);

for i = 1:size_x(2)
   temp2 = 0;
   for j = 1:i
      temp2 =  temp2 + data(j) * sinh(sigma * (j - i) * dt);
   end;
   temp(i) = C * sigma * cosh(sigma * i * dt) + sigma * temp2 * dt;
end;

plot(x, temp, 'g');
fit=(data-temp')*(data'-temp)/n

hold off
```



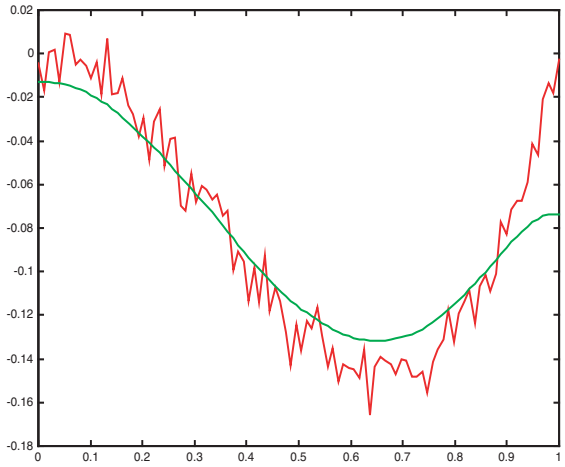
True *f*



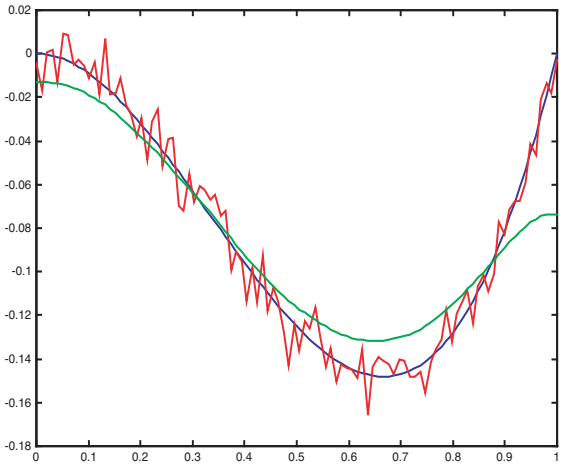
Noise added, noise level = 0.01



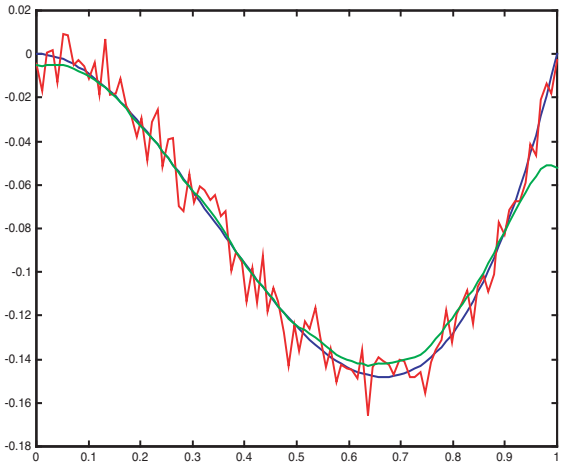
Denoised, lambda=0.2
Fit=9.0531e-04



Denoised, lambda=0.1
Fit=3.4322e-04



Denoised, lambda=0.1
Error=2.8119e-04



Denoised, lambda=0.05
Error=7.8683e-05

2. The sequence space model.

$$Y_i = f_0(x_i) + \epsilon_i,\ i = 1, \ldots, n.$$

with $\epsilon_1, \ldots, \epsilon_n$ independent $\mathcal{N}(0, \sigma^2)$, and $x_i \in \mathcal{X}$, $Y_i \in \mathbf{R}$, $i = 1, \ldots, n$.

Let

$$Q_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

Let $\psi_1, \ldots, \psi_n$ be an orthonormal basis of $L_2(Q_n)$ ($\psi_j : \mathcal{X} \to \mathbf{R}$, $j = 1, \ldots, n$). Define

$$\tilde{Y}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \psi_j(x_i),\ j = 1, \ldots, n,$$

$$\tilde{\epsilon}_j = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \psi_j(x_i),\ j = 1, \ldots, n,$$

$$\theta_j = \frac{1}{n} \sum_{i=1}^{n} f_0(x_i) \psi_j(x_i),\ j = 1, \ldots, n.$$

Then

$$\tilde{Y}_j = \theta_j + \tilde{\epsilon}_j,\ j = 1, \ldots, n,$$

with $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n$ independent $\mathcal{N}(0, \frac{\sigma^2}{n})$.

- Fourier
- Wavelets
- Trigonometric
- Polynomial
- ....

3. Sparse signals.

$$Y_j = \theta_j + \epsilon_j,\ j = 1, \ldots, n,$$

with $\epsilon_1, \ldots, \epsilon_n$ independent $\mathcal{N}(0, \frac{\sigma^2}{n})$,

Define

$$\|\vartheta\|_n^2 = \sum_{j=1}^n |\vartheta_j|^2,\ \vartheta \in \mathbf{R}^n.$$

Definition 3.1. *Let $0 \leq r < 2$. The signal θ is called* r-sparse *if*

$$\sum_{j=1}^n |\theta_j|^r \leq 1.$$

Lemma 3.2. *Suppose θ is r-sparse. Then*

$$\#\{|\theta_j| > \lambda\} \leq \lambda^{-r}.$$

□

Lemma 3.3. *Let*

$$\theta_j^* = \begin{cases} \theta_j & \text{if } |\theta_j| > \lambda \\ 0 & \text{if } |\theta_j| \leq \lambda \end{cases}.$$

If θ is r-sparse we have

$$\|\theta^* - \theta\|_n^2 \leq \lambda^{2-r}.$$

□

A sparse signal.....

4. A concentration inequality.

Lemma 4.1. *Let $\epsilon_1, \ldots, \epsilon_n$ be independent $\mathcal{N}(0,1)$. Then*

$$\max_{1 \le j \le n} |\epsilon_j| \le \sqrt{6 \log n}, \ a.s.,$$

for n sufficiently large.

Proof. We have for all $a > 0$, and $\gamma > 0$, by Chebyshev's inequality

$$\mathbf{P}(\epsilon_j > a) \le \frac{E \exp[\gamma \epsilon_j]}{\exp[\gamma a]}$$

$$= \exp[\frac{1}{2}\gamma^2 - \gamma a].$$

Take $\gamma = a$ to arrive at

$$\mathbf{P}(\epsilon_j > a) \le \exp[-\frac{1}{2}a^2].$$

So

$$\mathbf{P}\left(\max_{1 \le j \le n} |\epsilon_j| > a\right) \le 2n \exp[-\frac{1}{2}a^2].$$

Take $a = \sqrt{6 \log n}$ to get

$$\mathbf{P}\left(\max_{1 \le j \le n} |\epsilon_j| > \sqrt{6 \log n}\right) \le 2n \exp[-3 \log n] = 2 \exp[-2 \log n] = \frac{2}{n^2}.$$

Since

$$\sum_n \frac{2}{n^2} < \infty,$$

the result follows. □

Remark. The result can be improved to

$$\max_{1 \le j \le n} |\epsilon_j| \le \sqrt{2 \log n}, \ a.s.,$$

for n sufficiently large.

5. Hard and soft thresholding.

$$Y_j = \theta_j + \epsilon_j, \ j = 1, \ldots, n,$$

with $\epsilon_1, \ldots, \epsilon_n$ i.i.d. $\mathcal{N}(0, \frac{\sigma^2}{n})$.

Let $\lambda \ge 0$ be some threshold (= regularization parameter).

Hard thresholding:

$$\hat{\theta}_j(\text{hard}) = \begin{cases} Y_j & \text{if } |Y_j| > \lambda \\ 0 & \text{if } |Y_j| \le \lambda \end{cases}, \ j = 1, \ldots, n.$$

Soft thresholding:

$$\hat{\theta}_j(\text{soft}) = \begin{cases} Y_j - \lambda & \text{if } Y_j > \lambda \\ Y_j + \lambda & \text{if } Y_j < -\lambda \\ 0 & \text{if } |Y_j| \leq \lambda \end{cases}, \ j = 1, \ldots, n.$$

Lemma 5.1. $\hat{\theta}(\text{hard})$ *minimizes*

$$\sum_{j=1}^{n} (Y_j - \vartheta_j)^2 + \lambda^2 \#\{\vartheta_j \neq 0\}.$$

□

Lemma 5.2. $\hat{\theta}(\text{soft})$ *minimizes*

$$\sum_{j=1}^{n} (Y_j - \vartheta_j)^2 + 2\lambda \sum_{j=1}^{n} |\vartheta_j|.$$

□

We refer to $\#\{\vartheta_j \neq 0\} = \sum_{j=1}^{n} |\vartheta_j|^0$ as the ℓ_0-penalty and $\sum_{j=1}^{n} |\vartheta_j|$ as the ℓ_1-penalty.

The estimators $\hat{\theta}(\text{hard})$ and $\hat{\theta}(\text{soft})$ have similar oracle properties (see Section 6 for the explanation of this terminology). We will prove this for the soft thresholding estimator.

Define

$$N(\vartheta) = \#\{\vartheta_j \neq 0\}.$$

Lemma 5.5. *Let* $\hat{\theta} = \hat{\theta}(\text{soft})$, $\lambda = \lambda_n = \sqrt{2\sigma \log n/n}$. *On the set* $\{\max_{1 \leq j \leq n} |\epsilon_j| \leq \lambda_n\}$, *we have for all* $0 < \delta \leq 1$

$$\|\hat{\theta} - \theta\|_n^2 \leq (1 + \delta^2) \min_{\vartheta} \left\{ \|\vartheta - \theta\|_n^2 + \frac{64}{\delta^2} \lambda_n^2 N(\vartheta) \right\}.$$

Proof. Let θ^* be arbitrary. Write

$$\text{pen}(\vartheta) = 2\lambda_n \sum_{j=1}^{n} |\vartheta_j|$$

$$= \text{pen}_1(\vartheta) + \text{pen}_2(\vartheta),$$

with

$$\text{pen}_1(\vartheta) = 2\lambda_n \sum_{\theta_j^* \neq 0} |\vartheta_j|,$$

$$\text{pen}_1(\vartheta) = 2\lambda_n \sum_{\theta_j^* = 0} |\vartheta_j|.$$

Then

$$\|\hat{\theta}-\theta\|_n^2 \leq 2\sum_{j=1}^n \epsilon_j(\hat{\theta}_j-\theta_j^*) + \text{pen}(\theta^*) - \text{pen}(\hat{\theta}) + \|\theta^*-\theta\|_n^2$$

$$\leq 2\lambda_n \sum_{j=1}^n |\hat{\theta}_j-\theta_j^*| + \text{pen}(\theta^*) - \text{pen}(\hat{\theta}) + \|\theta^*-\theta\|_n^2$$

$$\leq \text{pen}_1(\hat{\theta}-\theta^*) + \text{pen}_2(\hat{\theta}-\theta^*) + \text{pen}(\theta^*) - \text{pen}(\hat{\theta}) + \|\theta^*-\theta\|_n^2$$

$$\leq 4\text{pen}_1(\hat{\theta}-\theta^*) + \|\theta^*-\theta\|_n^2$$

$$\leq 4\lambda_n\sqrt{N(\theta^*)}\|\hat{\theta}-\theta^*\|_n + \|\theta^*-\theta\|_n^2$$

$$\leq 4\lambda_n\sqrt{N(\theta^*)}\|\hat{\theta}-\theta\|_n + 4\lambda_n\sqrt{N(\theta^*)}\|\theta^*-\theta\|_n + \|\theta^*-\theta\|_n^2$$

$$= I + II + III.$$

If $(I+II) \leq \delta III$ we now have

$$\|\hat{\theta}-\theta\|_n^2 \leq (1+\delta)III = (1+\delta)\|\theta^*-\theta\|_n^2.$$

If $(I+II) \geq \delta III$ and $I \leq II$ we have

$$III \leq \frac{1}{\delta}(I+II) \leq \frac{2}{\delta}II,$$

which implies $\|\theta^*-\theta\|_n^2 \leq \frac{4}{\delta^2}16\lambda_n^2 N(\theta^*)$. But then

$$I+II+III \leq 2\frac{1+\delta}{\delta}II = 2\frac{1+\delta}{\delta}4\lambda_n\sqrt{N(\theta^*)}\|\theta^*-\theta\|_n$$

$$\leq 4(\frac{1+\delta}{\delta})^2 16\lambda_n^2 N(\theta^*).$$

If $(I+II) \geq \delta III$ and $I \geq II$ we get

$$I+II+III \leq \frac{1}{\delta}(I+II) \leq \frac{2}{\delta}I,$$

which gives

$$\|\hat{\theta}-\theta\|_n \leq \frac{2}{\delta}4\lambda_n\sqrt{N(\theta^*)}.$$

□

6. The oracle.

$$Y_j = \theta_j + \epsilon_j,\ j=1,\ldots,n,$$

with $\epsilon_1,\ldots,\epsilon_n$ i.i.d. $\mathcal{N}(0,\frac{\sigma^2}{n})$.

Let $\mathcal{J} \subset \{1,\ldots,n\}$. Define

$$\hat{\theta}_j(\mathcal{J}) = \begin{cases} Y_j & \text{if } j \in \mathcal{J} \\ 0 & \text{if } j \notin \mathcal{J} \end{cases}.$$

Lemma 6.1. *We have*

$$\mathbf{E}\|\hat{\theta}(\mathcal{J}) - \theta\|_n^2 = \sum_{j \notin \mathcal{J}} \theta_j^2 + \frac{|\mathcal{J}|\sigma^2}{n}.$$

□

Oracle θ^{oracle}:

$$\|\theta^{\text{oracle}} - \theta\|_n^2 = \min_{\mathcal{J} \subset \{1,\ldots,n\}} \left\{ \sum_{j \notin \mathcal{J}} \theta_j^2 + \frac{|\mathcal{J}|\sigma^2}{n} \right\}.$$

7. Discretization.

$$Y_i = f_0(x_i) + \epsilon_i,\ i = 1, \ldots, n.$$

with $\epsilon_1, \ldots, \epsilon_n$ independent $\mathcal{N}(0,1)$, and $x_i \in \mathcal{X}$, $Y_i \in \mathbf{R}$, $i = 1, \ldots, n$.

Let $\mathcal{F}$ be a finite collection of functions, and

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n |Y_i - f_0(x_i)|^2.$$

Define

$$\|f^* - f_0\|_n = \min_{f \in \mathcal{F}} \|f - f_0\|_n.$$

Lemma 7.1. *We have for all* $\delta > 0$,

$$\mathbf{P}\left(\|\hat{f} - f_0\|_n^2 \geq 2\|f^* - f_0\|_n^2 + 576\delta^2 + \frac{576 \log |\mathcal{F}|}{n} \right) \leq \exp[-n\delta^2].$$

Proof. We have

$$\|\hat{f} - f_0\|_n^2 \leq \frac{2}{n} \sum_{i=1}^n \epsilon_i (\hat{f}(x_i) - f^*(x_i)) + \|f^* - f_0\|_n^2,$$

and

$$\mathbf{P}\left(\max_{f \in \mathcal{F}, \|f - f^*\|_n^2 \geq 288\delta^2 + \frac{288 \log |\mathcal{F}|}{n}} \frac{\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) - f^*(x_i)}{\|f - f^*\|_n^2} \geq \frac{1}{12} \right)$$

$$\leq \exp\left[\log |\mathcal{F}| - \frac{n}{288} [288\delta^2 + \frac{288 \log |\mathcal{F}|}{n}] \right]$$

$$\leq \exp[-n\delta^2].$$

We also have, if

$$\|f - f_0\|_n^2 \geq 2\|f^* - f_0\|_n^2 + 576\delta^2 + \frac{576 \log |\mathcal{F}|}{n},$$

then

$$\|f - f^*\|_n^2 \geq 288\delta^2 + \frac{288 \log |\mathcal{F}|}{n}.$$

□

Let $\{\mathcal{F}_m\}_{m \in \mathcal{M}}$ be a collection of nested finite models, and let $\mathcal{F} = \cup_{m \in \mathcal{M}} \mathcal{F}_m$. Define

$$\text{pen}(f) = \min_{m:\ f \in \mathcal{F}_m} c \frac{\log |\mathcal{F}_m|}{n}.$$

Let

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(x_i)|^2 + \text{pen}(f) \right\},$$

and

$$f^* = \arg\min_{f \in \mathcal{F}} \left\{ \|f - f_0\|_n^2 + \text{pen}(f) \right\}.$$

Lemma 7.2. *We have*

$$\mathbf{E}[\|\hat{f} - f_0\|_n^2 + \text{pen}(\hat{f})] \leq 2[\|f^* - f_0\|_n^2 + \text{pen}(f^*)] + \frac{c'}{n}.$$

□

8. General penalties.

$$Y_i = f_0(x_i) + \epsilon_i,\ i = 1, \ldots, n.$$

with $\epsilon_1, \ldots, \epsilon_n$ independent $\mathcal{N}(0,1)$, and $x_i \in \mathcal{X}$, $Y_i \in \mathbf{R}$, $i = 1, \ldots, n$.

Let

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(x_i)|^2 + \text{pen}(f) \right\},$$

and

$$f^* = \arg\min_{f \in \mathcal{F}} \left\{ \|f - f_0\|^2 + \text{pen}(f) \right\}.$$

Definition 8.1. *The δ-entropy $H(\delta, \mathcal{F})$ is the logarithm of the minimum number of balls with radius δ necessary to cover $\mathcal{F}$.*

Lemma 8.2. *For*

$$\sqrt{n}\delta_n^2 \geq c \left(\int_0^{\delta_n} H^{1/2}(u, \{\|f - f^*\|_n^2 + \text{pen}(f) \leq \delta_n^2\}) du \vee \delta_n \right),$$

we have

$$\mathbf{E}[\|\hat{f} - f_0\|_n^2 + \text{pen}(\hat{f})] \leq 2[|f^* - f_0\|_n^2 + \text{pen}(f^*) + \delta_n^2] + \frac{c'}{n}.$$

□

9. Application to the classical penalty.

$$Y_i = f_0(x_i) + \epsilon_i,\ i = 1, \ldots, n.$$

with $\epsilon_1, \ldots, \epsilon_n$ independent $\mathcal{N}(0,1)$. and $x_i = i/n$, $i = 1, \ldots, n$.

Let

$$\text{pen}(f) = \min_{\lambda} \left\{ \lambda^2 \int_0^1 |f'(x)|^2 dx + \frac{c}{n\lambda} \right\}.$$

We will apply Lemma 8.1.

10. Robust regression. Let Y_i depend on some covariable x_i, $i = 1, \ldots, n$. Assume $Y_1, \ldots, Y_n$ are independent. Let $\gamma : \mathbf{R} \to \mathbf{R}$ be a convex loss function satisfying the Lipschitz condition

$$|\gamma(x) - \gamma(y)| \leq |x - y|,\ x, y \in \mathbf{R}.$$

Examples.

- $\gamma(x) = |x|$, $x \in \mathbf{R}$,

- $\gamma(x) = \beta |x| \text{l}\{x < 0\} + (1 - \beta)|x| \text{l}\{x > 0\}$, $x \in \mathbf{R}$. Here $0 < \beta < 1$ is fixed.

We consider the estimator

$$\hat{f}_n = \arg \min_{f = \sum_{j=1}^n \theta_j \psi_j} \left\{ \frac{1}{n} \sum_{i=1}^n \gamma(Y_i - f(x_i)) + \lambda_n \sum_{j=1}^n |\theta_j| \right\}.$$

We will derive an oracle inequality for $\hat{f}_n$, using the same arguments as in Lemma 5.5.

11. Density estimation. Let $X_1, \ldots, X_n$ be i.i.d. with distribution P on $\mathcal{X}$, and suppose the denisty $p_0 = dP/d\mu$ w.r.t. some σ-finite measure μ exisits. Let

$$\Lambda = \{f \in L_2(P) : \int f dP = 0\}.$$

Define

$$b(f) = \log \int \text{e}^f d\mu,$$

and $\mathcal{F} = \{f \in \Lambda : \ b(f) < \infty\}$. We assume that

$$f_0 = \log p_0 - \int \log p_0 dP \in \mathcal{F}.$$

Consider now an orthonormal system $\{\psi_j\} \subset \mathcal{F}$ in $L_2(\mu)$. Let $p_f = \exp[f - b(f)]$, $f \in \mathcal{F}$, and consider the (penalized maximum likelihood) estimator

$$\hat{f}_n = \arg \max_{f=\sum_j \theta_j \psi_j} \left\{ \int \log p_f dP_n - \lambda_n \sum_j |\theta_j| \right\}.$$

We will derive an oracle inequality for $\hat{f}_n$, using the same arguments as in Lemma 5.5.

12. Classification. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. copies of (X, Y), where $X \in \mathcal{X}$ is an instance and $Y \in \{0, 1\}$ is a label. A classifier is a subset $G \in \mathcal{X}$. Using the classifier G, we predict $Y = 1$ whenever $X \in \mathcal{G}$. Bayes rule is to use the classifier

$$G_0 = \{x : \ \eta(x) > 1/2\},$$

with

$$\eta(x) = E(Y|X = x), \ x \in \mathcal{X}.$$

This classifier minimizes the prediction error

$$R(G) = P(Y \mathrm{l}_G(X) < 0) = E|Y - \mathrm{l}_G(X)|.$$

Consider the excess risk

$$R(G) - R(G_0) = \int_{G \triangle G_0} |2\eta - 1| dQ.$$

Here, Q denotes the distribution of X.

The empirical risk is

$$R_n(G) = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \mathrm{l}_G(X_i)|.$$

Let $H_B(\delta, \mathcal{G}, Q)$ denote the δ-entropy with bracketing, for the measure Q, of a class $\mathcal{G}$ of subsets of $\mathcal{X}$.

Theorem 12.1. (Mammen and Tsybakov(1999)) *Let*

$$\hat{G}_n = \arg \min_{G \in \mathcal{G}} R_n(G),$$

where $\mathcal{G}$ is a class of subsets of $\mathcal{X}$, satisfying for some constants $c > 0$ and $0 < \rho < 1$,

$$H_B(\delta, \mathcal{G}, Q) \leq c\delta^{-\rho}, \ \delta > 0.$$

Moreover, suppose that for some $\kappa \geq 1$ *and* $\sigma_0 > 0$,

$$R(G) - R(G_0) \geq Q^{\kappa}(G \triangle G_0)/\sigma_0, \ G \subset \mathcal{X}.$$

Let

$$R(G^*) = \min_{G \in \mathcal{G}} R(G).$$

Then

$$\mathbf{E}(R(\hat{G}_n) - R(G_0)) \leq \text{Const.} \left\{ R(G^*) - R(G_0) + n^{-\frac{\kappa}{2\kappa+\rho-1}} \right\}.$$

□

In Tsybakov and van de Geer (2003) one can find an oracle inequality for a penalized empirical risk estimator, which shows adaptivity to both the margin parameter κ as well as to the complexity parameter ρ. The method is based on the same arguments as in Lemma 5.5.

13. Some references for Part 2.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proceedings 2nd International Symposium on Information Theory*, P.N. Petrov and F. Csaki, Eds., Akademia Kiado, Budapest, 267-281

Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Prob. Theory Rel. Fields* **113** 301-413.

Devroye, L. , Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York, Berlin, Heidelberg.

Donoho, D.L., and Johnstone, I.M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425-455

Donoho, D.L. (1995). De-noising via soft-thresholding. *IEEE Transactions in Information Theory* **41**, 613-627

Edmunds, E., and Triebel, H. (1992). Entropy numbers and approximation numbers in function spaces. II. *Proceedings of the London Mathematical Society (3)* **64**, 153-169

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, Approximation and Statistical Applications.* Lecture Notes in Statistics, vol. 129. Springer, New York, Berlin, Heidelberg.

Hastie, T., Tibshirani, R.,and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer, New York

Koenker, R., and Bassett Jr. G. (1978). Regression quantiles. *Econometrica* **46**, 33-50

Koenker, R., Ng, P.T. and Portnoy, S.L. (1992). Nonparametric estimation of conditional quantile functions. L_1 *Statistical Analysis and Related Methods*, Ed. Y. Dodge, Elsevier, Amsterdam, 217-229

Koenker, R., Ng, P.T. and Portnoy, S.L. (1994). Quantile smoothing splines. *Biometrika* **81**, 673-680

Koltchinskii, V. and Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.* **30** 1-50.

Korostelev, A. P. and Tsybakov, A. B. (1993). *Minimax Theory of Image Reconstruction.* Lecture Notes in Statistics **82**, Springer, New York, Berlin, Heidelberg.

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

Loubes, J.-M., and van de Geer S. (2002). Adaptive estimation, using soft thresholding type penalties. *Statistica Neerlandica* **56**, 453-478

Ledoux, M., and Talagrand, M. (1991). *Probability in Banach Spaces, Isoperimetry and Processes.* Springer, Berlin

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808 - 1829.

Pinkus, A. (1985) *n-widths in Approximation Theory.* Springer, New York

Portnoy, S. (1997). Local asymptotics for quantile smoothing splines. *Ann. Statist.* **25**, 414-434

Portnoy, S., and Koenker, R. (1997). The Guassian hare and the Laplacian tortoise: computability of squared error versus absolute-error estimators, with discussion. *Stat. Science* **12**, 279-300

Schölkopf, B. and Smola, A. (2002). *Learning with Kernels*, MIT Press, Cambridge.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464

Tibshirani, R. (1996). Regression analysis and selection via the LASSO. *Journal Royal Statist. Soc. B* **58**, 267-288

Tsybakov, A.B. and van de Geer, S. (2003). Square root penalty: adaptation to the margin in classification, and in edge estimation. Prépublication PMA-820, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VII.

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

van de Geer, S. (2001). Least squares estimation with complexity penalties. *Mathematical Methods of Statistics* **10**, 355-374.

van de Geer, S. (2002). M-estimation using penalties or sieves. *J. Statist. Planning Inf.* **108**, 55-69

van de Geer, S. (2003). Adaptive quantile regression. Techn. Report MI 2003-05, University of Leiden.

van der Vaart, A.W., and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes, with Applications to Statistics.* Springer, New York

Vapnik, V. N. (1998). *Statistical Learning Theory.* Wiley, New York.