Basic tools from empirical processes

Applications to statistics

Sébastien Loustau

LAREMA, Université d'Angers

Dynstoch meeting, June 2010, Angers

Non serious motivation

Angers 16 — 19 June 2010
"The principal aim of the DYNSTOCH network is to make a major contribution to the theory of **statistical inference** for **stochastic processes** by taking advantage..."

Non serious motivation

Angers 16 — 19 June 2010
"The principal aim of the DYNSTOCH network is to make a major contribution to the theory of **statistical inference** for **stochastic processes** by taking advantage..."

... not so far from
"The principal aim of the DYNSTOCH network is to make a major contribution to the theory of **stochastic processes** for **statistical inference** by taking advantage..."

Non serious motivation

Angers 16 — 19 June 2010
"The principal aim of the DYNSTOCH network is to make a major contribution to the theory of **statistical inference** for **stochastic processes** by taking advantage..."

... not so far from
"The principal aim of the DYNSTOCH network is to make a major contribution to the theory of **stochastic processes** for **statistical inference** by taking advantage..."

Empirical processes for statistical inference !

Serious motivation

$Z_1, \ldots, Z_n$ i.i.d. from P. Given $\mathcal{G}$, we aim at:

$$f^* = \arg\min_{f \in \mathcal{G}} \mathbb{E}_P l(f, Z) = \arg\min_{f \in \mathcal{G}} R(f),$$

where l is some loss function.
We study the performance (consistency, rates of convergence) of

$$\hat{f}_n = \arg\min_{f \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} l(f, Z_i) = \arg\min_{f \in \mathcal{G}} R_n(f),$$

in terms of the excess risk:

$$R(\hat{f}_n) - R(f^*).$$

Where do empirical processes appear ?

$$\begin{aligned} R(\hat{f}_n) - R(f^*) &\leq R(\hat{f}_n) - R_n(\hat{f}_n) + R_n(f^*) - R(f^*) \\ &\leq 2 \sup_{f \in \mathcal{G}} |R_n(f) - R(f)|. \end{aligned}$$

$\Rightarrow$ Aim: to control uniformly the empirical process $R_n(f) - R(f)$ indexed by $\mathcal{G}$.

Examples

- Density estimation:

$$\hat{p}_n = \arg\max_{p \in \mathcal{P}} \frac{1}{n} \sum_{i=1}^{n} \log p(Z_i).$$

- Regression:

$$\hat{g}_n = \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i))^2.$$

- Classification:

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+.$$

Outlines

1. Uniform Law of Large Numbers
   Entropy/Chaining/Symmetrization/Hoeffding's inequality/Consistency of M-estimators.
2. Increments of empirical processes
   Finer results at a neighbourhood of fixed function give rates of convergence of M-estimators.
3. Statistical Learning theory
   Examples of Rademacher of RKHS Balls, Besov Balls give rates of convergence of SVM estimators.

Notations, definitions

- $Z_1, \ldots, Z_n$ i.i.d. with law P, and for g measurable,

$$Pg := \mathbb{E}_P g(Z) \text{ and } P_n g := \frac{1}{n}\sum_{i=1}^{n} g(Z_i).$$

- $\mathcal{G}$ **satisfies ULLN** if $\sup_{g\in\mathcal{G}} |P_n g - Pg| \longrightarrow 0$ almost surely.
- $N_p(\mathcal{G}, \delta, Q) :=$ smallest number of balls of radius δ to cover $\mathcal{G}$:

$$\forall g \in \mathcal{G}, \exists i \in \{1, \ldots, N_p\} : \|g - g_i\|_{L^p(Q)} \leq \delta.$$

  $H_p(\mathcal{G}, \delta, Q) := \log N_p(\mathcal{G}, \delta, Q)$ is called the **entropy** of $\mathcal{G}$.
- $N_p^B(\mathcal{G}, \delta, Q) :=$ smallest number of δ-brackets to cover $\mathcal{G}$:

$$\forall g \in \mathcal{G}, \exists i \in \{1, \ldots, N_p^B\} : g_i^L \leq g \leq g_i^U and \|g_i^L - g_i^U\|_{L^p(Q)} \leq \delta.$$

  $H_p^B(\mathcal{G}, \delta, Q) := \log N_p^B(\mathcal{G}, \delta, Q)$ is called the **entropy with bracketing** of $\mathcal{G}$.

Preliminary result

Lemma

$H_1^B(\mathcal{G}, \delta, P) < \infty$, $\forall \delta > 0 \Rightarrow \mathcal{G}$ satisfies ULLN.

Proof :

Let $g \in \mathcal{G}$. then $\exists$ $[g_j^L, g_j^U]$ δ-bracket, $j \in \{1, \dots N\}$:

$$\begin{aligned}(P_n - P)g &\leq P_n g_j^U - Pg = (P_n - P)g_j^U + P(g_j^U - g)\\ &\leq (P_n - P)g_j^U + \delta,\end{aligned}$$

and similarly $(P_n - P)g \geq (P_n - P)g_j^L - \delta$.

Since $([g_j^L, g_j^U])_{j=1\dots N}$ is finite, we have (LLN) for n great enough:

$$\max_{j=1\dots N} |(P_n - P)g_j^U| \leq \delta \text{ and } \max_{j=1\dots N} |(P_n - P)g_j^L| \leq \delta \; a.s.$$

Then

$$\sup_{g \in \mathcal{G}} |P_n g - Pg| \leq 2\delta \; a.s.$$

Chaining method

Let $\mathcal{G} \subset L^2(Q)$: $\sup_{\mathcal{G}} \|g\|_Q \leq R$.
Idea : approximate $g \in \mathcal{G}$ by a finite class called a chain.
Consider $(g_j^s)_{j=1\dots N_s}$ a $2^{-s}R$-covering set of $\mathcal{G}$, for $s = 1, \dots, S$:

$$\forall g \in \mathcal{G}, \exists g^s \in \{g_1^s, \dots, g_{N_s}^s\} : \|g - g_s\|_Q \leq 2^{-s}R.$$

Then we can write :

$$g = g - g^S + \sum_{s=1}^{S}(g^s - g^{s-1})$$

S large enough to get $\|g - g^S\|_Q$ small enough
$\sum_{s=1}^{S}(g^s - g^{s-1})$ only involves finite number of functions.

Application: intermediate result

Lemma

Let $(\xi_1, \ldots, \xi_n)$ fixed. Suppose W_i, $i = 1, \ldots n$ are such that:

$$\mathbb{P}(|\sum_{i=1}^{n} W_i \gamma_i| \geq a) \leq C_1 \exp(-\frac{a^2}{C_2 \sum_{i=1}^{n} \gamma_i^2}).$$

Assume $\sup_{\mathcal{G}} \|g\|_{Q_n} \leq R$ where $Q_n = \frac{1}{n} \sum \delta_{\xi_i}$. Then for

$$\sqrt{n}\delta \geq C(\int_{\delta/8K}^{R} \sqrt{H_2(u, \mathcal{G}, Q_n)} du \vee R),$$

we have:

$$\mathbb{P}(\sup_{g \in \mathcal{G}} |\frac{1}{n} \sum_{i=1}^{n} W_i g(\xi_i)| \geq \delta \wedge \frac{1}{n} \sum_{i=1}^{n} W_i^2 \leq K^2) \leq C \exp(-\frac{n\delta^2}{C^2 R^2}).$$

Proof using chaining

Consider $(g_j^s)_{j=1\dots N_s}$ $2^{-s}R$ covering set of $\mathcal{G}$ w.r.t. $L^2(Q_n)$, for $s=1,\dots S$. Rewrite, for $g\in\mathcal{G}$:

$$\frac{1}{n}\sum_{i=1}^n W_i g(\xi_i) = \frac{1}{n}\sum_{i=1}^n W_i(g(\xi_i)-g^S(\xi_i)) + \frac{1}{n}\sum_{i=1}^n W_i g^S(\xi_i).$$

Choosing $S=\min\{s\geq 1 : 2^{-s}R\leq \frac{\delta}{2K}\}$ ensures (with Cauchy-Schwartz), on the event $\{\frac{1}{n}\sum W_i^2\leq K^2\}$:

$$\frac{1}{n}\sum_{i=1}^n W_i(g(\xi_i)-g^S(\xi_i)) \leq K\|g-g^S\|_{Q_n}\leq \frac{\delta}{2}.$$

It remains to control the probability:

$$\mathbb{P}(\max_{j=1,\dots N_S}|\frac{1}{n}\sum_{i=1}^n W_i g_j^S(\xi_i)|\geq \frac{\delta}{2}).$$

Write $g^S = \sum_{s=1}^S (g^s - g^{s-1})$ and for $\sum \eta_s \leq 1$:

$$\begin{aligned} & \mathbb{P}(\sup_{g\in\mathcal{G}} |\frac{1}{n}\sum_{s=1}^S \sum_{i=1}^n W_i(g^s - g^{s-1})(\xi_i)| \geq \frac{\delta}{2}) \\ \leq & \sum_{s=1}^S \mathbb{P}(\sup_{g\in\mathcal{G}} |\frac{1}{n}\sum_{i=1}^n W_i(g^s - g^{s-1})(\xi_i)| \geq \frac{\delta}{2}\eta_s) \\ \leq & \sum_{s=1}^S C_1 \exp(H_2(2^{-s}R, \mathcal{G}, Q_n)) \exp(-\frac{n\delta^2\eta_s^2}{C_2 2^{-2s} R^2}). \end{aligned}$$

With a good choice of (η_s), we get the result.

Symmetrization

Goal : replace the study of an empirical process to the study of a symmetrized version.
Idea : consider a ghost sample $(Z_i')_{i=1}^n$ i.i.d. from Z and independent of $(Z_i)_{i=1}^n$. Then:

$$\epsilon_i(f(Z_i) - f(Z_i')) \sim f(Z_i) - f(Z_i').$$

where $(\epsilon_i)_{i=1}^n$ are i.i.d. Rademacher variables.
Then we have in expectation :

$$\mathbb{E} \sup_{g \in \mathcal{G}} |P_n - P|(g) \leq 2\mathbb{E} \sup_{g \in \mathcal{G}} |\frac{1}{n} \sum_{i=1}^n \epsilon_i g(Z_i)|.$$

or in probability :

$$\mathbb{P}(\sup_{g \in \mathcal{G}} |P_n - P|(g) \geq \delta) \leq 4\mathbb{P}(\sup_{g \in \mathcal{G}} |\frac{1}{n} \sum_{i=1}^n \epsilon_i g(Z_i)| \geq \frac{\delta}{4})$$

Hoeffding's inequality

Theorem

Consider $Z_1, \dots Z_n$ centered independent r.v. such that $a_i \leq Z_i \leq b_i$, $i = 1, \dots n$. Then $\forall a > 0$:

$$\mathbb{P}(\sum_{i=1}^{n} Z_i \geq a) \leq \exp(-2\frac{a^2}{\sum_{i=1}^{n}(b_i - a_i)^2}).$$

Hence we get for ϵ_i i.i.d. Rademacher:

$$\mathbb{P}(|\sum_{i=1}^{n} \epsilon_i \gamma_i| \geq a) \leq 2\exp(-\frac{a^2}{2\sum_{i=1}^{n}\gamma_i^2}).$$

ULLN under entropy condition

Theorem

Assume $\sup_{g\in\mathcal{G}} \|g\|_\infty < R$ and:

$$\frac{1}{n}H_2(\delta, \mathcal{G}, P_n) \xrightarrow{\mathbb{P}} 0, \ \forall \delta > 0.$$

Then $\mathcal{G}$ satisfies ULLN.

Proof:

Apply Symmetrization:

$$\mathbb{P}(\sup_{g\in\mathcal{G}} |P_n - P|(g) \geq \delta) \leq 4\mathbb{P}(\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(X_i) \geq \frac{\delta}{4})$$

Hoeffding's inequality:

$$\mathbb{P}(|\sum_{i=1}^{n} \epsilon_i \gamma_i| \geq a) \leq \exp(-\frac{a^2}{2\sum_{i=1}^{n} \gamma_i^2}).$$

Apply intermediate result conditionnaly on $Z_1, \ldots, Z_n$ on the set

$$A_n = \{\sqrt{n}\delta \geq C(R\sqrt{H_2(\frac{\delta}{32}, \mathcal{G}, P_n)} \vee R)\},$$

we get:

$$\mathbb{P}(\sup_{g\in\mathcal{G}} |\frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(X_i)| \geq \frac{\delta}{4}) \leq C\exp(-\frac{n\delta^2}{C^2R^2}) + \mathbb{P}(A_n^C) \overset{n\to\infty}{\to} 0.$$

Applications to MLE estimator

Observe $X_1, \ldots, X_n$ i.i.d. with density $p_0 \in \mathcal{P}$ with respect to μ σ-finite measure. Suppose there exists :

$$\hat{p}_n = \arg\max_{p \in \mathcal{P}} P_n(\log p).$$

We want the a.s. convergence of the quantity:

$$h(\hat{p}_n, p_0) = \sqrt{\frac{1}{2} \int (\sqrt{\hat{p}_n} - \sqrt{p_0})^2 d\mu}.$$

For convex $\mathcal{P}$, we have:

$$h^2(\hat{p}_n, p_0) \leq (P_n - P)(\frac{2\hat{p}_n}{\hat{p}_n + p_0}).$$

Then it remains to get ULLN for the class

$$\mathcal{G} = \{\frac{2p}{p + p_0}, \, p \in \mathcal{P}\}.$$

Applications to MLE estimator (2)

Examples :

- The class of monotone Lebesgue densities:

$$\mathcal{P} = \{p \text{ is a decreasing density on } [0,1]\}.$$

Here use the convexity of $\mathcal{P}$.

- The class of Lebesgue smooth densities:

$$\mathcal{P} = \{p : [0,1] \to \mathbb{R}^+ : \int p d\mu = 1, \int_0^1 (p^{(m)}(x))^2 dx \leq M^2\}.$$

Here $H_1^B(\mathcal{P}, \delta, \mu) \leq A\delta^{-\frac{1}{m}}$.

Increments of empirical process

Where do empirical processes appear ? (a finer bound)

$$\begin{aligned} R(\hat{f}_n) - R(f^*) &\leq R(\hat{f}_n) - R_n(\hat{f}_n) + R_n(f^*) - R(f^*) \\ &\leq \sup_{f \in \mathcal{G}(\delta)} |(R_n - R)(f - f^*)|. \end{aligned}$$

$\Rightarrow$ We study the behaviour of $\nu_n(g - g_0) = \sqrt{n}(P_n - P)(g - g_0)$ on $\mathcal{G}(\delta) = \{g \in \mathcal{G} : \|g - g_0\|_P \leq \delta\}$.
Agenda :

- Bernstein's inequality
- Uniform Bernstein over $\mathcal{G}$
- application to $\mathcal{G}(\delta)$ for $\delta \to 0$.

Bernstein's inequality

Theorem

Let $\|g\|_\infty \leq K$ and $\|g\|_P \leq R$. Then:

$$\mathbb{P}(\nu_n(g) \geq a) \leq \exp(-\frac{a^2}{2(\frac{aK}{\sqrt{n}} + R^2)}).$$

It gives subgaussian or subexponential tails and we are interested in subgaussian:

$$\mathbb{P}(\nu_n(g) \geq a) \leq \exp(-\frac{a^2}{4R^2}) \text{ holds for } a \leq \frac{\sqrt{n}R^2}{K},$$

in particular when $R \to 0$.

Uniform Bernstein under entropy condition

Theorem

Let $\mathcal{G}$ such that $\sup_{\mathcal{G}} \|g\|_\infty \leq K$, $\sup_{\mathcal{G}} \|g\|_P \leq R$ and $\int_0^1 \sqrt{H_2^B(\mathcal{G}, u, P)} du < \infty$. Then take

$$C_0(\int_0^R \sqrt{H_2^B(\mathcal{G}, u, P)} du \vee R) \leq a \leq C_1 \frac{\sqrt{n} R^2}{K},$$

then

$$\mathbb{P}(\sup_{g \in \mathcal{G}} |\nu_n(g)| \geq a) \leq \exp(-\frac{a^2}{C^2(C_1+1)R^2}).$$

Proof : Chaining similar to the intermediate result.

Application to $\mathcal{G}(\delta_n)$ when $\delta_n \longrightarrow 0$

Theorem

Suppose $H_2^B(\mathcal{G}, \delta, P) \leq A\delta^{-\alpha}$, $0 < \alpha < 2$. Then for $\delta \geq n^{-\frac{1}{2+\alpha}}$, we have

$$\mathbb{P}(\sup_{g \in \mathcal{G}(\delta)} |\nu_n(g) - \nu_n(g_0)| \geq c\delta^{1-\frac{\alpha}{2}}) \leq C \exp(-\frac{C'\delta^{-\alpha}}{C''}).$$

Proof:

- $\delta \longrightarrow 0$ not to fast to have subgaussian tails.
- $a \sim \int_0^\delta \sqrt{H_2(\mathcal{G}, u, P)} du \sim \delta^{1-\frac{\alpha}{2}}$ to ensure uniform Bernstein.

Application to $\mathcal{G}(\delta_n)$ when $\delta_n \longrightarrow 0$

Consequence:

$$(i)\mathbb{P}\Big(\sup_{g\in\mathcal{G}(n^{-\frac{1}{2+\alpha}})} |\nu_n(g)-\nu_n(g_0)| \geq Tn^{-\frac{2-\alpha}{2+\alpha}}\Big) \leq \exp\Big(-\frac{Tn^{\frac{\alpha}{2+\alpha}}}{C}\Big).$$

$$(ii)\mathbb{P}\Big(\sup_{g\in\mathcal{G}^C(n^{-\frac{1}{2+\alpha}})} \frac{|\nu_n(g)-\nu_n(g_0)|}{\|g-g_0\|^{1-\frac{\alpha}{2}}} \geq T\Big) \leq \exp\Big(-\frac{T}{C}\Big).$$

$\Rightarrow$ We arrive at:

$$\sup_{g\in\mathcal{G}} \frac{|\nu_n(g)-\nu_n(g_0)|}{\|g-g_0\|^{1-\frac{\alpha}{2}} \vee n^{-\frac{2}{2+\alpha}}} = O_{\mathbb{P}}(1).$$

Proof of (ii): Peeling

Goal: to study the tails of $\frac{\nu_n(g)}{w(g)}$ from tails of $\nu_n(g)$.
Idea: consider $(m_s)_{s=1}^S$ decreasing sequence and peel $\mathcal{G}$ as follows:

$$\mathcal{G} \subseteq \bigcup_{s=1}^{S} \mathcal{G}_s, \text{ where } \mathcal{G}_s = \{g \in \mathcal{G} : m_s \leq w(g) \leq m_{s-1}\}.$$

Then write:

$$\begin{aligned} \mathbb{P}(\sup_{g\in\mathcal{G}} \frac{\nu_n(g)}{w(g)} \geq a) &\leq \sum_{s=1}^{S} \mathbb{P}(\sup_{g\in\mathcal{G}(s)} \frac{\nu_n(g)}{w(g)} \geq a) \\ &\leq \sum_{s=1}^{S} \mathbb{P}(\sup_{g:w(g)\leq m_{s-1}} \nu_n(g) \geq a m_s). \end{aligned}$$

Here to get (ii), $w(g) = \|g - g_0\|$, and $m_s = 2^{-s}$.

Application to get rates of convergence of MLE

Observe $X_1, \ldots, X_n$ i.i.d. with density $p_0 \in \mathcal{P}$. We want to upper bound the quantity:

$$h(\hat{p}_n, p_0) = \sqrt{\frac{1}{2}\int (\sqrt{\hat{p}_n} - \sqrt{p_0})^2 d\mu}.$$

Using for instance:

$$h^2(\hat{p}_n, p_0) \leq (P_n - P)(\frac{\sqrt{\hat{p}_n}}{\sqrt{p_0}}).$$

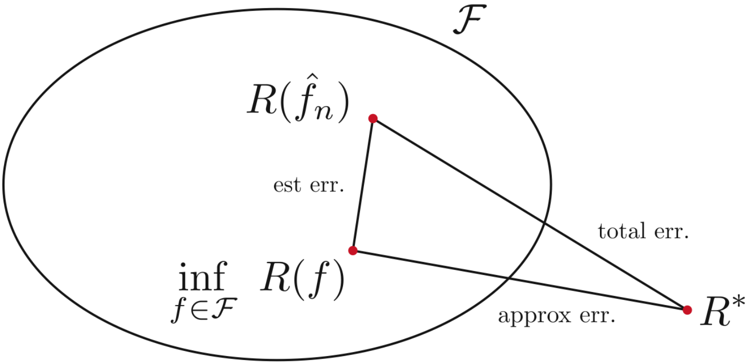
If $H_2^B(\delta, \frac{\mathcal{P}^{1/2}}{p_0^{1/2}}, P) \leq \delta^{-\alpha}$, we have:

$$h^2(\hat{p}_n, p_0) \leq (P_n - P)(\frac{\sqrt{\hat{p}_n}}{\sqrt{p_0}}) \leq O_P(n^{-\frac{1}{2}})h^{1-\frac{\alpha}{2}}(\hat{p}_n, p_0) \vee O_P(n^{-\frac{2}{2+\alpha}}),$$

and gives

$$h(\hat{p}_n, p_0) = O_P(n^{-\frac{1}{2+\alpha}}).$$

How to choose $\mathcal{F}$?



A penalized M-estimator for classification

Consider the SVM minimization:

$$\min_{f\in\mathcal{H}}\left[\frac{1}{n}\sum_{i=1}^{n}(1-Y_i f(X_i))_+ + \alpha_n\|f\|_{\mathcal{H}}^2\right],$$

where:

- $(X_i, Y_i)\in\mathbb{R}^d\times\{-1,+1\}$, $i=1,\ldots n$ are i.i.d.,
- $\mathcal{H}$ is a given functional space,
- α_n smoothing parameter to determine.

Rates of convergence

- Estimation error:

$$R(\hat{f}_n, f^*) \leq C \inf_{f \in \mathcal{H}} \left(R(f, f^*) + \alpha_n \|f\|_{\mathcal{H}}^2 \right) + \delta_n(\alpha_n).$$

- Approximation error:

$$a(\alpha_n) = \inf_{f \in \mathcal{H}} \left(R(f, f^*) + \alpha_n \|f\|_{\mathcal{H}}^2 \right).$$

Then you have:

$$R(\hat{f}_n, f^*) \leq Ca(\alpha_n) + \delta_n(\alpha_n) \overset{\alpha_n \sim n^*}{\lesssim} n^{-?}$$

Local Rademacher

We are interested in

$$\mathbb{E} \sup_{f \in B_{\mathcal{H}}(R): \mathbb{E} f(X)^2 \leq \delta} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| := \mathbb{E}\mathcal{R}\mathrm{ad}_n(R, \delta),$$

where ϵ_i are i.i.d. $P(\epsilon_i = 1) = P(\epsilon_i = -1) = \frac{1}{2}$.

- In the RKHS situation, we have:

$$\mathbb{E}\mathcal{R}\mathrm{ad}_n(R, \delta) \leq \frac{1}{\sqrt{n}} \inf_{d \in \mathbb{N}} \left(\sqrt{d\delta} + R \sqrt{\sum_{j>d} \lambda_j} \right),$$

  where (λ_j) eigenspectrum of the integral operator $L_K : f \mapsto \int f(x) K(x, \cdot) dx$.
- What happens if $\mathcal{H} = \mathcal{B}_{spq}(\mathbb{R}^d)$?

Besov case

Theorem

Suppose X admits a bounded density ρ with compact support. Then if $s > \frac{d}{p}$ and $1 \leq p \leq 2$,

$$\forall \delta > 0,\ \mathbb{E} \sup_{f \in B(R): \mathbb{E} f(X)^2 \leq \delta} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \leq \frac{C}{\sqrt{n}} R^{\frac{d}{2u}} \delta^{\frac{s - \frac{d}{p}}{2u}},$$

where $u = s + d\left(\frac{1}{2} - \frac{1}{p}\right)$.

Consequence: if $f^* \in \mathcal{B}_{rpq}$,

$$R(\hat{g}_n) - R(f^*) = O_{\mathbb{P}}(n^{-\frac{r}{2s-r}\frac{2u}{2u+d}}),$$

where we choose $\alpha_n \sim n^{-\frac{2u}{2u+d}}$.

References

- S. Van De Geer Empirical Processes and M-estimation, 2000.
- J. Wellner and A. Van der Vaart Weak convergence and empirical processes, 2000.
- S. Mendelson Local Rademacher Complexities, 2005.
- P. Massart Some applications of concentration inequalities to statistics, 2000.
- S. Loustau Penalized ERM over Besov spaces, 2009.