Concentration, self-bounding functions

S. Boucheron[1] and G. Lugosi[2] and P. Massart[3]

[1]Laboratoire de Probabilités et Modèles Aléatoires
Université Paris-Diderot

[2]Economics
University Pompeu Fabra

[3]Département de Mathématiques
Université Paris-Sud

Cachan, 03/02/2011

Context

$X_1, \ldots, X_n$: $\mathcal{X}$-valued, independent random variables.

$$F : \mathcal{X}^n \to \mathbb{R}$$

$$Z = F(X_1, \ldots, X_n)$$

Goal : upper-bounds on

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right]$$

$$\mathbb{P}\{Z \geq \mathbb{E}Z + t\} \quad \text{and} \quad \mathbb{P}\{Z \leq \mathbb{E}Z - t\}$$

$t > 0$

Context ...

Non-asymptotic tail bounds for *functions of many independent random variables that do not depend too much on any of them*.

1. high dimensional geometry ;
2. random combinatorics ;
3. statistics ;

A variety of methods

1. Martingales
2. Talagrand's induction method
3. Transportation method
4. Entropy method
5. Chatterjee's method (exchangeable pairs)

Inspiration: Gaussian concentration

Theorem (Tsirelson, Borell, Gross,..., 1975)

$X_1, \ldots, X_n \sim_{i.i.d.} \mathcal{N}(0,1)$
$F : \mathbb{R}^n \to \mathbb{R}$
L-Lipschitz (w.r.t.) Euclidean distance
$Z = F(X_1, \ldots, X_n)$

$$\mathrm{var}[Z] \leq L^2 \quad \textit{Poincaré's inequality}$$

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] \leq \frac{\lambda^2 L^2}{2}$$

$$\mathbb{P}\{Z \geq \mathbb{E}Z + t\} \leq e^{-t^2/(2L^2)}$$

Efron-Stein inequalities (1981)

$Z = F(X_1, X_2, ..., X_n)$, (independent R.V)

$X'_1, ..., X'_n \sim X_1, ..., X_n$ but $\perp\!\!\!\perp$ from $X_1, ..., X_n$.
For each $i \in \{1, ..., n\}$
$Z'_i = F(X_1, ..., X_{i-1}, X'_i, X_{i+1}, ...X_n)$.
$X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$.
F_i: a function of $n-1$ arguments
$Z_i = F_i(X_1, ..., X_{i-1}, X_{i+1}, X_n) = F_i(X^{(i)})$.

Theorem (Jackknife estimates of variance are biased)

$V_+ = \sum_{i=1}^n \mathbb{E}\left[(Z - Z'_i)_+^2 \mid X_1, ..., X_n\right]$
$V = \sum_i (Z - Z_i)^2$

$$\mathrm{Var}[Z] \leqslant \mathbb{E}[V_+] \leqslant \mathbb{E}[V]$$

Exponential Efron-Stein inequalities

Theorem (Sub-Gaussian behavior)

If $V_+ \leq v$ *then, for* $\lambda \geq 0$

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] \leq \frac{\lambda^2 v}{2}$$

Theorem (B., Lugosi and Massart, 2003)

For $0 \leq \lambda \leq 1/\theta$,

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] \leq \frac{\lambda\theta}{(1-\lambda\theta)} \log \mathbb{E}\left[e^{\lambda V_+/\theta}\right]$$

Entropy method

Entropy

Y an $\mathcal{X}$-valued random variable
f non-negative (measurable) function over $\mathcal{X}$

$$\mathrm{Ent}\,[f] = \mathbb{E}\,[f(Y)\log f(Y)] - \mathbb{E}[f(Y)]\log\mathbb{E}[f(Y)]\,.$$

Why ?

if $Y = \exp(\lambda(Z - \mathbb{E}Z))$, let $G(\lambda) = \frac{1}{\lambda}\log\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right]$

$$\frac{1}{\lambda^2}\frac{\mathrm{Ent}\left[e^{\lambda(Z-\mathbb{E}Z)}\right]}{\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right]} = \frac{dG(\lambda)}{d\lambda}$$

Basis of Herbst's argument : bounds on Entropy can be translated into differential inequalities for logarithmic moment generating functions.

Gross logarithmic Sobolev inequality

Theorem (Gross,. . . , 1975)

$X_1, \ldots, X_n \sim_{i.i.d.} \mathcal{N}(0,1)$
$F : \mathbb{R}^n \to \mathbb{R}$ *differentiable*
$Z = F(X_1, \ldots, X_n)$

$$\mathrm{var}[Z] \leq \mathbb{E}\left[\left\|\nabla F(X_1, \ldots, X_n)\right\|^2\right]$$

$$\mathrm{Ent}\left[Z^2\right] \leq 2\mathbb{E}\left[\|\nabla F\|^2\right]$$

Bounds on entropy

Subadditivity

$X_1, \ldots, X_n \perp\!\!\!\perp$ random variables. $Z = f(X_1, \ldots, X_n) \geq 0$

$$\mathrm{Ent}^{(i)}[Z] = \mathbb{E}^{(i)}[Z \log Z] - \mathbb{E}^{(i)} Z \log \mathbb{E}^{(i)} Z$$

$$\mathrm{Ent}[f(X_1, \ldots, X_n)] \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Ent}^{(i)}[Z]\right]$$

Upper-bounding Entropy of a function of a single random variable

Expected value minimizes expected Bregman divergence with respect to convex function $x \mapsto x \log x$

$$\mathrm{Ent}[Z] \leq \inf_{u>0} \mathbb{E}[Z(\log Z - u) - (Z - u)]$$

Entropy method in a nutshell

Summary

The entropy method converts a modified logarithmic Sobolev inequality into a differential inequality involving the logarithm of the moment generating function of Z.

Starting point

Theorem (a modified logarithmic sobolev inequality.)

let $\phi(x) = e^x - x - 1$. *For any* $\lambda \in \mathbb{R}$,

$$\lambda \mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda Z}\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z} \phi\left(-\lambda(Z - Z_i)\right)\right].$$

Use different conditions to upper-bound $\sum_{i=1}^{n} \phi\left(-\lambda(Z - Z_i)\right)$ ...

Folklore

$Z \geq 0$ and $\mathrm{Var}[Z] \leq a\mathbb{E}Z \Rightarrow \mathrm{Var}[\sqrt{Z}] \leq a$

If $Z_n \sim \mathrm{Pois}(n\mu)$, then $\sqrt{Z_n} - \mathbb{E}\sqrt{Z_n} \to \mathcal{N}(0, 1/4)$ (Cramer's Delta)

Lemma

If $X \sim \mathrm{Pois}$, *let* $v = (\mathbb{E}X)\mathbb{E}[1/(4X+1)]$, *for* $\lambda \geq 0$,

$$\log \mathbb{E}\left[e^{\lambda(\sqrt{X} - \mathbb{E}[\sqrt{X}])}\right] \leq v\lambda(e^{\lambda} - 1) .$$

$$\text{for } t > 0 \qquad \mathbb{P}\left\{\sqrt{X} \geq \mathbb{E}\sqrt{X} + t\right\} \leq \exp\left(-\frac{t}{2}\log\left(1 + \frac{t}{2v}\right)\right) .$$

Proof

Poisson Poincaré inequality (Klaassen 1985)

$X \sim \text{Pois}$, $Z = f(X)$

$$\text{Var}[Z] \leq \mathbb{E}X \times \mathbb{E}[|Df(X)|^2]\,,$$

with $Df(X) = f(X+1) - f(X)$.

Consequence:

$$\text{Var}[\sqrt{X}] \leq v\,.$$

Poisson logarithmic Sobolev inequality (L. Wu, Bobkov-Ledoux)

$$\text{Ent}[Z] \leq \mathbb{E}X \times \mathbb{E}\left[Df \times D\log f\right]$$

Self-bounding property (I)

$f : \mathcal{X}^n \to \mathbb{R}$ is said to have the self-bounding property if $\exists f_i : \mathcal{X}^{n-1} \to \mathbb{R}$:

1. $\forall x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $\forall i = 1, \ldots, n$,

$$0 \leq f(x) - f_i(x^{(i)}) \leq 1$$

2.

$$\sum_{i=1}^{n} \left(f(x) - f_i(x^{(i)}) \right) \leq f(x) \, .$$

where $x^{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

Examples

1. Suprema of positive bounded empirical processes.
   $X_i = (X_{i,s})_{s \in \mathcal{T}}$, $\mathcal{T}$ finite, $0 \leq X_{i,s} \leq 1$. X_i independent.

$$Z = \sup_{s \in \mathcal{T}} \sum_{i=1}^{n} X_{i,s}$$

2. Suprema of bounded empirical processes $X_{i,s} \leq 1$ (relaxing the second assumption)
3. Largest eigenvalue of a Gram matrix

$$\sup_{u \colon \|u\|_2 = 1} u^T \sum_{i=1}^{n} X_i X_i^T u$$

Binomial-Poisson tails

If $|\mathcal{T}| = 1$, Bennet inequality holds

- $$h(u) = (1+u)\log(1+u) - u, \quad u \geq -1$$
  and
  $$\phi(v) = \sup_{u \geq -1} (uv - h(u)) = e^v - v - 1 \, .$$
- $$\log \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}Z)}\right] \leq \phi(\lambda)\mathbb{E}Z \qquad \forall \lambda \in \mathbb{R} \, .$$
- $$\mathbb{P}\{Z \geq \mathbb{E}Z + t\} \leq \exp\left(-\mathbb{E}Z h\left(\frac{t}{\mathbb{E}Z}\right)\right) \quad \forall t > 0$$

Chi-square tails

- $X_i \sim \mu_i \chi_1^2$ weighted chi-square random variables
- $Z = \sum_{i=1}^n X_i$
- $v = 2\sum_{i=1}^n \mu_i^2$ and $c = \max_i \mu_i / 2$
- $$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] \leq \frac{v\lambda^2}{2(1-c\lambda)}$$
- Bernstein inequality

$$\mathbb{P}\{Z \geq \mathbb{E}Z + t\} \leq \exp\left(-\frac{t^2}{2(v+ct)}\right)$$

- Bennett inequality entails Bernstein inequality (with scale factor 1/3)

Self-bounding property and concentrations inequalities

$$h(u) = (1+u)\log(1+u) - u, \quad u \geq -1$$

and

$$\phi(v) = \sup_{u \geq -1} (uv - h(u)) = e^v - v - 1 \, .$$

Theorem (B., Lugosi and Massart 2002-3)

If Z satisfies the self-bounding property,

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] \leq \phi(\lambda)\mathbb{E}Z \qquad \forall \lambda \in \mathbb{R} \, .$$

$$\mathbb{P}\{Z \geq \mathbb{E}Z + t\} \leq \exp\left(-\mathbb{E}Z h\left(\frac{t}{\mathbb{E}Z}\right)\right) \quad \forall t > 0$$

$$\mathbb{P}\{Z \leq \mathbb{E}Z - t\} \leq \exp\left(-\mathbb{E}Z h\left(\frac{-t}{\mathbb{E}Z}\right)\right) \quad \forall 0 < t \leq \mathbb{E}Z \, .$$

... with applications

Definition (Conditional Rademacher averages)

$\epsilon_1, \ldots, \epsilon_n$ Rademacher variables

$$(x_1, \ldots, x_n) \mapsto F(x_1, \ldots, x_n) = \mathbb{E}\left[\sup_{s \in \mathcal{T}} \sum_{i=1}^{n} \epsilon_i x_{i,s}\right]$$

Symmetrization inequalities

(Giné and Zinn, 1984)

$$\frac{1}{2}\mathbb{E}\left[F(X_1, \ldots, X_n)\right] \leq \mathbb{E}\left[\sup_{s \in \mathcal{T}} \sum_{i=1}^{n} X_{i,s}\right] \leq 2\mathbb{E}\left[F(X_1, \ldots, X_n)\right]$$

Theorem (B., Lugosi and Massart 2003)

Conditional Rademacher averages are self-bounding.

Consequences

$$\mathrm{Var}[F(X_1,\ldots,X_n)] \leq \mathbb{E}[F(X_1,\ldots,X_n)]$$

while

$$\mathrm{Var}[\sup_{s\in\mathcal{T}} \sum_{i=1}^{n} X_{i,s}] \leq \sup_{s\in\mathcal{T}} n\,\mathrm{Var}[X_{1,s}] + 2\mathbb{E}[\sup_{s\in\mathcal{T}} \sum_{i=1}^{n} X_{i,s}]\,.$$

Conditional Rademacher averages : kind of weighted Bootstrap estimates.

Variations on a theme

Definition (weakly (a, b)-self-bounding)

$f : \mathcal{X}^n \to [0, \infty)$ is *weakly* (a, b)*-self-bounding* if
$\exists f_i : \mathcal{X}^{n-1} \to [0, \infty) : \forall x \in \mathcal{X}^n$

$$\sum_{i=1}^{n} \left(f(x) - f_i(x^{(i)})\right)^2 \leq af(x) + b\,.$$

Definition (strongly (a, b)-self-bounding)

$f : \mathcal{X}^n \to [0, \infty)$ is *strongly* (a, b)*-self-bounding* if
$\exists f_i : \mathcal{X}^{n-1} \to [0, \infty) : \forall i = 1, \ldots, n, \text{and} x \in \mathcal{X}^n,$

$$0 \leq f(x) - f_i(x^{(i)}) \leq 1\,, \text{ and } \sum_{i=1}^{n} \left(f(x) - f_i(x^{(i)})\right) \leq af(x) + b\,.$$

Definition (Submodular function)

$$f : 2^n \to \mathbb{R}$$
$$f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$$

Submodularity implies neither monotonicity nor non-negativity

Example

The capacity of cuts in a directed graph is submodular.

Lemma (Vondrák)

Non-negative 1-Lipschitz submodular function are $(2, 0)$*-self-bounding.*

Efron-Stein inequality

$$\mathrm{Var}[Z] \leq \mathbb{E}\left[\sum_{i=1}^{n}(f(X) - f_i(X^{(i)}))^2\right]$$

Remark

Both definitions imply that $Z = f(X)$ satisfies

$$\mathrm{Var}[Z] \leq a\mathbb{E}Z + b\,.$$

Theorem (Maurer 2006)

$X = (X_1, ..., X_n)$ $\mathcal{X}$-valued independent random variables.
$f : \mathcal{X}^n \to [0, \infty)$ weakly (a, b)-self-bounding function $(a, b \geq 0)$.
Let $Z = f(X)$.
If $\forall i \leq n, \forall x \in \mathcal{X}^n, f_i(x^{(i)}) \leq f(x)$, then
for all $0 \leq \lambda \leq 2/a$,

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] \leq \frac{(a\mathbb{E}Z + b)\lambda^2}{2(1 - a\lambda/2)}$$

and for all $t > 0$,

$$\mathbb{P}\{Z \geq \mathbb{E}Z + t\} \leq \exp\left(-\frac{t^2}{2\left(a\mathbb{E}Z + b + at/2\right)}\right) .$$

Lower tails

Theorem (McDiarmid and Reed 2008, B., Lugosi, Massart, 2009)

$X = (X_1, ..., X_n)$ $\mathcal{X}$-valued independent random variables. Let $f : \mathcal{X}^n \to [0, \infty)$ be a weakly (a, b)-self-bounding function $(a, b \geq 0)$. Let $Z = f(X)$ and define $c = (3a - 1)/6$. If, $f(x) - f_i(x^{(i)}) \leq 1$ for each $i \leq n$ and $x \in \mathcal{X}^n$, then for $0 < t \leq \mathbb{E}Z$,

$$\mathbb{P}\{Z \leq \mathbb{E}Z - t\} \leq \exp\left(-\frac{t^2}{2\left(a\mathbb{E}Z + b + c\_t\right)}\right) .$$

If $a \geq 1/3$, sub-Gaussian behavior.

Upper tails

Theorem (McDiarmid and Reed 2008, B., Lugosi, Massart, 2009)

$X = (X_1, ..., X_n)$ $\mathcal{X}$-valued independent random variables. Let $f : \mathcal{X}^n \to [0, \infty)$ be a weakly (a, b)-self-bounding function ($a, b \geq 0$). Let $Z = f(X)$ and define $c = (3a - 1)/6$. Then for all $\lambda \geq 0$,

$$\log \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}Z)}\right] \leq \frac{(a\mathbb{E}Z + b)\lambda^2}{2(1 - c_+\lambda)}$$

and for all $t > 0$,

$$\mathbb{P}\{Z \geq \mathbb{E}Z + t\} \leq \exp\left(-\frac{t^2}{2\left(a\mathbb{E}Z + b + c_+ t\right)}\right) .$$

If $a \leq 1/3$, sub-Gaussian behavior.

Proofs

Remark

The entropy method converts a modified logarithmic Sobolev inequality into a differential inequality involving the logarithm of the moment generating function of Z.

Starting point

Theorem (a modified logarithmic sobolev inequality.)

For any $\lambda \in \mathbb{R}$,

$$\lambda \mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda Z}\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z} \phi\left(-\lambda(Z - Z_i)\right)\right].$$

Use different conditions to upper-bound $\sum_{i=1}^{n} \phi\left(-\lambda(Z - Z_i)\right)$ ...

Proofs (...)

Establishing differential inequalities for $G(\lambda) = \log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right]$

1. (a,b)-self-bounding:
   $\mathrm{Ent}\left[e^{\lambda Z}\right] \leq \phi(-\lambda)\mathbb{E}\left[(aZ+b)\, e^{\lambda Z}\right]$
2. (a,b)-weakly self-bounding, $\lambda \geq 0$:
   $\mathrm{Ent}\left[e^{\lambda Z}\right] \leq \frac{\lambda^2}{2}\mathbb{E}\left[(aZ+b)e^{\lambda Z}\right]$.
3. (a,b)-weakly self-bounding, $\lambda \leq 0$:
   $\mathrm{Ent}\left[e^{\lambda Z}\right] \leq \phi(-\lambda)\mathbb{E}\left[(aZ+b)e^{\lambda Z}\right]$.

Key differential inequality

$$[\lambda - a\phi(-\lambda)]\, G'(\lambda) - G(\lambda) \leq v\phi(-\lambda), \tag{1}$$

where $v = a\mathbb{E}Z + b$.

Around the Herbst argument

Lemma

Let $f : I \to \mathbb{R}, \nearrow, C^1$ *where interval* $I \ni 0$ *with* $f(0) = 0$. *Assume* $x \neq 0 \Rightarrow f(x) \neq 0$.
Let g *be* C^0 *on* I *and* G *be* C^∞ *on* I *with* $G(0) = G'(0) = 0$ *and for every* $\lambda \in I$,

$$f(\lambda)G'(\lambda) - f'(\lambda)G(\lambda) \leq f^2(\lambda)g(\lambda) \; .$$

Then, for every $\lambda \in I$, $G(\lambda) \leq f(\lambda) \int_0^\lambda g(x)dx$.

Comparisons

Let ρ be C^0 on $I \ni 0$. Let $a \geq 0$.
Let $H\colon I \to \mathbb{R}$, be C^∞ satisfying

$$\lambda H'(\lambda) - H(\lambda) \leq \rho(\lambda)\left(aH'(\lambda) + 1\right)$$

with $aH'(\lambda) + 1 > 0 \quad \forall \lambda \in I$ and $H'(0) = H(0) = 0$.
Let $\rho_0 : I \to \mathbb{R}$, assume that $G_0 : I \to \mathbb{R}$ is C^∞ with $\forall \lambda \in I$,

$$aG_0'(\lambda) + 1 > 0 \quad \text{and} \quad G_0'(0) = G_0(0) = 0 \quad \text{and} \quad G_0''(0) = 1 \ .$$

Assume also that G_0 solves the differential equation

$$\lambda G_0'(\lambda) - G_0(\lambda) = \rho_0(\lambda)\left(aG_0'(\lambda) + 1\right) \ .$$

If $\rho(\lambda) \leq \rho_0(\lambda)$ for every $\lambda \in I$, then $H \leq G_0$.

Sketch of proof

Key differential inequality

$$\lambda G'(\lambda) - G(\lambda) \leq \phi(-\lambda)(1 + aG'(\lambda)))$$

1. $2G_\gamma(\lambda) = \lambda^2/(1-\gamma\lambda)$ solves $\lambda H'(\lambda) - H(\lambda) \leq \lambda^2(1 + \gamma H'(\lambda)))$
2. Choosing $\gamma = a$ works for $\lambda \geq 0$
3. May not be the best choice for ρ_γ ...
4. Optimizing $\rho_\gamma = (\lambda G'_\gamma(\lambda) - G_\gamma(\lambda))/(1 + aG'_\gamma(\lambda)))$ leads to the desired result

Application: Convex distance

Definition

$A \subseteq \mathcal{X}^n$

B_n^2 unit ball in $\mathbb{R}^n$ endowed with euclidean metric

$$d_T(X, A) = \inf_{y \in A} \sup_{\alpha \in B_2^n} \sum_{i=1}^{n} \alpha_i \mathbb{I}_{X_i \neq y_i}$$

Theorem

$\mathcal{M}(A)$ *: Probability distributions supported by A*

$$d_T(X, A) = \sup_{\alpha \in B_2^n} \inf_{\nu \in \mathbb{M}(A)} \sum_{i=1}^{n} \alpha_i \nu\{X_i \neq Y_i\}$$

Modus operandi

Lemma

For any $A \in \mathcal{X}^n$ and $x \in \mathcal{X}^n$, the function $f(x) = d_T(x, A)^2$ satisfies

$$0 \leq f(x) - f_i(x^{(i)}) \leq 1$$

where f_i is defined by

$$f_i(x^{(i)}) = \inf_{x_i' \in \mathcal{X}} f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)\,. \tag{2}$$

Moreover, f is weakly $(4, 0)$-self-bounding.

Efron-Stein estimates of the variance of d_T

Efron-Stein estimate of variance of $d_T(\cdot, A)$

$$V_+ = \sum_{i=1}^{n} \left(\sqrt{f(x)} - \sqrt{f_i(x^{(i)})} \right)^2$$

Lemma

For all $A \subset \mathcal{X}^n$, V_+ is bounded by 1.

$$V_+ \leq 1.$$

$$\mathrm{Var}\left[d_T(X, A) \right] \leq 1.$$

A consequence of the minmax characterization of d_T

1. $\mathcal{M}(A)$: the set of probability measures on A.
2. we may re-write d_T as

$$d_T(x, A) = \inf_{\nu \in \mathcal{M}(A)} \sup_{\alpha : \|\alpha\|_2 \leq 1} \sum_{j=1}^{n} \alpha_j \mathbb{E}_\nu [\mathbb{I}_{x_j \neq Y_j}] \tag{3}$$

   where $Y = (Y_1, \ldots, Y_n)$ is distributed according to ν.
3. By the Cauchy-Schwarz inequality,

$$d_T(x, A)^2 = \inf_{\nu \in \mathcal{M}(A)} \sum_{j=1}^{n} \left(\mathbb{E}_\nu [\mathbb{I}_{x_j \neq Y_j}] \right)^2 .$$

Weak self-boundedness of d_T^2

Denote the pair (ν, α) at which the saddle point is achieved by $(\widehat{\nu}, \widehat{\alpha})$. For all x,

$\sum_{i=1}^n \left(\sqrt{f(x)} - \sqrt{f_i(x^{(i)})}\right)^2 \leq 1$ since $\left(\sqrt{f(x)} - \sqrt{f_i(x^{(i)})}\right)^2 \leq \widehat{\alpha}_i^2$.

$$\begin{aligned}
&\sum_{i=1}^n \left(f(x) - f_i(x^{(i)})\right)^2 \\
&= \sum_{i=1}^n \left(\sqrt{f(x)} - \sqrt{f_i(x^{(i)})}\right)^2 \left(\sqrt{f(x)} + \sqrt{f_i(x^{(i)})}\right)^2 \\
&\leq \sum_{i=1}^n \widehat{\alpha}_i^2 4 f(x) \\
&\leq 4 f(x) \, .
\end{aligned}$$

Talagrand's convex distance inequality

Theorem (Talagrand 1995)

$$\mathbb{P}\{A\}\mathbb{E}\left[e^{d_T^2(X,A)/4}\right] \leq 1$$

Proof.

$$\mathbb{P}\{X \in A\} = \mathbb{P}\left\{d_T(X,A)^2 \leq \mathbb{E}\left[d_T^2(X,A)\right] - t\right\} \leq \exp\left(-\frac{\mathbb{E}[d_T(X,A)^2]}{8}\right).$$

For $0 \leq \lambda \leq 1/2$,

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] \leq \frac{\lambda^2\, 2\mathbb{E}Z}{1-2\lambda}.$$

Choosing $\lambda = 1/10$ leads to the desired result. □

Suprema of non-centered empirical processes

Well-understood scenarios

1. Suprema of positive bounded empirical processes: self-bounding property
2. Suprema of centered bounded empirical processes : Talagrand's inequality (revisited by Ledoux, Massart, Rio, Klein, Bousquet, ...).

Bennett inequality with variance factor coinciding with the Efron-Stein estimate of variance

Talagrand-. . . -Bousquet inequality

$Z = \sup_{s\in\mathcal{T}} \sum_{i=1}^{n} X_{i,s}$
$X_1, \ldots, X_n \perp\!\!\!\perp$ identifically distributed
$\mathbb{E}X_{i,s} = 0$ and $-1 \leq X_{i,s} \leq 1$
$\sigma^2 = \sup_{s\in\mathcal{T}} \sum_{i=1}^{n} \mathrm{var}[X_{i,s}]$

Efron-Stein estimate of variance

$$\mathrm{var}[Z] \leq v = 2\mathbb{E}Z + \sigma^2$$

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] \leq v\phi(\lambda)$$

For $x > 0$

$$\mathbb{P}\left\{Z \geq \mathbb{E}Z + \sqrt{2vx} + \frac{x}{3}\right\} \leq e^{-x}$$

Another scenario : Excess Empirical Risk

$\widehat{s}, \bar{s}$

$R(s) = \mathbb{E}[X_{i,s}]$ Risk of $s \in \mathcal{T}$
$R_n(s) = \frac{1}{n}\sum_{i=1}^n X_{i,s}$
$\bar{s}$: $R(\bar{s}) = \mathbb{E}X_{i,\bar{s}} = \inf_{s\in\mathcal{T}} \mathbb{E}X_{i,s} = \inf_{s\in\mathcal{T}} R(s)$
$\widehat{s}$: $nR_n(\widehat{s}) = \sum_{i=1}^n X_{i,\widehat{s}} = \inf_{s\in\mathcal{T}} \sum_{i=1}^n X_{i,s} = \inf_{s\in\mathcal{T}} R_n(s)$

Excess risk and empirical counterpart

Excess risk $R(\widehat{s}) - R(\bar{s})$
Excess empirical risk (EER)

$$Z = n(R_n(\bar{s}) - R_n(\widehat{s})) = \sup_{s\in\mathcal{T}} \sum_{i=1}^n (X_{i,\bar{s}} - X_{i,s}) = \sum_{i=1}^n \left(X_{i,\bar{s}} - X_{i,\hat{s}}\right)$$

Variance bounds for EER

- Consequences of Efron-Stein inequalities :

$$\begin{aligned}\mathrm{Var}\left[Z\right] &\leq 2\left(\mathbb{E}\left[\sum_{i=1}^{n-1}(X_{i,\bar{s}} - \mathbb{E}X'_{i,\bar{s}}) - (X_{i,\widehat{s}} - \mathbb{E}X'_{i,\widehat{s}})\right] + \mathbb{E}\left[\sum_{i=1}^{n}(X_{i,\bar{s}} - X_{i,\widehat{s}})^2\right]\right)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}\left[Z\right] &\leq 2\left(\mathbb{E}\left[\sum_{i=1}^{n}(X_{i,\bar{s}} - X_{i,\widehat{s}})^2\right] + \mathbb{E}\left[\sum_{i=1}^{n}(X'_{i,\bar{s}} - X'_{i,\widehat{s}})^2\right]\right)\end{aligned}$$

Consequences of Talagrand's inequalities (and peeling)

Assumptions

$\exists d$ a distance over $\mathcal{T}$,
$\exists \psi, \omega : [0,1] \to \mathbb{R}_+, \nearrow, \psi(x)/x, \omega(x)/x \searrow :$

$$\sqrt{n}\mathbb{E}\left[\sup_{s:\ d(s,\bar{s})\le r} \left|(R(s) - R_n(s)) - (R_n(\bar{s}) - R(\bar{s}))\right|\right] \le \psi(r)$$

$$\mathbb{E}\left[(X_{i,s} - X_{i,\bar{s}})^2\right] \le d(s,\bar{s})^2 \le \omega\left(\sqrt{R(s) - R(\bar{s})}\right)^2$$

Definition

r_* is the positive solution of $\sqrt{n}r^2 = \psi(\omega(r))$

Consequences, cont'd

With probability larger than $1 - \delta$

$$\max\left(R(\hat{s}) - R(\bar{s}), R_n(\bar{s}) - R(\hat{s})\right) \leq \kappa\left(r_*^2 + \frac{\omega(r_*)^2}{n r_*^2}\log\frac{1}{\delta}\right)$$

r_*^2 is called the rate of the estimation problem

$$\max\left(\mathbb{E}\left[R(\hat{s}) - R(\bar{s})\right], \mathbb{E}\left[R_n(\bar{s}) - R(\hat{s})\right]\right) \leq \kappa' r_*^2$$

Combining ...

$$\mathrm{var}\left[Z\right] = \mathrm{var}\left[n(R_n(\bar{s}) - R_n(\widehat{s}))\right] \quad \leq \quad n\kappa''\left(\omega(r_*)^2\right)$$

Bernstein inequality for EER

Theorem (B., Bousquet, Lugosi, Massart, 2005)

$$\left\|(Z - \mathbb{E}Z)_+\right\|_q \leq \sqrt{3q}\left\|\sqrt{V_+}\right\|_q$$

Bernstein like inequality

$$\left\|\left(n(R_n(\bar{s}) - R_n(\widehat{s}))\right)\right\|_q \leq \kappa\left(\sqrt{nq}\omega(r_*) + \sqrt{n}\omega\left(\frac{\omega(r_*)}{\sqrt{n}r_*}\right)q\right)$$

Variance factor $n\omega(r_*)^2$

Scale factor $\sqrt{n}\omega\left(\frac{\omega(r_*)}{\sqrt{n}r_*}\right)$

Works for some statistical learning problems (learning VC-classes under good noise conditions).

References

1. B. and Massart : A high dimensional Wilks phenomenon. *Probability Theory and Related Fields* online. (2010)
2. B., Lugosi and Massart : On concentration of self-bounding functions. *Electronic Journal of Probability* 14 (2009) 1884-1899.

and

1. Maurer. Concentration inequalities for functions of independent variables, *Random Structures and Algorithms*, 29 (2006) 121–138.
2. McDiarmid and Reed. Concentration for self-bounding functions and an inequality of talagrand. *Random Structures and Algorithms*, 29 (2006) 549-577.
3. B. Lugosi and Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16 (2000), 277-292.