Concentration Inequalities with Machine Learning Applications

Stéphane Boucheron

LRI UMR CNRS 8623 Université Paris-Sud

Organization

- From exponential inequalities to the concentration of measure phenomenon
- Concentration inequalities using the entropy method
- Learning-theoretical applications
- Moment inequalities using the generalized entropy method

From exponential inequalities to concentration

Lecture I: from exponential inequalities to concentration

- Introduction and Motivations
- Path to Bernstein inequality
- Martingales with bounded increments
- Efron-Stein inequality

Lecture I: Roadmap

The introduction describes traditional exponential inequalities (Hoeffding/Bernstein) as non-asymptotic counterparts to limit theorems for sums of independent random variables. Concentration inequalities are presented as upper-bounds on tail probabilities for functions of many independent random variables. The scope of concentration inequalities is illustrated on a combinatorial optimization problem.

The path to Bernstein inequality is described in detail, stressing the fact that good bounds on the Log-Laplace transform of a random variable provide exponential bounds on the tail probabilities. The main topic of this course will be the derivation of Bernstein-like inequalities for general functions.

Martingales methods provide a general recipe for constructing Bernstein-like inequalities. The exponential super-martingale associated with martingales with bounded increments allows to refine the celebrated bounded-differences inequality. Despite and because of their generality, using martingale methods may be quite difficult. This has prompted the search for more user-friendly methods as (for example) the entropy method.

The first step in the entropy method is illustrated by the Efron-Stein inequality. The latter inequality provides a general and often tight upper-bound on the the variance of general functions of independent random variables. The Efron-Stein bound is first illustrated on a combinatorial optimization problem.

Appetizer

Basic concern of Probability Theory : sums of independent bounded random variables

$$(X_i)_{i \leq n} \text{ i.i.d. in } [-1,1] \quad \mathbb{E}[X_i] = 0 \quad \mathbb{E}[X_i^2] \leq v$$

$$Z \triangleq \sum_{i=1}^{n} X_i$$

Appetizer

Basic concern of Probability Theory : sums of independent bounded random variables

$$(X_i)_{i \leq n} \text{ i.i.d. in } [-1,1] \quad \mathbb{E}[X_i] = 0 \quad \mathbb{E}[X_i^2] \leq v$$

$$Z \triangleq \sum_{i=1}^{n} X_i$$

Law of Large Numbers (LLN):

$$\forall \epsilon > 0, \quad \lim_n \mathbb{P}\left\{|Z - \mathbb{E}[Z]| > n\epsilon\right\} = 0$$

Central Limit Theorem (CLT, Invariance principle)

$$\forall x \in \mathbb{R} \quad \lim_n \mathbb{P}\left\{\frac{1}{\sqrt{n}}(Z - \mathbb{E}[Z]) < x\right\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi v}} e^{-\frac{x^2}{2v^2}} dx$$

LLN and CLT : asymptotic statements.

☞ Need for statements dealing with fixed values of n.

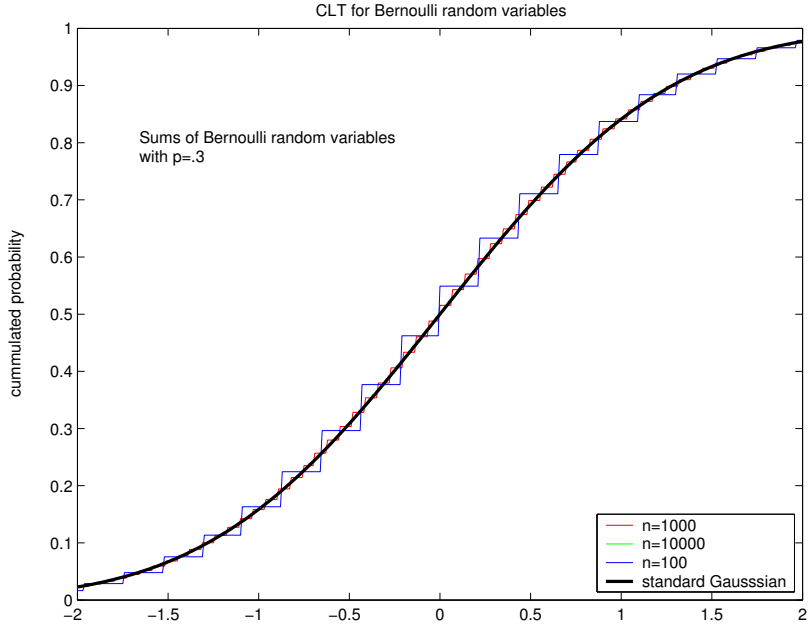
Appetizer

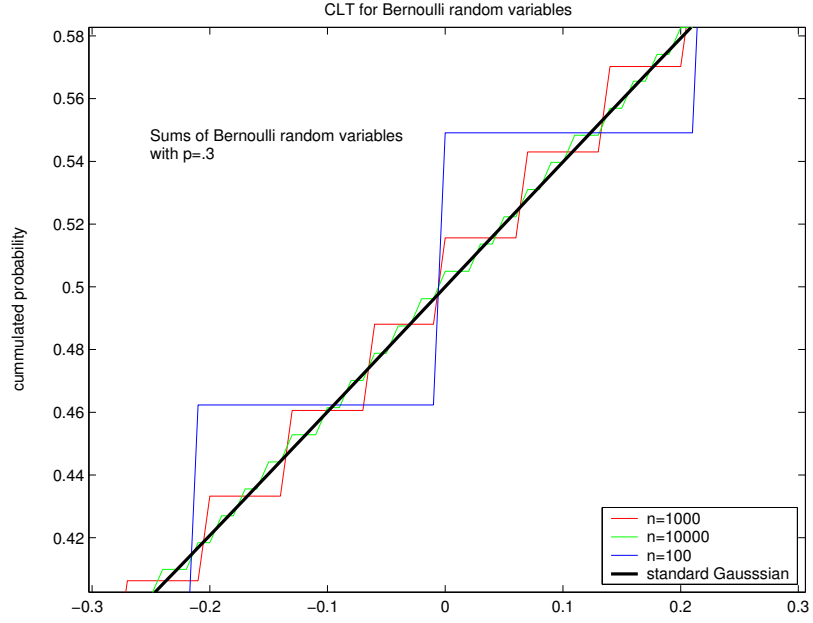
Central Limit Theorem (CLT, Invariance principle)

$$\forall x \in \mathbb{R} \quad \lim_n \mathbb{P}\left\{ \frac{1}{\sqrt{n}} (Z - \mathbb{E}[Z]) < x \right\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi v}} e^{-\frac{x^2}{2v^2}} dx$$

LLN and CLT : asymptotic statements.

☞ Need for statements dealing with fixed values of n.





Appetizer

Probability Theory provides informations about the rate of convergence in the CLT. (Berry-Esseen, 1943)

$$\mathbb{E}[|X_i|^3] < \infty \quad \rightarrow \sup_x \left| \mathbb{P}\left\{ \frac{1}{\sqrt{n}} (Z - \mathbb{E}[Z]) < x \right\} - \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi v}} e^{-\frac{x^2}{2v^2}} dx \right| \leq \frac{C}{\sqrt{n}}$$

Berry-Esseen Theorems say nothing meaningful about

$$\left| \mathbb{P}\left\{ (Z - \mathbb{E}[Z]) > nx \right\} - \int_{x}^{\infty} \frac{1}{\sqrt{2\pi v}} e^{-\frac{nx^2}{2v^2}} dx \right|$$

Large deviations depend upon the P.D. of X_i. !

Appetizer

Probability Theory provides informations about the rate of convergence in the CLT. (Berry-Esseen, 1943)

$$\mathbb{E}[|X_i|^3] < \infty \quad \rightarrow \sup_x \left| \mathbb{P}\left\{ \frac{1}{\sqrt{n}} (Z - \mathbb{E}[Z]) < x \right\} - \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi v}} e^{-\frac{x^2}{2v^2}} dx \right| \leq \frac{C}{\sqrt{n}}$$

Berry-Esseen Theorems say nothing meaningful about

$$\left| \mathbb{P}\left\{ (Z - \mathbb{E}[Z]) > nx \right\} - \int_{x}^{\infty} \frac{1}{\sqrt{2\pi v}} e^{-\frac{nx^2}{2v^2}} dx \right|$$

Large deviations depend upon the P.D. of X_i. !

If X is Gaussian:

$$\begin{aligned} \mathbb{P}\{X \geq t\} &= \int_t^\infty \frac{1}{x\sqrt{2\pi v}} x e^{-\frac{x^2}{2v}} dx \\ &= \left[\frac{-v}{x} \frac{1}{\sqrt{2\pi v}} e^{-\frac{x^2}{2v}} \right]_t^\infty + \int_t^\infty \frac{\sqrt{v}}{\sqrt{2\pi}} \frac{1}{x^2} e^{-\frac{x^2}{2v}} dx \\ &\leq \frac{\sqrt{v}}{\sqrt{2\pi t}} e^{-\frac{t^2}{2v}} \end{aligned}$$

Appetizer

Gaussian tail bounds:

$$\mathbb{P}\{X \geq t\} \quad \leq \quad \frac{\sqrt{v}}{\sqrt{2\pi}t} e^{-\frac{t^2}{2v}}$$

What about sums of i.i.d. random variables ? **Exponential inequalities** provide a widely applicable answer to those questions:

If Z is a sum of n independent centered $[-1,1]$-variables:

$$\mathbb{P}\{Z \geq \mathbb{E}[Z] + t\} \leq \exp\left(-\frac{t^2}{2n}\right) \quad \text{Hoeffding}$$

$$\mathbb{P}\{Z \geq \mathbb{E}[Z] + t\} \leq \exp\left(-\frac{t^2}{2(\mathrm{Var}[Z] + t/3)}\right) \quad \text{Bernstein.}$$

Gaussian-like tail bounds.
In Bernstein inequality, n does not appear in the exponent,

Bernstein inequality looks dimension-free.

Motivations

"While exponential inequalities for sums of independent random variables are at the core of classical probabilities, the new abstract inequalities are far-reaching extensions that apply to considerably more general functions."

M. Talagrand.*Ann. Probability*, 1996

Motivations

"While exponential inequalities for sums of independent random variables are at the core of classical probabilities, the new abstract inequalities are far-reaching extensions that apply to considerably more general functions."

M. Talagrand.*Ann. Probability*, 1996

"Any function of many independent random variables that depend on many of them but not too much on each of them is essentially constant."

M. Talagrand. *Inventiones Mathematicae*, 1996

Motivations

"While exponential inequalities for sums of independent random variables are at the core of classical probabilities, the new abstract inequalities are far-reaching extensions that apply to considerably more general functions."

M. Talagrand.*Ann. Probability*, 1996

"Any function of many independent random variables that depend on many of them but not too much on each of them is essentially constant."

M. Talagrand. *Inventiones Mathematicae*, 1996

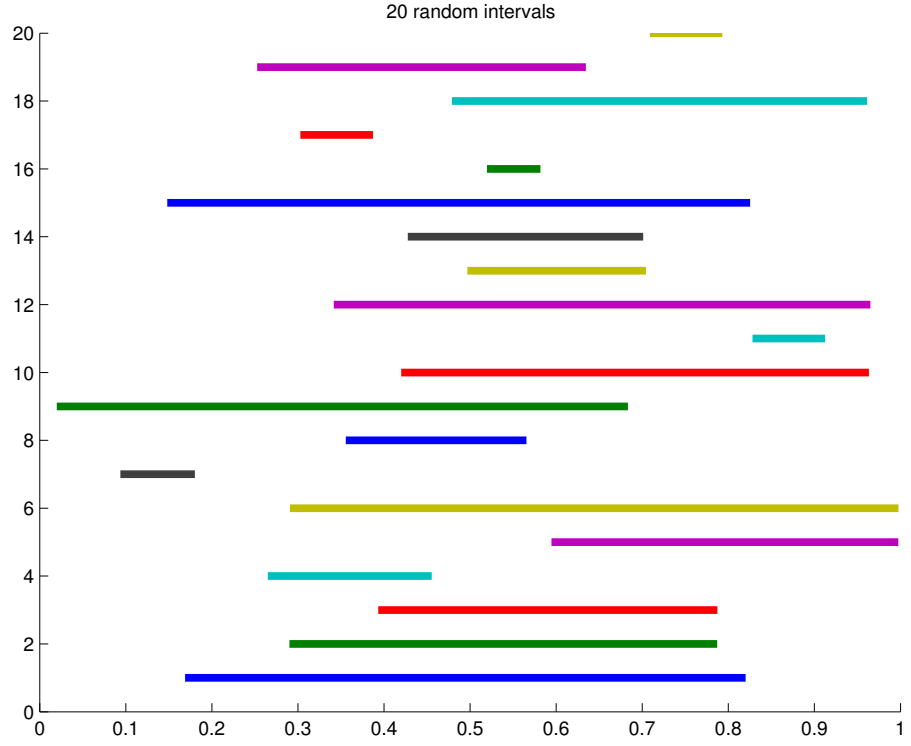
- What's in *depend not too much on each of the variables* ?
- How general should be the *considerably more general functions* ?
- Combinatorial optimization problems provide a cavalcade of illustrations of the concentration of measure phenomenon.

Random interval packings

X_i are intervals from $[0, 1]$: extremities are picked by picking random numbers from $[0, 1]$

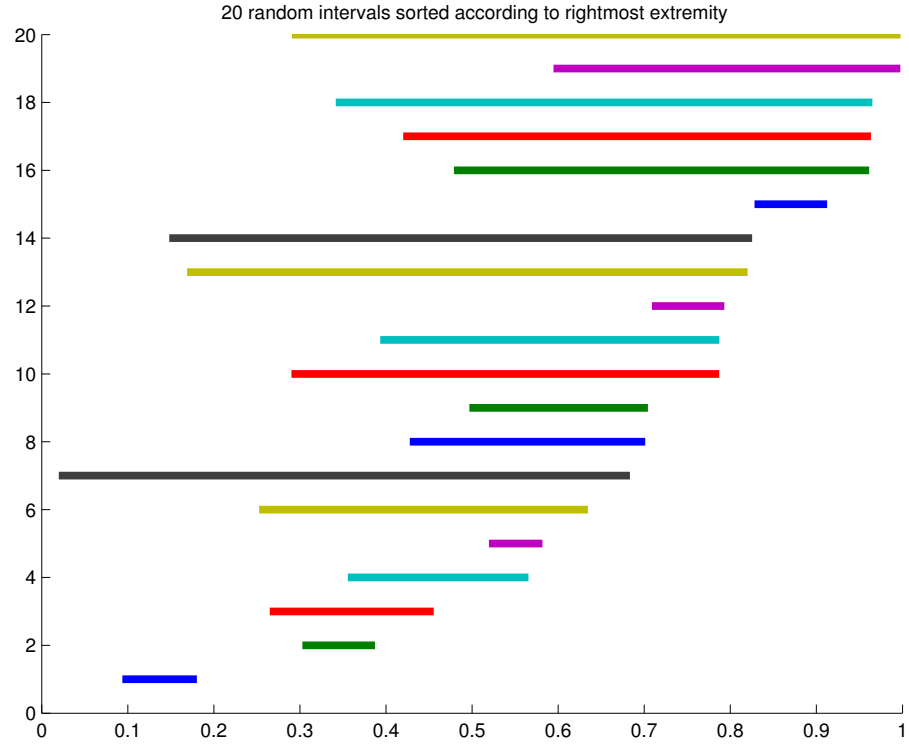
A packing is a set of pairwise disjoint intervals.

Z: maximum cardinality of a packing extracted from $X_1, \ldots, X_n$



Random interval packings

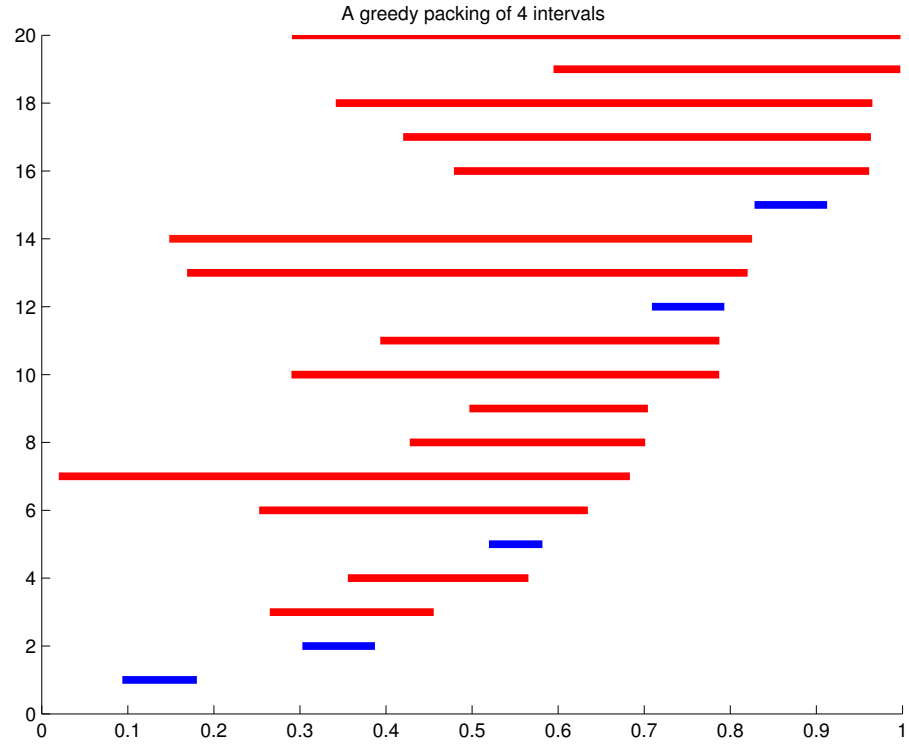
Z: maximum cardinality of a packing extracted from $X_1, \ldots, X_n$



☞: a maximum packing can be constructed (in $\Theta(n \log n)$ operations) by a greedy algorithm.

1 sort intervals according their right-most extremities in ascending order.

Random interval packings



2 the initial packing contains the first interval

3 scan the remaining intervals in ascending order

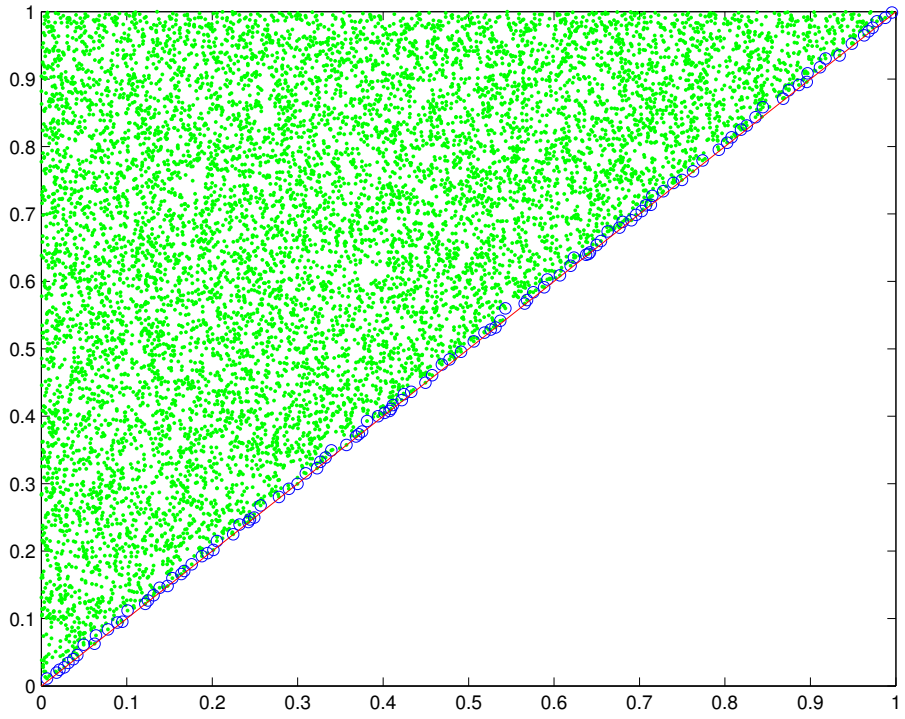
4 add the current interval to the packing if its left-most extremity is larger than the right-most extremity of the last interval in the packing.

Random interval packings

- How does $\mathbb{E}[Z]$ behave as $\sqrt{n}$ goes to infinity?
- Law of large numbers? CLT? Asymptotic theorems ?
- Tail behavior for fixed n ?
- No obvious way to represent Z as a sum of independent random variables.
- Z is a kind of “much more general function” of many independent random variables does not depend to much on each of them.

Random interval packings

1 interval $\leftrightarrow$ 1 point (x, y) in $[0, 1]^2$ with $x \leq y$.



Example: 10000 random intervals, $Z = 114$, $\mathbb{E}[Z] \approx \frac{2}{\sqrt{\pi}}\sqrt{n} \approx 113$

Z satisfies a Central Limit Theorem:
$(Z - \mathbb{E}[Z])/n^{1/4}$ is asymptotically Gaussian with variance $2(4 - \pi)/\pi^{3/2}$.

Random interval packings

☞ If $Z \geq t$, then there exists a sequence of t intervals $X_{i_1}, X_{i_2}, \ldots, X_{i_t}$ that witnesses this fact.
This is just a sequence of t disjoint intervals !

A set property (P) is hereditary if every subset of a set satisfying (P) also satisfies (P)

Z is a configuration function: Z is the largest cardinality of a subset of the random variables that satisfy some hereditary property.

Concentration inequalities assert that:

$$\begin{aligned} \mathbb{P}\left\{Z \geq \mathbb{E}[Z] + t\right\} &\leq \exp\left(-\frac{t^2}{2(\mathbb{E}[Z] + t/3)}\right) \\ \mathbb{P}\left\{Z \leq \mathbb{E}[Z] - t\right\} &\leq \exp\left(-\frac{t^2}{2\mathbb{E}[Z]}\right) \quad . \end{aligned}$$

☞ The concentration of measure perspective deals with deviations around expectation/median by focusing on the structure of increments of the functional under consideration.

Path to Bernstein inequality

$X_1, \dots X_n$, $X_i \in [-1, 1]$, independent, $\mathbb{E}[X_i] = 0$, $\mathrm{Var}[X_i] = v$.

Markov inequality.

Z an $\mathbb{R}$-valued random variable.
f a positive measurable function, such that f is non-decreasing on $[t, \infty)$.

$$\begin{aligned}
\mathbb{P}\{Z > t\} &= \mathbb{P}\Big\{f(Z) \geq f(t)\Big\} \quad f \text{ is non-decreasing.} \\
&= \mathbb{E}\Big[\mathbb{1}_{f(Z)\geq f(t)}\Big] \\
&\leq \mathbb{E}\Big[\mathbb{1}_{f(Z)\geq f(t)}\,\frac{f(Z)}{f(t)}\Big] \\
&\leq \mathbb{E}\Big[\frac{f(Z)}{f(t)}\Big]
\end{aligned}$$

Path to Bernstein inequality

$X_1, \dots X_n$, $X_i \in [-1, 1]$, independent, $\mathbb{E}[X_i] = 0$, $\mathrm{Var}[X_i] = v$.

Markov inequality.

Z an $\mathbb{R}$-valued random variable.
f a positive measurable function, such that f is non-decreasing on $[t, \infty)$.

$$\mathbb{P}\{Z > t\} \leq \mathbb{E}\left[\frac{f(Z)}{f(t)}\right]$$

Examples:

- $f(x) = x^2$, $\quad \mathbb{P}\Big\{Z - \mathbb{E}[Z] > t\Big\} \leq \mathbb{E}\left[\frac{(Z-\mathbb{E}[Z])^2}{t^2}\right] = \frac{\mathrm{Var}(Z)}{t^2}$
- $f(x) = \exp(\lambda x)$ where $\lambda > 0$ $\quad \mathbb{P}\Big\{Z > t\Big\} \leq \exp\Big(- (\lambda t - \log \mathbb{E}\big[e^{-\lambda Z}\big])\Big)$
- $f(x) = |x|^q$ $\quad \mathbb{P}\Big\{Z > t\Big\} \leq \mathbb{E}\left[\frac{|Z|^q}{t^q}\right] \leq \left(\frac{\|Z\|_q}{t}\right)^q$

☞ Markov inequality relates tail-behavior and integrability.

Path to Bernstein inequality

$X_1, \dots X_n$, $X_i \in [-1, 1]$, independent, $\mathbb{E}[X_i] = 0$, Var$[X_i] = v$.

Exponential Markov inequalities for sums of independent random variables. $Z = \sum_{i=1}^{n} X_i$

$$\begin{aligned}
\mathbb{P}\Big\{Z > t\Big\} &\leq e^{-\lambda t}\mathbb{E}\Big[e^{\lambda \sum_i X_i}\Big] \\
&\leq e^{-\lambda t}\mathbb{E}\Big[\prod_i e^{\lambda X_i}\Big] \\
&\leq e^{-\lambda t}\prod_i \mathbb{E}\Big[e^{\lambda X_i}\Big] \quad \text{by independence}
\end{aligned}$$

Path to Bernstein inequality

$X_1, \dots X_n$, $X_i \in [-1, 1]$, independent, $\mathbb{E}[X_i] = 0$, $\mathsf{Var}[X_i] = v$.

Exponential Markov inequalities for sums of independent random variables. $Z = \sum_{i=1}^{n} X_i$

$$\mathbb{P}\Big\{Z > t\Big\} \leq e^{-\lambda t} \prod_i \mathbb{E}\Big[e^{\lambda X_i}\Big]$$

$$\begin{aligned}
\prod_i \Big(\mathbb{E}\Big[e^{\lambda X_i}\Big]\Big) &= \prod_i \left(\mathbb{E}\Big[\sum_{k=0}^{\infty} \frac{\lambda^k X_i^k}{k!}\Big]\right) \\
&\leq \prod_i \left(\mathbb{E}\Big[1 + \sum_{k=2}^{\infty} \frac{\lambda^k |X_i|^k}{k!}\Big]\right) \quad \text{centering} \\
&\leq \prod_i \left(1 + v \sum_{k=2}^{\infty} \frac{\lambda^k}{k!}\right) \quad \text{variance, boundedness} \\
&= \prod_i \Big(1 + v(e^\lambda - \lambda - 1)\Big) \\
&\leq \exp\Big(\mathsf{Var}(Z)(e^\lambda - \lambda - 1)\Big)
\end{aligned}$$

Path to Bernstein inequality

$X_1, \dots X_n$, $X_i \in [-1, 1]$, independent, $\mathbb{E}[X_i] = 0$, $\mathrm{Var}[X_i] = v$.

Exponential Markov inequalities for sums of independent random variables. $Z = \sum_{i=1}^{n} X_i$
With $\tau^*(\lambda) = e^\lambda - \lambda - 1$,

$$\mathbb{P}\Big\{ Z > t \Big\} \leq e^{-\lambda t} \exp\Big(\mathrm{Var}(Z) \tau^*(\lambda) \Big)$$

Convex duality: $\tau(x) = \sup_\lambda [\lambda x - \tau^*(\lambda)] = (x+1)\log(x+1) - x$

$$\begin{aligned}
\mathbb{P}\Big\{ Z > t \Big\} &\leq \exp\left(- \sup_\lambda \big[\lambda t - \mathrm{Var}(Z) \tau^*(\lambda) \big] \right) \\
&\leq \exp\left(-\mathrm{Var}(Z) \tau\left(\frac{t}{\mathrm{Var}(Z)} \right) \right) \quad \text{Bennett} \\
&\leq \exp\left(- \frac{t^2}{2\left(\mathrm{Var}(Z) + t/3\right)} \right) \quad \text{Bernstein}
\end{aligned}$$

Martingale inequalities

Doob's embedding.

$$M_i = \mathbb{E}\Big[f(X_1, X_2, \ldots, X_n) \mid X_1^i\Big] = \mathbb{E}[Z \mid X_1^i]$$

M_i is $\sigma(X_1, \ldots, X_i)$-measurable.

$$M_0 = \mathbb{E}[Z] \qquad M_n = Z$$

$$\mathbb{E}\Big[M_{i+1} \mid X_1^i\Big] = M_i$$

The sequence $\left(M_i\right)_{i=0}^n$ is an $\left(\sigma(X_1^n)\right)_{i=0,\ldots,n}$-adapted martingale.

☞ Doob's embedding allows to represent a general function of many random variables as the last value of a martingale. The increments $M_{i+1} - M_i$ reflect the sensitivity of f with respect to its arguments.

Martingale inequalities

Martingale with bounded increments.
Assumption: for all i, $|M_{i+1} - M_i| \leq 1$ and $\mathbb{E}[Z] = 0$

Increasing process associated with martingale (M_i).

$$\langle M \rangle_i = \sum_{j=1}^{i} \mathbb{E}[(M_j - M_{j-1})^2 \mid X_1^{j-1}]$$

$$\mathrm{Var}(M_n) = \mathrm{Var}(Z) = \mathbb{E}[\langle M \rangle_n]$$

The process $\left(\exp\left(\lambda M_i - \tau^*(\lambda)\langle M \rangle_i\right)\right)_i$ is an $(\sigma(X_1^i))_i$-adapted super-martingale:

$$\mathbb{E}\left[\left(\exp\left(\lambda M_i - \tau^*(\lambda)\langle M \rangle_i\right)\right) \mid X_1^{i-1}\right] \leq \exp\left(\lambda M_{i-1} - \tau^*(\lambda)\langle M \rangle_{i-1}\right)$$

Martingale inequalities

Bounded-differences inequality (McDiarmid).

Assumption: $\langle M \rangle_n \leq c$ with probability 1.

$$\mathbb{E}\Big[e^{\lambda(Z-\mathbb{E}[Z])}\Big] \leq \exp\big(c\tau^*(\lambda)\big)$$

As increments are bounded by 1, we have $\langle M \rangle_n \leq n$.

Azuma inequality (Hoeffding inequality for martingales with bounded increments)

$$\mathbb{P}\Big\{Z - \mathbb{E}[Z] \geq t\Big\} \leq \exp\Big(-\frac{t^2}{2n}\Big)$$

☹ A worst-case upper bound on $\langle M \rangle_n$ may be excessively conservative.
☞ It is highly desirable to take advantage of the integrability of $\langle M \rangle_n$ in order to derive tight tail bounds for martingale with bounded increments.

Martingale inequalities

Bernstein inequality for martingales. Stopping

T is a stopping time with respect to $(\sigma(X_1^i))_i$ if $T \leq i$ is $\sigma(X_1^i)$-measurable.

The optional sampling theorem entails that

$$\Big(\exp\Big(\lambda M_0 - \tau^*(\lambda)\langle Z\rangle_0\Big), \exp\Big(\lambda M_T - \tau^*(\lambda)\langle M\rangle_T\Big)\Big)$$

is a super-martingale.

For a (finite) stopping time T $\quad \mathbb{E}\Big[\exp\Big(\lambda Z_T - \tau^*(\lambda)\langle Z\rangle_T\Big)\Big] \leq 1$

Let V denote a fixed positive quantity

$$T = \min\Big\{n, \min\big\{j \ : \ \langle M\rangle_{j+1} > V\big\}\Big\} \quad \text{is a stopping time.}$$

Martingale inequalities

Bernstein inequality for martingales. Stopping

T is a stopping time with respect to $(\sigma(X_1^i))_i$ if $T \leq i$ is $\sigma(X_1^i)$-measurable.

$$\begin{aligned}
\mathbb{P}\Big\{Z \geq t \wedge T = n\Big\} &\leq \mathbb{P}\Big\{M_T \geq t\Big\} \\
&\leq \mathbb{P}\Big\{\lambda M_T - \tau^*(\lambda)\langle M\rangle_T \geq \lambda t - \tau^*(\lambda)\langle M\rangle_T\Big\} \\
&\leq \exp\Big(-\lambda t + \tau^*(\lambda)V\Big) \\
&\leq \exp\Big(-V\tau\Big(\frac{t}{V}\Big)\Big) \quad \text{Optimizing with respect to } \lambda \\
&\leq \exp\Big(-\frac{t^2}{2(V + t/3)}\Big)
\end{aligned}$$

Martingale inequalities

Bernstein inequality for martingales. Stopping

T is a stopping time with respect to $(\sigma(X_1^i))_i$ if $T \leq i$ is $\sigma(X_1^i)$-measurable.

$$\mathbb{P}\Big\{Z \geq t \wedge T = n\Big\} \leq \exp\Big(-\frac{t^2}{2(V + t/3)}\Big)$$

$$\mathbb{P}\Big\{Z \geq t\Big\} \leq \mathbb{P}\Big\{\langle M\rangle_n \geq V\Big\} + \exp\Big(-\frac{t^2}{2(V + t/3)}\Big)$$

☞ The stopping-time trick allows to search for trade-offs and just requires the martingale to be smooth on average. It is tight for moderate deviations.

☹ The conditional variance process may be as difficult to analyze as Z...

☞ Efron-Stein estimates will serve as surrogates for conditional variance process.

Conventions

- $X_1, \ldots, X_n$: independent random variables.

Conventions

- $X_1, \ldots, X_n$: independent random variables.
- $Z = f(X_1, \ldots, X_n)$: function of *many* independent random variables.

Conventions

- $X_1, \ldots, X_n$: independent random variables.
- $Z = f(X_1, \ldots, X_n)$: function of *many* independent random variables.
- $X'_1, \ldots, X'_n$: independent copies of $X_1, \ldots, X_n$.

Conventions

- $X_1, \ldots, X_n$: independent random variables.
- $Z = f(X_1, \ldots, X_n)$: function of *many* independent random variables.
- $X'_1, \ldots, X'_n$: independent copies of $X_1, \ldots, X_n$.
- Two perturbations of Z :

  X_i is replaced by an independent copy X'_i:

  $Z^{(i)} = f(X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n)$

  Z_i does not depend on X_i.

  $Z_i = f_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$.

Conventions

- $X_1, \ldots, X_n$: independent random variables.
- $Z = f(X_1, \ldots, X_n)$: function of *many* independent random variables.
- $X'_1, \ldots, X'_n$: independent copies of $X_1, \ldots, X_n$.
- Two perturbations of Z :

  X_i is replaced by an independent copy X'_i:

  $Z^{(i)} = f(X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n)$

  Z_i does not depend on X_i.

  $Z_i = f_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$.
- V^+ and V allow to monitor the sensitivity of Z

$$V^+ = \sum_i \mathbb{E}\left[(Z - Z^{(i)})^2 \mathbb{1}_{Z > Z^{(i)}} \mid X_1^n\right]$$

$$V = \sum_i (Z - Z_i)^2 .$$

Conventions

- $X_1, \ldots, X_n$: independent random variables.
- $Z = f(X_1, \ldots, X_n)$: function of *many* independent random variables.
- $X'_1, \ldots, X'_n$: independent copies of $X_1, \ldots, X_n$.
- Two perturbations of Z :

  X_i is replaced by an independent copy X'_i:

  $Z^{(i)} = f(X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n)$

  Z_i does not depend on X_i.

  $Z_i = f_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$.
- V^+ and V allow to monitor the sensitivity of Z

$$V^+ = \sum_i \mathbb{E}\left[(Z - Z^{(i)})^2 \mathbb{1}_{Z > Z^{(i)}} \mid X_1^n\right]$$

$$V = \sum_i (Z - Z_i)^2 .$$

- $\mathrm{Var}[Z \mid X_i, i \in I] = \mathbb{E}_{X_i \in I}\left[\mathbb{E}_{X_j, j \notin I}\left[(Z - \mathbb{E}[Z \mid X_i, i \in I])^2 \mid X_i, i \in I\right]\right]$

  $I \subseteq \{1, \ldots, n\}$

Efron-Stein inequality

$$\mathrm{Var}[Z] \leq \mathbb{E}[V^+] = \begin{cases} \frac{1}{2}\sum_i \mathbb{E}\left[\left(Z - Z^{(i)}\right)^2\right] = \sum_i \mathrm{Var}[Z \mid X_1^{i-1}, X_{i+1}^n] \\ \sum_i \mathbb{E}\left[\left(Z - Z^{(i)}\right)^2 \mathbb{1}_{Z > Z^{(i)}}\right] \end{cases}$$

As $\mathrm{Var}[Z \mid X_1^{i-1}, X_{i+1}^n] \leq \mathbb{E}[(Z - Z_i)^2]$, for any Z_i that is X_1^{i-1}, X_{i+1}^n-measurable,

$$\mathrm{Var}[Z] \leq \mathbb{E}[V^+] \leq \mathbb{E}[V]\,.$$

☞ E-S inequality is tight: for sums of independent random variables, it becomes an inequality.

☞ E-S inequality is (sometimes) called the Jacknife estimate of variance.

☞ E-S inequality is also called the tensorization property of variance.

☹ E-S inequality may be poor: let $Z = \prod_i X_i$ where X_i are Bernouilli random variables with expectation p. $\mathrm{Var}[Z] = p^n(1 - p^n)$ while the Efron-Stein estimate equals $np^n(1 - p)$!!

Efron-Stein inequality

$$\mathrm{Var}[Z] \leq \mathbb{E}[V^+] \leq \mathbb{E}[V]\,.$$

Proof of Efron-Stein inequality.(I)

w.l.o.g. assume $\mathbb{E}[Z] = 0$.

$$\mathbb{E}\left[Z^2\right] = \mathbb{E}\Big[\big(\sum_{i=1}^{n} \mathbb{E}[Z \mid X_1^i] - \mathbb{E}[Z \mid X_1^{i-1}]\big)^2\Big] \text{ martingale decomposition of } Z$$

$$= \sum_{i=1}^{n} \mathbb{E}\Big[\big(\mathbb{E}[Z \mid X_1^i] - \mathbb{E}[Z \mid X_1^{i-1}]\big)^2\Big] \text{ orthogonality martingale increments}$$

$$= \sum_{i=1}^{n} \mathbb{E}_{X_1^i}\left[\left(\mathbb{E}_{X_{i+1}^n}[Z \mid X_1^i] - \mathbb{E}_{X_{i+1}^n}[\mathbb{E}_{X_i}[Z \mid X_1^{i-1}]]\right)^2\right]$$

Efron-Stein inequality

$$\mathrm{Var}[Z] \leq \mathbb{E}[V^+] \leq \mathbb{E}[V]\,.$$

Proof of Efron-Stein inequality.(I)

w.l.o.g. assume $\mathbb{E}[Z] = 0$.

$$\mathbb{E}\left[Z^2\right] = \mathbb{E}\Big[\Big(\sum_{i=1}^{n} \mathbb{E}[Z \mid X_1^i] - \mathbb{E}[Z \mid X_1^{i-1}]\Big)^2\Big] \text{ martingale decomposition of } Z$$

$$= \sum_{i=1}^{n} \mathbb{E}\Big[\big(\mathbb{E}[Z \mid X_1^i] - \mathbb{E}[Z \mid X_1^{i-1}]\big)^2\Big] \text{ orthogonality martingale increments}$$

(conditional) Jensen inequality, x^2 convex

$$\leq \sum_{i=1}^{n} \mathbb{E}_{X_1^{i-1}, X_{i+1}^n}\Big[\mathbb{E}_{X_i}\big[(Z - \mathbb{E}_{X_i}[Z \mid X_1^{i-1}, X_{i+1}^n])^2 \mid X_1^{i-1}, X_{i+1}^n\big]\Big]$$

$$= \sum_{i=1}^{n} \mathrm{Var}[Z \mid X_1^{i-1}, X_{i+1}^n] = \mathbb{E}[V^+]$$

Efron-Stein inequality

Where does convexity show up ?

Jensen inequality in large spaces.
$\mathcal{X}$: a locally convex Haussdorf topological vector space, with dual $\mathcal{X}^*$

Rockafellar duality Lemma. f: a lower-semi-continuous **convex** function from $\mathcal{X}$ on $(-\infty, \infty]$

if g is defined on $\mathcal{X}^*$ as $g(y) = \sup_{x \in \mathcal{X}} y(x) - f(x)$, then $f(x) = \sup_{y \in \mathcal{X}^*} y(x) - g(y)$

$$
\begin{array}{lrcll}
X \ \ \mathcal{X}\text{-valued r.v.} & \mathbb{E}\Big[f(X)\Big] & = & \mathbb{E}\Big[\sup_{y \in \mathcal{X}^*} \big(y(X) - g(y)\big)\Big] & \text{Rockafellar} \\
& & \geq & \sup_{y \in \mathcal{X}^*} \Big[\mathbb{E}\Big[\big(y(X) - g(y)\big)\Big]\Big] & \\
& & = & \sup_{y \in \mathcal{X}^*} \Big[y\Big(\mathbb{E}[X]\Big) - g(y)\Big] & \\
& & = & f\Big(\mathbb{E}[X]\Big) & \text{Rockafellar.}
\end{array}
$$

Efron-Stein inequality

$$\mathrm{Var}[Z] \leq \mathbb{E}[V^+] \leq \mathbb{E}[V]\,.$$

Proof of Efron-Stein inequality.(II)

Variance maps $\mathbb{L}_2$ toward $\mathbb{R}^+$. Variance is convex !

$$\text{duality} \quad \mathrm{Var}[Z] = \sup_{T \in \mathbb{L}_2^+} \Big[2\mathbb{E}[TZ] - \mathrm{Var}[T] \Big]$$

$$\begin{aligned}
\mathrm{Var}[Z \mid X_2] &= \mathbb{E}_{X_2} \left[\sup_{T \in \mathbb{L}_2(X_1)} \Big\{ 2\mathbb{E}_{X_1}[T(X_1)Z] - \mathrm{Var}[T] \Big\} \right] \\
&\geq \sup_{T \in \mathbb{L}_2(X_1)} \mathbb{E}_{X_2} \Big[2\mathbb{E}_{X_1}[T(X_1)Z] - \mathrm{Var}[T] \Big] \\
&= \sup_{T \in \mathbb{L}_2(X_1)} \mathbb{E}_{X_1} \left[2\mathbb{E}_{X_2}[Z] T_{X_1} - \mathrm{Var}[T] \right] \\
&= \mathrm{Var}[\mathbb{E}[Z \mid X_1]]
\end{aligned}$$

Efron-Stein inequality

$$\mathrm{Var}[Z] \leq \mathbb{E}[V^+] \leq \mathbb{E}[V]\,.$$

Proof of Efron-Stein inequality.(II)

$$\text{duality} \quad \mathrm{Var}[Z] = \sup_{T \in \mathbb{L}_2^+} 2\mathbb{E}[TZ] - \mathrm{Var}[T]$$

$$\rightarrow \mathrm{Var}[Z \mid X_1, \ldots X_{i-1}] \leq \mathrm{Var}[Z \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n]$$

$$\begin{aligned}
\mathrm{Var}[Z] &= \mathbb{E}\left[\left(Z - \mathbb{E}[Z \mid X_1]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}_{X_2}[Z \mid X_1] - \mathbb{E}[Z]\right)^2\right] \\
&= \mathbb{E}_{X_1}\left[\mathbb{E}_{X_2}\left[\left(Z - \mathbb{E}[Z \mid X_1]\right)^2\right]\right] + \mathbb{E}_{X_1}\left[\left(\mathbb{E}_{X_2}[Z \mid X_1] - \mathbb{E}[Z]\right)^2\right] \\
&\leq \mathrm{Var}[Z \mid X_1] + \mathrm{Var}[\mathbb{E}[Z \mid X_1]] \\
&\leq \mathrm{Var}[Z \mid X_1] + \mathrm{Var}[Z \mid X_2]
\end{aligned}$$

Application of Efron-Stein inequality

Back to Random interval packings

Z_i is the size of the largest packing that can be constructed when the ith interval is removed.

- $Z_i = Z$ if the ith interval does not belong to a witness of the value of Z,
- otherwise $Z_i \geq Z - 1$.

$$0 \leq Z - Z_i \leq 1 \quad \text{and} \quad \sum_i (Z - Z_i) \leq Z$$

$$V \leq Z \quad \rightarrow \quad \text{Var}[Z] \leq \mathbb{E}[Z]$$

☞ If $\mathbb{E}[Z] \nearrow \infty$ with n, E-S and Chebyshev imply a law of large numbers for $Z/\mathbb{E}[Z]$.

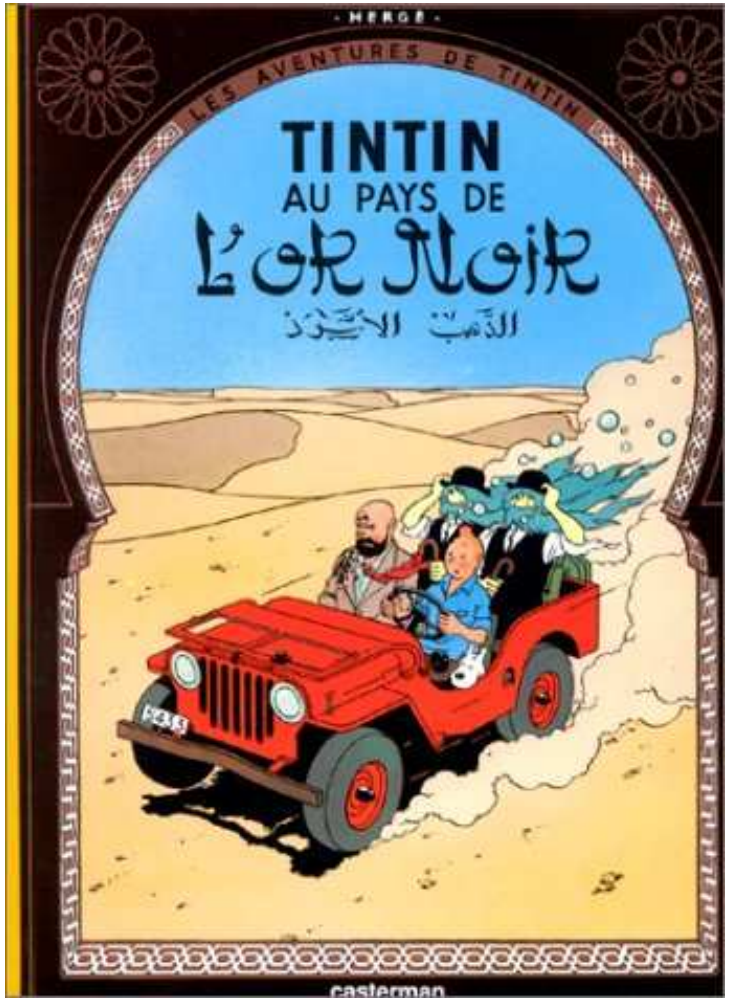
☞ As $n \to$ goes to infinity, $\mathbb{E}[Z]/\sqrt{n} \to 2/\sqrt{\pi}$ while $\text{Var}[Z]/\sqrt{n} \to \frac{2(4-\pi)}{\pi^{3/2}}$
The Efron-Stein estimate is within $\pi/(4-\pi) \approx 3.66$ from truth (asymptotically)!!

☞ $V \leq Z$ also holds for all configuration functions, this includes VC-dimension, fat-shattering dimension, VC-entropy, conditional Rademacher averages.

Lecture II: The entropy method

- From exponential inequalities to the concentration of measure phenomenon
- Concentration inequalities using the entropy method
  - Entropy
  - Tensorization
  - Modified Logarithmic Sobolev inequalities
  - Exponential Efron-Stein inequalities
  - Convex distance inequality
- Learning-theoretical applications
- Moment inequalities using the generalized entropy method

Lecture II: where are we ?



Lecture II: where are we ?

Concentration inequalities are aimed to extend the classical exponential inequalities for sums of independent random variables, to functions of independent random variables.Such functions show up in learning theory as the sumpremum of the deviations between the true risk and the empirical risk, the empirical VC-dimension, the empirical VC entropy, the eigenvalues of the Gram matrix.... They play an important role in the estimation of the generalization error and have proved useful in the design of model selection strategies.
Even when a function cannot be coded as a sum of independent random variables, it may be usually represented as a sum of martingale increments. If the increments happen to be bounded, martingale extensions of the classical inequalities allow (in principle) to derive exponential inequalities for the function of interest. There is indeed a tight relationship between the moments of the function under concern and the moments of the increasing process associated to the martingale representation.
Unfortunately the increasing process is not so easy to analyze. Efron-Stein estimates of the variance of the function of interest have proved to be reasonable surrogates for the increasing process. Indeed, in the second lecture, we will relate the exponential moments of the function under concern and the exponential moments of the Efron-Stein estimates. This will be carried out using the entropy method.
Efron-Stein estimates have often turned out to be surprisingly easy to analyze. This is exemplified by a derivation of Talagrand's convex distance inequality.

Lecture II: roadmap

The entropy of a random variable is defined and elementary properties are presented. Then the Gaussian logarithmic-Sobolev inequality is described. It is shown to imply a tight upper-bound on the log-Laplace transform of smooth functions of Gaussian random variables. A remarkable feature of the Gaussian logarithmic-Sobolev inequality lies in the fact that it is dimension-free.

The derivation of modified logarithmic-Sobolev inequalities in product spaces is carried in two steps: tensorization of entropy is derived by resorting to convexity arguments that proved useful when deriving the Efron-Stein inequality; then an *energy* that will prove to be adequate for general product spaces is defined. The Efron-Stein upper-bound on variance shows up in the right-hand-side of modified logarithmic-Sobolev inequalities.

Under various integrability conditions of the Efron-Stein upper-bound of variance, Bernstein-like inequalities for *functions of many independent random variables that depend not too much on each of them* are derived.

Bernstein-like inequalities for self-bounded functionals are derived. A derivation of Talagrand's convex distance inequality is finally presented.

Gaussian tails from logarithmic Sobolev inequalities

To get exponential tail bounds for Z
➔ control $\mathbb{E}[\exp(\lambda Z)]$ or $\log \mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))]$.

Control over log-Laplace transform of smooth functions of Gaussian random variables can be obtained from functional inequalities known as logarithmic Sobolev inequalities.

 Entropy of a (positive) function $f > 0$ with respect to a probability distribution

$$\mathrm{Ent}(f) \triangleq \mathbb{E}\big[f \log f\big] - \mathbb{E}[f] \log \mathbb{E}[f] = \mathbb{E}[\Phi(f)] - \Phi(\mathbb{E}[f]) \quad \text{with } \Phi(x) = x \log x \ .$$

Jensen inequality ➔ $\mathrm{Ent}(f) \geq 0$.

Gaussian tails from logarithmic Sobolev inequalities

To get exponential tail bounds for Z
➔ control $\mathbb{E}[\exp(\lambda Z)]$ or $\log \mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))]$.

Control over log-Laplace transform of smooth functions of Gaussian random variables can be obtained from functional inequalities known as logarithmic Sobolev inequalities.

 Entropy of a (positive) function $f > 0$ with respect to a probability distribution

$$\mathrm{Ent}(f) \triangleq \mathbb{E}\big[f \log f\big] - \mathbb{E}[f] \log \mathbb{E}[f] = \mathbb{E}[\Phi(f)] - \Phi(\mathbb{E}[f]) \quad \text{with } \Phi(x) = x \log x \ .$$

Jensen inequality ➔ $\mathrm{Ent}(f) \geq 0$.

$$\lim_{a \nearrow \infty} \mathrm{Ent}[(f + a)^2] = 4\mathrm{Var}[f] \ .$$

Gaussian tails from logarithmic Sobolev inequalities

To get exponential tail bounds for Z
➔ control $\mathbb{E}[\exp(\lambda Z)]$ or $\log \mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))]$.

Control over log-Laplace transform of smooth functions of Gaussian random variables can be obtained from functional inequalities known as logarithmic Sobolev inequalities.

Entropy of a (positive) function $f > 0$ with respect to a probability distribution

$$\mathrm{Ent}(f) \triangleq \mathbb{E}\big[f \log f\big] - \mathbb{E}[f] \log \mathbb{E}[f] = \mathbb{E}[\Phi(f)] - \Phi(\mathbb{E}[f]) \quad \text{with } \Phi(x) = x \log x \ .$$

Jensen inequality ➔ $\mathrm{Ent}(f) \geq 0$.

Gaussian Logarithmic-Sobolev Inequality (Gross, 1975)

$$\mathrm{Ent}(f^2) \leq 2\mathbb{E}\big[|| \bigtriangledown f||^2\big]$$

Entropy — Energy

Gaussian tails from logarithmic Sobolev inequalities

To get exponential tail bounds for Z
➔ control $\mathbb{E}[\exp(\lambda Z)]$ or $\log \mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))]$.

Control over log-Laplace transform of smooth functions of Gaussian random variables can be obtained from functional inequalities known as logarithmic Sobolev inequalities.

Entropy of a (positive) function $f > 0$ with respect to a probability distribution

$$\mathrm{Ent}(f) \triangleq \mathbb{E}\big[f \log f\big] - \mathbb{E}[f] \log \mathbb{E}[f] = \mathbb{E}[\Phi(f)] - \Phi(\mathbb{E}[f]) \quad \text{with } \Phi(x) = x \log x \ .$$

Jensen inequality ➔ $\mathrm{Ent}(f) \geq 0$.

Gaussian Logarithmic-Sobolev Inequality (Gross, 1975)

$$\mathrm{Ent}(f^2) \leq 2\mathbb{E}\big[\| \bigtriangledown f\|^2\big] \quad \rightarrow \quad \mathrm{Ent}\big(e^{\lambda f}\big) \leq \frac{\lambda^2}{2}\mathbb{E}\big[\| \bigtriangledown f\|^2 e^{\lambda f}\big]..$$

☞ This step critically depends on the chain rule for derivation.

Gaussian tails from logarithmic Sobolev inequalities

Herbst argument: derive a differential inequality satisfied by the log-Laplace transform.

Assumption: $\| \bigtriangledown f \|_\infty \leq 1$.

Notation : $H(\lambda) = \frac{1}{\lambda} \log \mathbb{E}\left[e^{\lambda Z} \right]$

$$\frac{1}{\lambda^2} \mathrm{Ent}(e^{\lambda Z}) = H'(\lambda) \mathbb{E}\left[e^{\lambda Z} \right] \quad \text{and} \quad \lim_{\lambda \to 0^+} H(\lambda) = \mathbb{E}[Z]$$

$$\begin{aligned} H'(\lambda) &\leq \frac{1}{2} && \text{Gaussian logarithmic-Sobolev inequality!} \\ H(\lambda) - H(0^+) &\leq \frac{\lambda}{2} && \text{Integration} \\ \log \mathbb{E}\left[e^{\lambda (Z - \mathbb{E}[Z])} \right] &\leq \frac{\lambda^2}{2} && \text{Rewriting.} \end{aligned}$$

$$\mathbb{P}\left\{ f(X_1, \ldots, X_n) \geq \mathbb{E}[f] + t \right\} \leq \exp\left(- \frac{t^2}{2 \| \bigtriangledown f \|_\infty^2} \right)$$

Gaussian tails from logarithmic Sobolev inequalities

Herbst argument: derive a differential inequality satisfied by the log-Laplace transform.

Assumption: $\| \bigtriangledown f \|_\infty \leq 1$.

Notation : $H(\lambda) = \frac{1}{\lambda} \log \mathbb{E}\left[e^{\lambda Z}\right]$

$$\frac{1}{\lambda^2} \mathsf{Ent}(e^{\lambda Z}) = H'(\lambda) \mathbb{E}\left[e^{\lambda Z}\right] \quad \text{and} \quad \lim_{\lambda \to 0^+} H(\lambda) = \mathbb{E}[Z]$$

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

$$\mathbb{P}\left\{f(X_1, \ldots, X_n) \geq \mathbb{E}[f] + t\right\} \leq \exp\left(-\frac{t^2}{2\| \bigtriangledown f \|_\infty^2}\right)$$

☞ Any smooth function of a Gaussian random variable enjoys Gaussian tail-behavior. The dimension of the Gaussian vector does not appear in tail-behavior.

The Gaussian concentration inequality is dimension-free.

Variation on Herbst's argument

☞ The Gaussian logarithmic Sobolev inequality provides with concentration inequalities when $\| \bigtriangledown F\|^2$ is (somewhat) exponentially integrable.

Decoupling inequality $\mathbb{E}\left[\lambda W e^{\lambda Z}\right] \leq \mathrm{Ent}\left[e^{\lambda Z}\right] + \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda W}\right]$

Variational characterization of Entropy: $\mathbb{E}\left[\frac{f}{\mathbb{E}[f]} g\right] \leq \mathrm{Ent}\left[\frac{f}{\mathbb{E}[f]}\right] + \log \mathbb{E}\left[e^{g}\right]$

Applying the decoupling inequality to the Gaussian logarithmic Sobolev inequality:

$$\mathrm{Ent}\left[e^{\lambda f}\right] \leq \frac{\theta\lambda}{2}\left(\mathrm{Ent}\left[e^{\lambda f}\right] + \mathbb{E}\left[e^{\lambda f}\right] \log \mathbb{E}\left[e^{\frac{\lambda \| \bigtriangledown f\|^2}{\theta}}\right]\right)$$

$$\downarrow$$

$$\frac{1}{\lambda^2} \frac{\mathrm{Ent}\left[e^{\lambda f}\right]}{\mathbb{E}\left[e^{\lambda f}\right]} \leq \frac{\theta}{2\lambda(1-\lambda\theta/2)} \log \mathbb{E}\left[e^{\frac{\lambda \| \bigtriangledown f\|^2}{\theta}}\right]$$

Variation on Herbst's argument

Notations:

$$H(\lambda) = \frac{1}{\lambda} \log \mathbb{E}\left[e^{\lambda f}\right] \quad \text{and} \quad G(\lambda) = \log \mathbb{E}\left[e^{\lambda \|\nabla f\|^2}\right],$$

$$\frac{1}{\lambda^2} \frac{\text{Ent}\left[e^{\lambda f}\right]}{\mathbb{E}\left[e^{\lambda f}\right]} \leq \frac{\theta}{2\lambda(1-\lambda\theta/2)} \log \mathbb{E}\left[e^{\frac{\lambda \|\nabla f\|^2}{\theta}}\right]$$

translates into:

$$H'(\lambda) \leq \frac{\theta}{2\lambda(1-\lambda\theta/2)} G(\lambda/\theta)$$

Convexity of G ➔ RHS is non-decreasing with respect to λ

$$\begin{aligned} H(\lambda) - H(0) &\leq \int_{s=0}^{\lambda} \frac{\theta}{2s(1-s\theta/2)} G(s/\theta) ds \\ &\leq \frac{\theta}{2(1-\lambda\theta/2)} G(\lambda/\theta) \end{aligned}$$

Variation on Herbst's argument

$$\log \mathbb{E}\left[e^{\lambda(f(X)-\mathbb{E}[f])}\right] \leq \frac{\lambda\theta}{2(1-\lambda\theta/2)} \log \mathbb{E}\left[e^{\lambda\|\nabla f\|^2/\theta}\right]$$

If $\| \bigtriangledown f\|_\infty \leq L$, taking θ toward 0, we recover sub-Gaussian behavior.

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

If $\| \bigtriangledown f\| = C\|X\|$ (sub-quadratic function), $\log \mathbb{E}\left[e^{\lambda\|\nabla f\|^2/\theta}\right] < \infty$ for a non-trivial range of values of λ.

☞ In the sequel, we will try to reproduce this line of reasoning for product distributions.

Toward modified Logarithmic Sobolev inequalities

 The Gaussian logarithmic Sobolev inequality is (almost) a characterization of the Gaussian distribution.

 In order to derive tail bounds for other kinds of distributions using analogues of Herbst argument, we need to develop:

- modified logarithmic Sobolev inequalities tailored to the distributions we have in mind (exponential, Poisson, product distributions,...)
- analogues of Herbst argument.

☞ In Learning Theory we are interested in (modified) logarithmic Sobolev inequalities for product measures

Toward modified Logarithmic Sobolev inequalities

The Gaussian logarithmic Sobolev inequality is (almost) a characterization of the Gaussian distribution.

In order to derive tail bounds for other kinds of distributions using analogues of Herbst argument, we need to develop:

- modified logarithmic Sobolev inequalities tailored to the distributions we have in mind (exponential, Poisson, product distributions,...)
- analogues of Herbst argument.

☞ In Learning Theory we are interested in (modified) logarithmic Sobolev inequalities for product measures

General form: $\mathrm{Ent}(f^2) \leq \mathcal{E}[f]$,

where $\mathcal{E}[f]$ is an energy-like functional.

Toward modified Logarithmic Sobolev inequalities

The Gaussian logarithmic Sobolev inequality is (almost) a characterization of the Gaussian distribution.

In order to derive tail bounds for other kinds of distributions using analogues of Herbst argument, we need to develop:

- modified logarithmic Sobolev inequalities tailored to the distributions we have in mind (exponential, Poisson, product distributions,...)
- analogues of Herbst argument.

☞ In Learning Theory we are interested in (modified) logarithmic Sobolev inequalities for product measures

General form: $\mathrm{Ent}(f^2) \leq \mathcal{E}[f],$

where $\mathcal{E}[f]$ is an energy-like functional.

Balanced Bernouilli Log-Sobolev Inequality (Gross, 1975)

$$\mathrm{Ent}(f^2) \leq \frac{1}{2}\mathbb{E}\left[(Df)^2\right] \quad \text{with } Df \triangleq f(1) - f(-1).$$

Tensorization of entropy

Derivation of modified LS inequalities proceeds in two steps:

- Tensorization. (analogue of Efron-Stein inequality)

Notation $\mathrm{Ent}[Z \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots X_n]$

$$= \mathbb{E}_{X_1,\ldots,X_{i-1},X_{i+1},\ldots X_n}\Big[\mathrm{Ent}[Z(x_1, \ldots, x_{i-1}, X_i, x_{i+1}, \ldots x_n)]\Big]$$

$$\mathrm{Ent}[Z] \leq \sum_{i=1}^{n} \mathrm{Ent}[Z \mid X_1^{i-1}, X_{i+1}^{n}]$$

- Upper-bounding of entropies with respect to single random variables.

.

$$\log x \leq x - 1 \quad > \quad -\mathbb{E}[Y] \log \mathbb{E}[Y] \leq -\mathbb{E}[Y] \log u - \mathbb{E}[Y] + u \quad \forall u > 0.$$

Tensorization of entropy

The tensorization step is generic.
Its relies on a convexity property of the functional $Z \mapsto \mathrm{Ent}[Z]$ on $\mathbb{L}_+^1$.
Not to be confused with the convexity of $x \mapsto x \log x$.

☞ A representation formula for Entropy:

$$\mathrm{Ent}[Z] = \sup_{T \in \mathbb{L}_1^+} \mathbb{E}\left[Z \log \frac{T}{\mathbb{E}[T]} \right]$$

$$\mathrm{Ent}[Z \mid X_2] \geq \mathrm{Ent}[\mathbb{E}[Z \mid X_1]] \quad \text{2 variables}$$

$$\mathrm{Ent}[Z \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n] \geq \mathrm{Ent}[\mathbb{E}[Z \mid X_1, \ldots, X_{i-1}] \mid X_{i+1}, \ldots X_n]$$

The conditional distribution of Z with respect $X_1, \ldots, X_{i-1}$ is a convex combination of conditional distributions with respect to $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$

Tensorization of entropy

$$
\begin{aligned}
&\mathrm{Ent}[Z(X_1, X_2)] \\
&= \sup_T \mathbb{E}\left[Z \log \frac{T(X_1, X_2)}{\mathbb{E}[T]} \right] \\
&= \sup_T \mathbb{E}\left[Z \log \frac{T}{\mathbb{E}_{X_2}[T]} + Z \log \frac{\mathbb{E}_{X_2}[T]}{\mathbb{E}[T]} \right] \\
&\leq \sup_T \mathbb{E}\left[Z \log \frac{T}{\mathbb{E}_{X_2}[T]} \right] + \sup_T \mathbb{E}_{X_2}\left[\mathbb{E}_{X_1}\left[Z \log \frac{\mathbb{E}_{X_2}[T]}{\mathbb{E}[T]} \right] \right] \\
&\leq \mathbb{E}_{X_1}\left[\sup_{T \in \mathbb{L}_1^+(X_2)} \mathbb{E}_{X_2}\left[Z \log \frac{T}{\mathbb{E}_{X_2}[T]} \right] \right] + \mathbb{E}_{X_2} \sup_{T \in \mathbb{L}_1^+(X_1)} \left[\mathbb{E}_{X_1}\left[Z \log \frac{T}{\mathbb{E}_{X_1}[T]} \right] \right] \\
&\leq \mathrm{Ent}[Z \mid X_1] + \mathrm{Ent}[Z \mid X_2] \,.
\end{aligned}
$$

The general formula follows by induction on the number of variables.

$$
\underbrace{X_1, \ldots, X_{n-1}}_{Y_1} \qquad \underbrace{X_n}_{Y_2}
$$

Bounding entropy with respect to one random variable

When dealing with the variance, the proof of the Efron-Stein inequality reduced to the proof of the tensorization property of the variance. Here, the tensorization property of entropy leaves us with a sum of one-dimensional conditional entropies.

More work is needed in order to get an energy-like term.

$$\log x \leq x - 1 \quad \rightarrow \quad -\mathbb{E}[Y]\log\mathbb{E}[Y] \leq -\mathbb{E}[Y]\log u - \mathbb{E}[Y] + u \quad \forall u > 0.$$

$$\mathrm{Ent}[f(X)] \quad \leq \quad \mathbb{E}\left[f(X)\log\frac{f(X)}{u} - (f(X) - u)\right] \quad \forall u > 0\,.$$

Bounding entropy with respect to one random variable

$\log x \leq x - 1 \quad \rightarrow \quad -\mathbb{E}[Y] \log \mathbb{E}[Y] \leq -\mathbb{E}[Y] \log u - \mathbb{E}[Y] + u \quad \forall u > 0.$

$$\mathrm{Ent}[f(X)] \quad \leq \quad \mathbb{E}\left[f(X) \log \frac{f(X)}{u} - (f(X) - u) \right] \quad \forall u > 0\,.$$

Let g_i denote a function of the $n-1$ random variables: X_1^{i-1}, X_{i+1}^n.

$$\mathrm{Ent}[f \mid X_1^{i-1}, X_{i+1}^n] \leq \mathbb{E}_{X_1^{i-1}, X_{i+1}^n} \left[\mathbb{E}_{X_i} \left[f \log \frac{f}{g_i} - (f - g_i) \mid X_1^{i-1}, X_{i+1}^n \right] \right]$$

Bounding entropy with respect to one random variable

$$\log x \leq x - 1 \quad \rightarrow \quad -\mathbb{E}[Y]\log\mathbb{E}[Y] \leq -\mathbb{E}[Y]\log u - \mathbb{E}[Y] + u \quad \forall u > 0.$$

$$\mathrm{Ent}[f(X)] \quad \leq \quad \mathbb{E}\left[f(X)\log\frac{f(X)}{u} - (f(X) - u)\right] \quad \forall u > 0\,.$$

Let g denote a function of $n-1$ random variables.
Combining with the tensorization property of Entropy:

$$\mathrm{Ent}[f] \quad \leq \quad \sum_i \mathbb{E}_{X_1^{i-1}, X_{i+1}^n}\left[\mathbb{E}_{X_i}\left[f\log\frac{f}{g} - (f - g) \quad | \quad X_1^{i-1}, X_{i+1}^n\right]\right]$$

☞ entails the Gaussian logarithmic Sobolev inequality, the Poissonian logarithmic Sobolev inequality...

Bounding entropy with respect to one random variable

$$\log x \leq x - 1 \quad \rightarrow \quad -\mathbb{E}[Y]\log\mathbb{E}[Y] \leq -\mathbb{E}[Y]\log u - \mathbb{E}[Y] + u \quad \forall u > 0.$$

$$\mathrm{Ent}[f(X)] \quad \leq \quad \mathbb{E}\left[f(X)\log\frac{f(X)}{u} - (f(X) - u)\right] \quad \forall u > 0\,.$$

Specializing to functions $f(X_1^n) = e^{\lambda Z}$, letting $g(X_1^{i-1}, X_{i+1}^n) = e^{\lambda Z_i}$,

$$\begin{aligned}
\mathrm{Ent}\left[e^{\lambda Z}\right] &\leq \sum_i \mathbb{E}_{X_1^{i-1}, X_{i+1}^n}\left[\mathbb{E}_{X_i}\left[e^{\lambda Z}\left(\lambda Z - \lambda Z_i\right) - \left(e^{\lambda Z} - e^{\lambda Z_i}\right)\right]\right] \\
&\leq \sum_i \mathbb{E}_{X_1^{i-1}, X_{i+1}^n}\left[\mathbb{E}_{X_i}\left[e^{\lambda Z}\left(e^{\lambda(Z_i - Z)} - \lambda(Z_i - Z) - 1\right)\right]\right] \\
&\leq \sum_i \mathbb{E}_{X_1^{i-1}, X_{i+1}^n}\left[\mathbb{E}_{X_i}\left[e^{\lambda Z}\tau^*(\lambda(Z_i - Z))\right]\right]
\end{aligned}$$

Modified Log-Sobolev Inequalities (I)

$$\text{Ent}\left[e^{\lambda Z}\right] \leq \mathbb{E}\left[e^{\lambda Z} \sum_i \tau^*(\lambda(Z_i - Z))\right]$$

where Z_i is a measurable function of $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$,
and $\tau^*(x) \stackrel{\Delta}{=} e^x - x - 1$.

Modified Log-Sobolev Inequalities (I)

$$\text{Ent}\left[e^{\lambda Z}\right] \leq \mathbb{E}\left[e^{\lambda Z} \sum_i \tau^*(\lambda(Z_i - Z))\right]$$

where Z_i is a measurable function of $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, and $\tau^*(x) \stackrel{\Delta}{=} e^x - x - 1$.

If $Z \geq Z_i$ and $\lambda \geq 0$ as $e^x - x - 1 \leq x^2/2$ for $x \leq 0$

$$\text{Ent}\left[e^{\lambda Z}\right] \leq \frac{\lambda^2}{2}\mathbb{E}\left[e^{\lambda Z}\sum_i (Z - Z_i)^2\right]$$

Let $V = \sum_i (Z - Z_i)^2$,

$$\text{Ent}\left[e^{\lambda Z}\right] \leq \frac{\lambda^2}{2}\mathbb{E}\left[V e^{\lambda Z}\right] \rightarrow \log \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] \leq \frac{\lambda\theta}{2(1 - \lambda\theta/2)} \log \mathbb{E}\left[e^{\lambda V/\theta}\right]$$

☞ Not only the second moment of Z is related to some moment of V, (Efron-Stein) but also the log-Laplace transforms are connected.

Modified Log-Sobolev Inequalities (I)

$$\mathrm{Ent}\left[e^{\lambda Z}\right] \leq \mathbb{E}\left[e^{\lambda Z} \sum_i \tau^*(\lambda(Z_i - Z))\right]$$

Self-bounded functions

- If $Z_i \leq Z \leq Z_i + 1$, as τ^* is convex:

$$\begin{aligned} \tau^*(-\lambda(Z - Z_i)) &\leq (1 + Z - Z_i)\tau(0) + \tau^*(-\lambda)\,(Z - Z_i) \\ &= \tau^*(-\lambda)\,(Z - Z_i) \end{aligned}$$

$$\mathrm{Ent}\left[e^{\lambda Z}\right] \leq \frac{\lambda^2}{2}\mathbb{E}\left[e^{\lambda Z}\tau^*(-\lambda)\left(\sum_i Z - Z_i\right)\right]$$

- If furthermore $\sum_i (Z - Z_i) \leq Z$ $\quad \mathrm{Ent}\left[e^{\lambda Z}\right] \leq \tau^*(-\lambda)\mathbb{E}\left[Z e^{\lambda Z}\right]$

Modified Log-Sobolev Inequalities (I)

Stating and solving the differential inequality:

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

multiplying by $\exp(-\lambda \mathbb{E}[Z])$

regrouping

dividing by $F(\lambda) = \mathbb{E}\left[\exp\left(\lambda(Z - \mathbb{E}[Z])\right)\right]$

$$(\lambda - \tau^*(\lambda))\frac{F'(\lambda)}{F(\lambda)} - \log F(\lambda) \leq \tau^*(\lambda)\mathbb{E}[Z]$$

Note that $\mathbb{E}[Z]\tau^*(\lambda)$ is a solution of the differential equation

$$(\lambda - \tau^*(\lambda))f'(\lambda) - f(\lambda) = \tau^*(\lambda)\mathbb{E}[Z]$$

One can check that $\log F(\lambda) = \mathbb{E}[Z]\tau^*(\lambda)$ is the largest solution of the differential inequality, by showing that for all solutions $h(\lambda) = (e^\lambda - 1)g(\lambda)$ of

$$(\lambda - \tau^*(\lambda))h'(\lambda) - h(\lambda) \leq 0$$

are non-positive.

Modified Log-Sobolev Inequalities (I)

Self-bounded functions

- If $Z_i \leq Z \leq Z_i + 1$, as τ^* is convex:

$$\begin{aligned} \tau^*(-\lambda(Z - Z_i)) &\leq (1 + Z - Z_i)\tau(0) + \tau^*(-\lambda)\,(Z - Z_i) \\ &= \tau^*(-\lambda)\,(Z - Z_i) \end{aligned}$$

$$\mathrm{Ent}\Big[e^{\lambda Z}\Big] \leq \frac{\lambda^2}{2}\mathbb{E}\Big[e^{\lambda Z}\tau^*(-\lambda)\big(\sum_i Z - Z_i\big)\Big]$$

- If furthermore $\sum_i (Z - Z_i) \leq Z$ $\quad \mathrm{Ent}\Big[e^{\lambda Z}\Big] \leq \tau^*(-\lambda)\mathbb{E}\Big[Ze^{\lambda Z}\Big]$

$$\log \mathbb{E}\Big[e^{\lambda(Z - \mathbb{E}[Z])}\Big] \leq \mathbb{E}[Z]\tau^*(\lambda)$$

$$\mathbb{P}\big\{Z \geq \mathbb{E}[Z] + t\big\} \leq \exp\Big(-\frac{t^2}{2(\mathbb{E}[Z] + t/3)}\Big).$$

Modified Log-Sobolev Inequalities (II)

Rather than Z_i, we may consider $Z^{(i)}$, where
$Z^{(i)} = f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$
☞ Z and $Z^{(i)}$ are identically distributed.

$$
\begin{aligned}
\mathrm{Ent}\left[e^{\lambda Z}\right] &\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z}\tau^*(-\lambda(Z - Z^{(i)}))\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z}\tau^*(-\lambda(Z - Z^{(i)}))\mathbb{1}_{Z>Z^{(i)}}\right] \\
&\quad + \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z}\tau^*(-\lambda(Z - Z^{(i)}))\mathbb{1}_{Z<Z^{(i)}}\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z}\tau^*(-\lambda(Z - Z^{(i)}))\mathbb{1}_{Z>Z^{(i)}}\right] \\
&\quad + \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z^{(i)}}\tau^*(-\lambda(Z^{(i)} - Z))\mathbb{1}_{Z>Z^{(i)}}\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z}\left(e^{-\lambda(Z-Z^{(i)})} - 1\right)\lambda\big(Z^{(i)} - Z\big)\mathbb{1}_{Z>Z^{(i)}}\right]
\end{aligned}
$$

Modified Log-Sobolev Inequalities (II)

Symmetrized version of modified logarithmic Sobolev inequalities.

$$\text{Ent}\left[e^{\lambda Z}\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z} \psi(-\lambda(Z - Z^{(i)})) \mathbb{1}_{Z > Z^{(i)}}\right]$$

where $Z^{(i)}$ is a measurable function of $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, and X_i' while

$$\psi(x) = e^x \tau^*(-x) + \tau^*(x) = x(e^x - 1).$$

$\psi()$ is convex and for $x > 0$, $\psi(-x) \leq x^2$,

$$\text{for } \lambda > 0 \quad \text{Ent}\left[e^{\lambda Z}\right] \leq \lambda^2 \mathbb{E}\left[e^{\lambda Z} V^+\right]$$

Where $\mathbb{E}[V^+]$ is the Efron-Stein upper-bound on Var$[Z]$.

Modified Log-Sobolev Inequalities (II)

For $x \leq 0$, $\psi(x) \leq x^2$. The modified logarithmic Sobolev inequality implies that for $\lambda \geq 0$:

$$\begin{aligned}\mathrm{Ent}\left[e^{\lambda Z}\right] &\leq \lambda^2 \mathbb{E}\left[e^{\lambda Z} V_+\right] \\ &\leq \theta\lambda\left(\mathrm{Ent}\left[e^{\lambda Z}\right] + \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda V_+/\theta}\right]\right).\end{aligned}$$

which translates into:

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

☞ This has exactly the same form as the inequality encountered while carrying the extended Herbst argument !

Modified Log-Sobolev Inequalities (II)

An exponential Efron-Stein inequality. $\theta > 0$ and $\lambda \in (0, 1/\theta)$,

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

The exponential integrability of Z is actually related to the exponential integrability of the Efron-Stein estimate(s) of variance. If we can bound $\log \mathbb{E}\left[\exp\left(\frac{\lambda V_+}{\theta}\right)\right]$, then we can use an exponential Markov inequality in order to get a Bernstein-like inequality for Z.

Modus operandi: check whether

- V or V is constant.
- V or V^+ is upper-bounded by $aZ + b$,
- V or V^+ is simpler than Z.

Done or to be done for:

- $\sup_f R[f] - R_{\text{emp}}(f)$.
- Empirical VC or fat-shattering dimension.
- Empirical VC entropy.
- Number of support vectors.

Control by convex distance

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

$$d_\alpha(x_1^n, A) = \inf_{y_1^n \in A} d_\alpha(x_1^n, y_1^n)$$

$$= \inf_{y_1^n \in A} \sum_{i: x_i \neq y_i} |\alpha_i|$$

$$d_T(x_1^n, A) = \sup_{\alpha \in [0,\infty)^n : \|\alpha\| \leq 1} d_\alpha(x_1^n, A)$$ Talagrand's convex distance.

Control by convex distance

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

$$d_\alpha(x_1^n, A) = \inf_{y_1^n \in A} d_\alpha(x_1^n, y_1^n)$$

$$= \inf_{y_1^n \in A} \sum_{i: x_i \neq y_i} |\alpha_i|$$

$$d_T(x_1^n, A) = \sup_{\alpha \in [0,\infty)^n : \|\alpha\| \leq 1} d_\alpha(x_1^n, A)$$ Talagrand's convex distance.

- If $\mathbb{P}[X_1^n \in A] \geq 1/2, \qquad \mathbb{P}\left[d_T(X_1^n, A) \geq t\right] \leq 2e^{-t^2/4}$

- Control by convex distance provides sharp deviation inequalities around the median, for
  1. *configuration functions,*
  2. euclidean traveling salesman,
  3. bin-packing,
  4. minimum spanning trees.

Control by convex distance

Control by convex distance can be recovered from exponential Efron-Stein inequalities.

 $d_T(\cdot,\cdot)$ can be represented as a saddle point. $\mathcal{M}(A)$: set of probabilities on A.

$$
\begin{aligned}
d_T(X_1^n, A) &= \inf_{\nu \in \mathcal{M}(A)} \sup_{\alpha : \|\alpha\|_2 \leq 1} \sum_j \alpha_j \mathbb{E}_\nu [\mathbb{1}_{X_j \neq Y_j}] \\
&= \sup_{\alpha : \|\alpha\|_2 \leq 1} \inf_{\nu \in \mathcal{M}(A)} \sum_j \alpha_j \mathbb{E}_\nu [\mathbb{1}_{X_j \neq Y_j}]
\end{aligned}
$$

Control by convex distance

$d_T(\cdot,\cdot)$ can be represented as a saddle point. $\mathcal{M}(A)$: set of probabilities on A.

$$
\begin{aligned}
d_T(X_1^n, A) &= \inf_{\nu\in\mathcal{M}(A)} \sup_{\alpha:\|\alpha\|_2\leq 1} \sum_j \alpha_j \mathbb{E}_\nu[\mathbb{1}_{X_j\neq Y_j}] \\
&= \sup_{\alpha:\|\alpha\|_2\leq 1} \inf_{\nu\in\mathcal{M}(A)} \sum_j \alpha_j \mathbb{E}_\nu[\mathbb{1}_{X_j\neq Y_j}]
\end{aligned}
$$

Sion Minimax Theorem $f : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$

convex and lower-semi-continuous with respect to x

concave and upper-semi-continuous with respect to y

$\mathcal{X}$ convex and compact

$$
\inf_x \sup_y f(x,y) = \sup_y \inf_x f(x,y) = \min_x \sup_y f(x,y)\,.
$$

Control by convex distance

$(\widehat{\nu}, \widehat{\alpha})$: a saddle point for X_1^n.

$$Z^{(i)} = \inf_{\nu \in \mathcal{M}(A)} \sup_{\alpha} \sum_j \alpha_j \mathbb{E}_\nu[\mathbb{1}_{X_j^{(i)} \neq Y_j}] \geq \inf_{\nu \mathcal{M}(A)} \sum_j \widehat{\alpha}_j \mathbb{E}_\nu[\mathbb{1}_{X_j^{(i)} \neq Y_j}] .$$

$\tilde{\nu}$: distribution on A that achieves the infimum.

$$Z = \inf_\nu \sum_j \widehat{\alpha}_j \mathbb{E}_\nu[\mathbb{1}_{X_j \neq Y_j}] \leq \sum_j \widehat{\alpha}_j \mathbb{E}_{\tilde{\nu}}[\mathbb{1}_{X_j \neq Y_j}] .$$

$$Z - Z^{(i)} \leq \sum_j \widehat{\alpha}_j \mathbb{E}_{\tilde{\nu}}[\mathbb{1}_{X_j \neq Y_j} - \mathbb{1}_{X_j^{(i)} \neq Y_j}] = \widehat{\alpha}_i \mathbb{E}_{\tilde{\nu}}[\mathbb{1}_{X_i \neq Y_i} - \mathbb{1}_{X_i^{(i)} \neq Y_i}] \leq \widehat{\alpha}_i \ .$$

$$V_+ \leq \sum_i \widehat{\alpha}_i^2 = 1.$$

$\mathrm{Var}[d_T(X_1^n, A)] \leq 1$ Efron-Stein inequality !

$\mathbb{P}\left[d_T(X_1^n, A) - \mathbb{E}d_T(X_1^n, A) \geq t\right] \leq e^{-t^2/4}$ exponential Efron-Stein inequality !.

Control by convex distance

$$\mathrm{Var}[d_T(X_1^n, A)] \leq 1 \quad \text{Efron-Stein inequality !}$$

$$\mathbb{P}\left[d_T(X_1^n, A) - \mathbb{E}d_T(X_1^n, A) \geq t\right] \leq e^{-t^2/4} \quad \text{exponential Efron-Stein inequality !.}$$

$$\mathbb{P}[d_T(X_1^n, A) - \mathbb{E}d_T(X_1^n, A) \leq -t] \leq \frac{\mathrm{Var}[d_T(X_1^n, A)]}{t^2} \leq \frac{1}{t^2} \quad \text{Chebyshev inequality}$$

$$\mathbb{P}[A] \leq \frac{1}{\left(\mathbb{E}[d_T(\cdot, A)]\right)^2}$$

Control by convex distance

$$\mathrm{Var}[d_T(X_1^n, A)] \leq 1 \quad \text{Efron-Stein inequality !}$$

$$\mathbb{P}\left[d_T(X_1^n, A) - \mathbb{E}d_T(X_1^n, A) \geq t\right] \leq e^{-t^2/4} \quad \text{exponential Efron-Stein inequality !.}$$

$$\mathbb{P}[d_T(X_1^n, A) - \mathbb{E}d_T(X_1^n, A) \leq -t] \leq \frac{\mathrm{Var}[d_T(X_1^n, A)]}{t^2} \leq \frac{1}{t^2} \quad \text{Chebyshev inequality}$$

$$\mathbb{E}d_T(X_1^n, A) \leq \frac{1}{\sqrt{\mathbb{P}[A]}}$$

$$\mathbb{P}[d_T(X_1^n, A) \geq t + \sqrt{2}] \leq e^{-t^2/4} \quad \text{if } \mathbb{P}[A] \geq 1/2$$

Control by convex distance

$$\mathrm{Var}[d_T(X_1^n, A)] \leq 1 \quad \text{Efron-Stein inequality !}$$

$$\mathbb{P}\left[d_T(X_1^n, A) - \mathbb{E}d_T(X_1^n, A) \geq t\right] \leq e^{-t^2/4} \quad \text{exponential Efron-Stein inequality !.}$$

$$\mathbb{P}[d_T(X_1^n, A) - \mathbb{E}d_T(X_1^n, A) \leq -t] \leq \frac{\mathrm{Var}[d_T(X_1^n, A)]}{t^2} \leq \frac{1}{t^2} \quad \text{Chebyshev inequality}$$

$$\mathbb{E}d_T(X_1^n, A) \leq \frac{1}{\sqrt{\mathbb{P}[A]}}$$

$$t > \sqrt{2} > (t - \sqrt{2})^2 \geq t^2/2 - 4\log 2$$

$$\mathbb{P}[d_T(X_1^n, A) \geq t] \leq 2e^{-t^2/8} \quad \text{for } t > \sqrt{2} \text{ and } \mathbb{P}[A] \geq 1/2$$

Plan III : Statistical learning applications

- From exponential inequalities to the concentration of measure phenomenon
- Concentration inequalities using the entropy method
- Learning-theoretical applications
  - Fat-shattering VC-dimension
  - VC entropy
  - Conditional Rademacher averages
  - Supremum of empirical processes
  - ...
- Moment inequalities using the generalized entropy method

Where are we?

We hope (and have some evidence) that we have developped valuable tools for deriving tail bounds for general functions of independent variables.

$$
\begin{aligned}
Z &= f(X_1, \ldots, X_n) \\
Z^{(i)} &= f(X_1, \ldots, X_i', \ldots, X_n) \\
Z_i &= f_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\
V^+ &= \mathbb{E}'\Big[\sum_i (Z - Z^{(i)})^2 \mathbb{1}_{Z > Z^{(i)}} \mid X_1^n\Big] \\
V &= \sum_i (Z - Z_i)^2
\end{aligned}
$$

$$
\mathrm{Var}[Z] \leq \mathbb{E}[V^+] \leq \mathbb{E}[V]
$$

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

We need to check that for some learning problems, either V or V^+ is manageable.

Lecture III: Roadmap

Concentration inequalities have proved helpful in statistical learning theory because they are key ingredients in the derivation of risk bounds for empirical risk minimizers in classification and bounded regression.
Concentration inequalities for self-bounded functionals allow to prove that many of the quantities that have been considered in order to quantify the complexity of a class of functions, like the empirical VC-dimension, the empirical VC-entropy and conditional Rademacher averages are sharply concentrated.
This paves the way to data-dependent estimation of the complexity of function classes, which is of great importance in model selection.
The most important consequence of concentration inequalities concerns suprema of empirical processes. The most refined versions of Talagrand's concentration inequality for suprema of empirical processes may be considered as process versions of Bernstein inequality. They provide new insights on Vapnik-Chervonenkis inequalities.

Fat-shattering dimension

Fat-shattering dimension fat$(X_1^n, \gamma, \mathcal{F})$.

Given $\gamma > 0$, and $\mathcal{F}$, and sample $X_1, \dots X_n$, fat$(X_1^n, \gamma, \mathcal{F})$ is the largest d such that there exists $\{i_1, \dots, i_d\} \subseteq \{1, \dots, n\}$ with

$$\forall (\epsilon_{i_j})_{j \leq d} \in \{-1, 1\}^d, \quad \exists f \in \mathcal{F}, \quad \epsilon_{i_j} f(X_{i_j}) \geq \gamma \, .$$

$$Z = \text{fat}(X_1^n, \gamma, \mathcal{F})$$

Fat-shattering dimension captures the complexity of a function class at certain scale on a certain sample.

If a set of linear classifiers $\mathcal{F}$ separates a sample $X_1, \dots, X_n$ with margin $\gamma > 0$:

$$\forall (\epsilon_i)_{i \leq n} \in \{-1, 1\}^n, \quad \exists f \in \mathcal{F}, \quad \epsilon_i f(X_i) \geq \gamma \quad \text{then fat}(X_1^n, \gamma, \mathcal{F}) = n.,,$$

Depending on the choice of kernels, on underlying distribution, fat$(X_1^n, \gamma, \mathcal{F})$ may scale very differently with n. Is the fat-shattering dimension relatively stable?

$$\frac{\text{fat}(X_1^n, \gamma, \mathcal{F})}{\mathbb{E}\Big[\text{fat}(X_1^n, \gamma, \mathcal{F})\Big]} \to 1 \ ?$$

Fat-shattering dimension

Fat-shattering dimension fat$(X_1^n, \gamma, \mathcal{F})$.

Given $\gamma > 0$, and $\mathcal{F}$, and sample $X_1, \dots X_n$, fat$(X_1^n, \gamma, \mathcal{F})$ is the largest d such that there exists $\{i_1, \dots, i_d\} \subseteq \{1, \dots, n\}$ with

$$\forall (\epsilon_{i_j})_{j \leq d} \in \{-1, 1\}^d, \quad \exists f \in \mathcal{F}, \quad \epsilon_{i_j} f(X_{i_j}) \geq \gamma \, .$$

$$Z = \text{fat}(X_1^n, \gamma, \mathcal{F})$$

Z is a configuration function: a γ-shattered sub-sample witnesses the value of Z

With $\quad Z_i = \text{fat}(X_1^{i-1}, X_{i+1}^n, \gamma, \mathcal{F}), \quad 0 \leq Z - Z_i \leq 1 \quad \& \quad \sum_i Z - Z_i \leq Z.$

Z is sub-Poissonian !!! $\quad \log \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] \leq \mathbb{E}[Z] \tau^*(\lambda) \, .$

$$\mathbb{P}\left[Z \geq \mathbb{E}Z + t\right] \leq \exp\left[-\frac{t^2}{2\mathbb{E}Z + 2t/3}\right]$$

For every $0 < t \leq \mathbb{E}Z, \quad \mathbb{P}\left[Z \leq \mathbb{E}Z - t\right] \leq \exp\left[-\frac{t^2}{2\mathbb{E}Z}\right] \, .$

VC-entropy

$\mathcal{F}$: a class of $\{-1,1\}$-valued functions.
X_1^n a sample.

Trace of $\mathcal{F}$ on X_1^n : $\left\{(b_i)_{i\leq n}; (b_i)\in\{-1,1\}^n,\ \exists f\in\mathcal{F}, b_i=f(X_i)\right\}$

The VC-entropy of $\mathcal{F}$ in X_1^n is defined as:

$$Z=\log_2\left|\left\{(b_i)_{i\leq n}; (b_i)\in\{-1,1\}^n,\ \exists f\in\mathcal{F}, b_i=f(X_i)\right\}\right|$$

Z_i is defined as

$$Z_i=\log_2\left|\left\{(b_j)_{j\leq n, j\neq i}; (b_j)\in\{-1,1\}^{n-1},\ \exists f\in\mathcal{F}, b_j=f(X_j)\ \forall j\right\}\right|$$

Obvious: $0\leq Z-Z_i\leq 1$,

VC-entropy

P a probability on a finite set.

Shannon entropy of P $\quad H(P) = \sum_i -P(i) \log_2 P(i)$

Shannon entropy is positive and maximal when P is uniform.

Notation	$H(X) = H(\text{dist}(X))$
Conditional entropy	$H(X \mid Y) = \mathbb{E}\Big[H(\text{dist}(X \mid Y))\Big]$
Chain rule.	$H(X,Y) = H(Y) + H(X \mid Y)$
Conditionning	$H(X \mid Y) \leq H(X)$ ...decreasses Shannon entropy.

VC-entropy

Han inequality.

$$\begin{aligned}\sum_i H(X_1^n) - H(X_1^{i-1}, X_{i+1}^n) &= \sum_{i=1}^n H(X_i \mid X_1^{i-1}, X_{i+1}^n) \\ &\leq \sum_{i=1}^n H(X_i \mid X_1^{i-1}) \\ &\leq H(X_1^n)\,.\end{aligned}$$

The uniform distribution on the trace of $\mathcal{F}$ on X_1^n defines a random element Y_1^n of $\{-1,1\}^n$

$$Z = H(Y_1^n)$$

Z_i is the Shannon-entropy of the uniform distribution on the trace of $\mathcal{F}$ on X_1^{i-1}, X_{i+1}^n.

$$Z_i \geq H(Y_1^{i-1}, Y_{i+1}^n)\,.$$

☞ Han inequality entails $\sum_i Z - Z_i \leq Z$.

VC-entropy

The VC-entropy is a self bounded functional. It enjoys a sub-Poissonian behavior.

$$V \leq Z \quad \Rightarrow \quad \mathrm{Var}[Z] \leq \mathbb{E}[Z]$$

☞ VC-entropy may be almost-surely constant (half-spaces when samples are in general position with probability one.)

☞ VC-entropy may be approximately Gaussian: samples of size n from the uniform distribution on a shattered set of size αn where α is fixed and n tends to infinity...

If we just know that the Z is a VC-entropy, the inequality is tight.

VC-entropy

$\mathcal{F}$ a VC-class of classifiers.

$$
\begin{aligned}
&\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\sum_i \mathbb{1}_{f(X_i)\neq Y_i} - n\mathbb{E}\big[\mathbb{1}_{f(X)\neq Y}\big]\Big|\Big] \\
&\quad= \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\sum_i \mathbb{1}_{f(X_i)\neq Y_i} - \sum_i \mathbb{E}'\big[\mathbb{1}_{f(X_i')\neq Y_i'}\big]\Big|\Big] \\
&\quad\leq \mathbb{E}\mathbb{E}'\Big[\sup_{f\in\mathcal{F}}\Big|\sum_i \mathbb{1}_{f(X_i)\neq Y_i} - \mathbb{1}_{f(X_i')\neq Y_i'}\Big|\Big] \\
&\quad= \mathbb{E}\mathbb{E}'\mathbb{E}_\epsilon\Big[\sup_{f\in\mathcal{F}}\Big|\sum_i \epsilon_i\big(\mathbb{1}_{f(X_i)\neq Y_i} - \mathbb{1}_{f(X_i')\neq Y_i'}\big)\Big|\Big] \\
&\quad\leq \mathbb{E}\mathbb{E}_\epsilon\Big[\sup_{f\in\mathcal{F}}\Big|\sum_i \epsilon_i \mathbb{1}_{f(X_i)\neq Y_i}\Big|\Big] + \mathbb{E}'\mathbb{E}_\epsilon\Big[\sup_{f\in\mathcal{F}}\Big|\sum_i \epsilon_i \mathbb{1}_{f(X_i')\neq Y_i'}\Big|\Big] \\
&\quad= 2\mathbb{E}\Big[\mathbb{E}_\epsilon\Big[\sup_{f\in\mathcal{F}}\Big|\sum_i \epsilon_i \mathbb{1}_{f(X_i)\neq Y_i}\Big| \mid X_1^n\Big]\Big]
\end{aligned}
$$

The conditional Rademacher average only depends on the trace of $\mathcal{F}$ on X_1^n ....

VC-entropy

$$\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left|\sum_i \epsilon_i \mathbb{1}_{f(X_i)\neq Y_i}\right| \mid X_1^n\right] \leq \sqrt{2\log|\text{trace}(\mathcal{F}, X_1^n)|}\times\sup_{f\in\mathcal{F}}\sqrt{\sum_i \mathbb{1}_{f(X_i)\neq Y_i}}$$

For a fixed $f\in\mathcal{F}$ $\mathbb{E}_\epsilon\left[e^{\lambda\left|\sum_i \epsilon_i \mathbb{1}_{f(X_i)\neq Y_i}\right|}\right] \leq 2\exp\left(\frac{\lambda^2}{2}\sum_i \mathbb{1}_{f(X_i)\neq Y_i}\right)$ Hoeffding!

Revisiting the union bound: $\exp\left(\lambda\mathbb{E}\left[\max_{f\in\text{trace}(\mathcal{F},X_1^n)}\left|\sum_i \epsilon_i \mathbb{1}_{f(X_i)\neq Y_i}\right|\right]\right)$

$$\begin{aligned}
&\leq \mathbb{E}\left[\exp\left(\lambda\max_{f\in\text{trace}(\mathcal{F},X_1^n)}\left|\sum_i \epsilon_i \mathbb{1}_{f(X_i)\neq Y_i}\right|\right)\right]\\
&\leq \mathbb{E}\left[\sum_{f\in\text{trace}(\mathcal{F},X_1^n)}\exp\left(\lambda\left|\sum_i \epsilon_i \mathbb{1}_{f(X_i)\neq Y_i}\right|\right)\right]\\
&\leq \sum_{f\in\text{trace}(\mathcal{F},X_1^n)}\mathbb{E}\left[\exp\left(\lambda\left|\sum_i \epsilon_i \mathbb{1}_{f(X_i)\neq Y_i}\right|\right)\right]\\
&\leq 2|\text{trace}(\mathcal{F}, X_1^n)|\max_{f\in\text{trace}(\mathcal{F},X_1^n)}\exp\left(\frac{\lambda^2}{2}\sum_i \mathbb{1}_{f(X_i)\neq Y_i}\right)
\end{aligned}$$

The bound follows by optimizing with respect to λ.

VC-entropy

$$
\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\sum_i \mathbb{1}_{f(X_i)\neq Y_i} - n\mathbb{E}\big[\mathbb{1}_{f(X)\neq Y}\big]\Big|\Big] \\
&\leq \mathbb{E}\Big[\sqrt{2\log|\mathsf{trace}(\mathcal{F},X_1^n)|}\times\sup_{f\in\mathcal{F}}\sqrt{\sum_i \mathbb{1}_{f(X_i)\neq Y_i}}\Big] \\
&\leq \sqrt{\mathbb{E}\Big[2\log|\mathsf{trace}(\mathcal{F},X_1^n)|\Big]}\times\sqrt{\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_i \mathbb{1}_{f(X_i)\neq Y_i}\Big]} \\
&\leq \sqrt{\mathbb{E}\Big[2\log|\mathsf{trace}(\mathcal{F},X_1^n)|\Big]}\times\sqrt{\sup_{f\in\mathcal{F}} n\mathbb{E}\big[\mathbb{1}_{f(X)\neq Y}\big]+\mathbb{E}[Z]}
\end{aligned}
$$

$$
\mathbb{E}[Z] \leq \mathbb{E}\Big[2\log|\mathsf{trace}(\mathcal{F},X_1^n)|\Big] + \sup_{f\in\mathcal{F}} n\mathbb{E}\big[\mathbb{1}_{f(X)\neq Y}\big]
$$

May be highly relevant if $\mathcal{F}$ is constituted by classifiers with law error rate (order $1/n$).

Line of reasoning is at the origin of (some) localization procedures.

Rademacher complexity

The empirical fat-shattering dimension, and the empirical VC-entropy, are two capacity concepts in statistical learning theory. Their relevance stems from their relationship to the average value of $\sup_{f\in\mathcal{F}}\left|\sum_i \mathbb{1}_{f(X_i)\neq Y_i} - n\mathbb{E}\left[\mathbb{1}_{f(X)\neq Y}\right]\right|$.

This relationship is proved using conditional Rademacher averages (symmmetrization). Why shouldn't we prove that conditional Rademacher averages are concentrated around their mean value.

$\mathcal{F}$: countable class of measurable centered real-valued functions of $[-1,1]$-valued functions..

$$Z = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_i \epsilon_i f(X_i)\right| \mid X_1^n\right]$$

ϵ_i: independent centered $\{-1,1\}$-valued Random variables.

$$Z_i = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{j\neq i} \epsilon_j f(X_j)\right| \mid X_1^{i-1}, X_{i+1}^n\right]$$

Rademacher complexity

$$
\begin{aligned}
Z_i &= \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \Big| \sum_{j\neq i} \epsilon_j f(X_j) + \mathbb{E}_{\epsilon_i} \epsilon_i f(X_j) \Big| \mid X_1^n \Big] \\
&\leq \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \Big| \sum_{j\neq i} \epsilon_j f(X_j) + \epsilon_i f(X_j) \Big| \mid X_1^n \Big] \quad \text{Jensen} \\
&= Z
\end{aligned}
$$

$$
Z \leq \mathbb{E}_\epsilon \Big[\sup_{f\in\mathcal{F}} \big| \sum_{j\neq i} \epsilon_j f(X_j) \big| + 1 \mid X_1^n \Big] = Z_i + 1 \,.
$$

$$
\sum_i (Z - Z_i) \quad \leq \quad Z \quad \text{triangle inequality.}
$$

Rademacher complexity

Rademacher complexities are self-bounded functionals. They also enjoy sub-Poissonian behavior.

$$V \leq Z \quad \rightarrow \quad \mathrm{Var}[Z] \leq \mathbb{E}[Z]$$

For VC-classes, $\mathbb{E}[Z] \approx C\sqrt{\mathrm{vc}\, n}$.
E-S inequality implies that the typical fluctuations of Rademacher complexities. are of order at most $(dn)^{1/4}$.

Again this may be too conservative.

Suprema of positive bounded empirical processes

$$Z = \sup_{f \in \mathcal{F}} \sum_i f(X_i)$$

where $\mathcal{F}$ is a set of positive functions, ...
Assumption $\sup_f \sup_x f(x) \leq 1$

$$Z_i = \sup_{f \in \mathcal{F}} \sum_{j \neq i} f(X_j)$$

$$0 \leq Z - Z_i \leq 1 \quad \text{and} \quad \sum_i Z - Z_i \leq Z\,.$$

Z is sub-Poissonian !!! $\quad \log \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] \leq \mathbb{E}[Z]\tau^*(\lambda)\,.$

$$\mathbb{P}\left[Z \geq \mathbb{E}Z + t\right] \leq \exp\left[-\frac{t^2}{2\mathbb{E}Z + 2t/3}\right]$$

For every $0 < t \leq \mathbb{E}Z$, $\quad \mathbb{P}\left[Z \leq \mathbb{E}Z - t\right] \leq \exp\left[-\frac{t^2}{2\mathbb{E}Z}\right]\,.$

Suprema of positive bounded empirical processes

Gram matrix (Kernel-machines) $G_{i,j} = \langle X_i, X_j \rangle$

Let $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \geq 0$ denote the ordered sequence of eigenvalues of G.

What can we say about $\sum_{i=1}^{r} \lambda_i$?

It may be representeed as a supremum of a positive bounded empirical processes
Π_V: projection on subspace V.

$$\sum_{i=1}^{r} \lambda_i = \sup_{V:\dim(V)=r} \|\Pi_V(X_i)\|^2$$

If the distribution of Xi has bounded support, this is the supremum of a bounded positive empirical process...

Suprema of bounded empirical processes

Suprema of empirical processes

$$\sup_{f \in \mathcal{F}} |\sum_i \Big(f(X_i) - \mathbb{E}[f(X_i)] \Big)|$$

where $\mathcal{F}$ may be a set of classifiers, regression functions, ...
Assumption $\sup_f \sup_x |f(x)| \leq 1$

The analysis of empirical processes are at the root of the Vapnik-Chervonenkis theory of learning. More generally, they prove to be central in the annalysis of M-estimators.

The supremum is one among many quantities associated with an empirical process. Another is the modulus of continuity.

As the supremum of an empirical process is the sup-norm of a sum of random vectors, a natural question is: are there extensions of Bernstein inequalities for suprema of empirical processes?

Using specifically modified logarithmic Sobolev inequalities, Rio and Bousquet have proved that Bernstein inequality scales up to the vector-valued setting.

Suprema of bounded empirical processes

Suprema of bounded empirical processes provide an exemple of a two-steps approach of concentration inequalities:

- Compute a tractable upper-bound for Efron-Stein estimates.
- Analyse the concentration properties of this Efron-Stein estimate.

$V^+ \leq \sup_{f\in\mathcal{F}} \sum_i \mathbb{E}'\left[(f(X_i) - f(X_i'))^2\right]$ If $\mathbb{E}[f] = 0$ for all $f \in \mathcal{F}$

$$V^+ \leq \sup_{f\in\mathcal{F}} \sum_i f^2(X_i) + \sup_{f\in\mathcal{F}} \mathbb{E}\left[f^2(X_i)\right]$$

The stochastic part of the upper-bound on V^+ is a supremum of a bounded positive process !

$$\log \mathbb{E}\left[e^{\lambda V^+}\right] \leq \lambda\left(\mathbb{E}[\sup_{f\in\mathcal{F}} \sum_i f^2(X_i)] + \sup_{f\in\mathcal{F}} \mathbb{E}\left[f^2(X_i)\right]\right) + \mathbb{E}[\sup_{f\in\mathcal{F}} \sum_i f^2(X_i)]\tau^*(\lambda).$$

Suprema of bounded empirical processes

Suprema of bounded empirical processes

$$
\begin{aligned}
\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_i f^2(X_i)\Big] &\leq \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_i f^2(X_i)-\mathbb{E}[f^2(X_i)]\Big]+\sup_{f\in\mathcal{F}}\sum_i \mathbb{E}\Big[f^2(X_i)\Big]\\
&\leq 2\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_i \epsilon_i f^2(X_i)\Big]+\sup_{f\in\mathcal{F}}\sum_i \mathbb{E}\Big[f^2(X_i)\Big] \quad \text{symmetrization}\\
&\leq 4\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_i \epsilon_i f(X_i)\Big]+\sup_{f\in\mathcal{F}} \mathbb{E}\Big[f^2(X_i)\Big] \quad \text{contraction}\\
&\leq 4\mathbb{E}\Big[\sup_{f\in\mathcal{F}}|\sum_i f(X_i)-\mathbb{E}[f(X_i)]|\Big]+\sup_{f\in\mathcal{F}}\sum_i \mathbb{E}\Big[f^2(X_i)\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\log\mathbb{E}\Big[e^{\lambda V^+}\Big]\\
&\leq (2\lambda+\tau^*(\lambda))\sum_i\Big(\sup_{f\in\mathcal{F}}\mathbb{E}\Big[f^2(X_i)\Big]\Big)+4(\tau^*(\lambda)+\lambda)\mathbb{E}[\sup_{f\in\mathcal{F}}\sum_i|f(X_i)-\mathbb{E}[f(X_i)]|]
\end{aligned}
$$

Suprema of bounded empirical processes

Plugging

$$\log \mathbb{E}\left[e^{\lambda V^+}\right]$$
$$\leq (2\lambda + \tau^*(\lambda))\Big(\sup_{f\in\mathcal{F}} \sum_i \mathbb{E}\Big[f^2(X_i)\Big]\Big) + 4(\tau^*(\lambda)+\lambda)\mathbb{E}[\sup_{f\in\mathcal{F}} \sum_i |f(X_i) - \mathbb{E}[f(X_i)]|]$$

in exponential Efron-Stein inequality.

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])_+}\right]$$
$$\frac{\theta\lambda}{1-\theta\lambda}\left(\Big(\frac{2\lambda}{\theta} + \tau^*\big(\frac{\lambda}{\theta}\big)\Big)\sup_{f\in\mathcal{F}} \sum_i \mathbb{E}\Big[f^2(X_i)\Big] + 4\big(\frac{\lambda}{\theta} + \tau^*\big(\frac{\lambda}{\theta}\big)\big)\mathbb{E}[Z]\right)$$

Suprema of bounded empirical processes

The official concentration inequality for suprema of bounded empirical processes. Talagrand (94,96), Leddoux (97), Massart(2000), Rio(2001), Bousquet (2002)

$$v = \sup_{f \in \mathcal{F}} \sum_i \mathbb{E}\left[f^2(X_i)\right] + 2\mathbb{E}[Z]$$

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

This holds for very small classes as well as for large classes ...

☞ What distinguishes large and small classes is $\mathbb{E}[Z]$ not the concentration phenomenon.

☞ $\mathbb{E}[Z]$ may be analyzed using different techniques (chaining).

Moment inequalities

Organization

- From exponential inequalities to the concentration of measure phenomenon
- Concentration inequalities using the entropy method
- Learning-theoretical applications
- Moment inequalities using the generalized entropy method

Plan IV : when increments are not bounded

- From exponential inequalities to the concentration of measure phenomenon
- Concentration inequalities using the entropy method
- Learning-theoretical applications
- Moment inequalities
  - Beyond exponential inequalities
  - Main moment inequalities
  - ϕ-Sobolev inequalities
  - From ϕ-Sobolev inequalities to moment inequalities
  - Rosenthal inequalities from ϕ-Sobolev inequalities
  - Conditional Rademacher averages for general processes

Where are we?

The entropy method (lecture II) has allowed us to scale up Efron-Stein inequalities (lecture I) to the level of exponential moments. In lecture III, we have illustrated the power of exponential Efron-Stein inequalities by showing that some quantities that play a central role in Vapnik-Chervonenkis theory of learning are indeed concentrated around their mean value.

Such results proved vvery helpful when understanding classification problems. However, the concentration inequalities described in lecture II, took full advantage of some boundedness assumptions.

In some settings (regression, SVM-soft-margin classification), assuming that everything is bounded seems very restrictive. Hence there is a need to interpolate between Efron-Stein and exponential Efron-Stein inequalities.

Lecture IV: Roadmap

Exponential inequalities prove powerless when dealing with sums of poorly integrable independent random variables. In order to get tail inequalities for such sums, we traditionally resort to moment inequalities known as Rosenthal-Pinelis inequalities. The aim of this lecture is to describe the generalized entropy method. This generalized entropy method provides upper-bound on the qth norm of functions of many independent random variables.

The approach is to relate the qth norm of some functional of independent random variables, with the qth norm of the square root of V^+ (the quantity which shows up in the Efron-Stein upper-bound), and then to upper-bound $\|\sqrt{V^+}\|_q$.

The relationship between $\|(Z - \mathbb{E}[Z])_+\|_q$ and $\|\sqrt{V^+}\|_q$, interpolates between Efron-Stein inequality and modified logarithmic Sobolev inequalities. It is proved by resorting to functionals called ϕ-entropies that generalize both entropy and variance. Those functionals called ϕ-entropies enjoy a duality property that warrants that they also enjoy the tensorization property. Using tensorization and optimization, it is thus possible to show that product probability distributions enjoy ϕ-Sobolev inequalities that may be used to derive moment inequalities.

Inspirations

- The Martingale approach is well-adapted to derive moment inequalities.
  Recall: $(\mathcal{F}_i) = \sigma(X_1^i)$-algebra.
  $\Big(M_i = \mathbb{E}[Z|\mathcal{F}_i]\Big)_i$ is an $\mathcal{F}_i$-adapted martingale.

$$\langle Z \rangle = \sum_{i=1}^{n} (M_i - M_{i-1})^2 \, .$$

Burkholder inequalities

$$\|Z - \mathbb{E}[Z]\|_q \leq (q-1)\sqrt{\|\langle Z \rangle\|_{q/2}} = (q-1)\left\|\sqrt{\langle Z \rangle}\right\|_q \, .$$

Inspirations

- Burkholder inequalities

$$\|Z - \mathbb{E}[Z]\|_q \leq (q-1)\sqrt{\|\langle Z \rangle\|_{q/2}} = (q-1)\left\|\sqrt{\langle Z \rangle}\right\|_q \quad .$$

- Efron-Stein inequalities

$$V_+ \triangleq \sum_i \mathbb{E}\left[\left(Z - Z^{(i)}\right)^2 \mathbb{1}_{Z > Z^{(i)}} \mid X_1^n\right] \quad .$$

$$\mathrm{Var}(Z) \leq \mathbb{E}\left[V^+\right] .$$

What about other moments?

Inspirations

- Burkholder inequalities

$$\|Z - \mathbb{E}[Z]\|_q \leq (q-1)\sqrt{\|\langle Z\rangle\|_{q/2}} = (q-1)\left\|\sqrt{\langle Z\rangle}\right\|_q \quad .$$

- Efron-Stein inequalities

$$\mathrm{Var}(Z) \leq \mathbb{E}\left[V^+\right].$$

What about other moments?

- Exponential Efron-Stein inequality.

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq \frac{\lambda\theta}{1-\lambda\theta} \log \mathbb{E}\left[e^{\frac{\lambda V^+}{\theta}}\right] \quad \text{for} \quad \lambda \in (0, 1/\theta).$$

Requires exponential integrability of V^+.

Inspirations

- Burkholder inequalities

$$\|Z - \mathbb{E}[Z]\|_q \leq (q-1)\sqrt{\|\langle Z\rangle\|_{q/2}} = (q-1)\left\|\sqrt{\langle Z\rangle}\right\|_q \quad .$$

- Efron-Stein inequalities

$$\mathrm{Var}(Z) \leq \mathbb{E}\left[V^+\right].$$

What about other moments?

- Exponential Efron-Stein inequality.

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq \frac{\lambda\theta}{1-\lambda\theta}\log \mathbb{E}\left[e^{\frac{\lambda V^+}{\theta}}\right] \quad \text{for} \quad \lambda \in (0, 1/\theta).$$

Requires exponential integrability of V^+.

- Goal: Burkholder-type inequalities relating the qth norm of Z with the $q/2$th norm V or V^+

The case for moment inequalities

- Goal: relate the integrability of Z with the integrability of V^+.
  Viewing Burkholder inequalities, it is reasonable to assume that Z is as integrable as $\sqrt{V^+}$.
  It seems highly desirable to get a sharp dependence on q.

The case for moment inequalities

- Goal: relate the integrability of Z with the integrability of V^+.
  Viewing Burkholder inequalities, it is reasonable to assume that Z is as integrable as $\sqrt{V^+}$.
  It seems highly desirable to get a sharp dependence on q.
- Moment inequalities might be interesting per se:

$$\mathbb{P}\{Z \geq t\} \leq \min_q \left(\frac{\|Z\|_q}{t} \right)^q \leq \inf_\lambda \frac{\mathbb{E}\left[e^{\lambda Z}\right]}{e^{\lambda t}}$$

The case for moment inequalities

- Goal: relate the integrability of Z with the integrability of V^+.
  Viewing Burkholder inequalities, it is reasonable to assume that Z is as integrable as $\sqrt{V^+}$.
  It seems highly desirable to get a sharp dependence on q.
- Moment inequalities might be interesting per se:

$$\mathbb{P}\{Z \geq t\} \leq \min_q \left(\frac{\|Z\|_q}{t} \right)^q \leq \inf_\lambda \frac{\mathbb{E}\left[e^{\lambda Z}\right]}{e^{\lambda t}}$$

- The relationship between $\|Z\|_q$ and q reflects the tail behavior of Z.
  "Equivalence of moments" principle asserts that $\|Z\|_q$ grows slower than q^d iff $Z^{1/d}$ is exponentially integrable.

The case for moment inequalities

- Goal: relate the integrability of Z with the integrability of V^+.
  Viewing Burkholder inequalities, it is reasonable to assume that Z is as integrable as $\sqrt{V^+}$.
  It seems highly desirable to get a sharp dependence on q.
- Moment inequalities might be interesting per se:

$$\mathbb{P}\{Z \geq t\} \leq \min_q \left(\frac{\|Z\|_q}{t}\right)^q \leq \inf_\lambda \frac{\mathbb{E}\left[e^{\lambda Z}\right]}{e^{\lambda t}}$$

- The relationship between $\|Z\|_q$ and q reflects the tail behavior of Z.
  "Equivalence of moments" principle asserts that $\|Z\|_q$ grows slower than q^d iff $Z^{1/d}$ is exponentially integrable.
- Moment inequalities with a **tight** dependence on q provide Bernstein-like inequalities.

if $\|(Z - \mathbb{E}[Z])_+\|_q \leq \sum_{i=1}^{d} A_i q^{i/2}$ $\quad \mathbb{P}\{Z > \mathbb{E}[Z] + t\} \leq \exp\left(-\log 2 \min_{i \leq d} \left(\frac{dt}{2A_i}\right)^{2/i}\right)$.

Main inequalities

$q \in \mathbb{N}, q \geq 2.$

- Gaussian type behavior, bounded-difference inequality.

$$\text{if } V^+ \leq c, \text{ then } \quad q \geq 2, \quad \left\| (Z - \mathbb{E}[Z])_+ \right\|_q \leq \sqrt{qc} \, .$$

Main inequalities

$q \in \mathbb{N}, q \geq 2.$

- Gaussian type behavior, bounded-difference inequality.

$$\text{if } V^+ \leq c, \text{ then } \quad q \geq 2, \quad \left\| (Z - \mathbb{E}[Z])_+ \right\|_q \leq \sqrt{qc}\,.$$

- Burkholder-like inequality

$$\left\| (Z - \mathbb{E}[Z])_+ \right\|_q \leq \sqrt{3q} \left\| \sqrt{V^+} \right\|_q ,$$

Main inequalities

$q \in \mathbb{N}, q \geq 2.$

- Gaussian type behavior, bounded-difference inequality.

$$\text{if } V^+ \leq c, \text{ then } \quad q \geq 2, \quad \left\| (Z - \mathbb{E}[Z])_+ \right\|_q \leq \sqrt{qc}\,.$$

- Burkholder-like inequality

$$\left\| (Z - \mathbb{E}[Z])_+ \right\|_q \leq \sqrt{3q} \left\| \sqrt{V^+} \right\|_q,$$

- Burkholder-like inequality (II)
  If $Z_i \leq Z$ for all i,

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

Scheme of Proof

Modified ϕ-Sobolev inequalities.

The milestone in the proof of the Burkholder-like inequalities is a relationship with the following flavor:

$q \geq 2$ and α satisfies $q/2 \leq \alpha \leq q-1$. Then

$$\mathbb{E}\left[(Z-\mathbb{E}[Z])_+^q\right] - \mathbb{E}\left[(Z-\mathbb{E}[Z])_+^\alpha\right]^{q/\alpha} \leq \frac{q(q-\alpha)}{2}\mathbb{E}\left[V(Z-\mathbb{E}[Z])_+^{q-2}\right],$$

Letting $\phi(x) = x^{q/\alpha}$, this translates into

$$\mathbb{E}\left[\phi\left((Z-\mathbb{E}[Z])_+^\alpha\right)\right] - \phi\left(\mathbb{E}\left[(Z-\mathbb{E}[Z])_+^\alpha\right]\right) \leq \frac{q(q-\alpha)}{2}\mathbb{E}\left[V(Z-\mathbb{E}[Z])_+^{q-2}\right],$$

For $q=2$ and $\alpha=1$, this is (almost) Efron-Stein !!!

Scheme of Proof

In order to establish Burkholder-like inequality, we proceed by induction on q. At each step we take $\alpha = q - 1$, that is $\phi(x) = x^{q/(q-1)}$.

☞ when dealing with moments of high order, we consider functions ϕ that get closer and closer to $x \log x$.

$$\mathbb{E}\left[(Z - \mathbb{E}[Z])_+^q\right] \quad \leq \quad \mathbb{E}\left[(Z - \mathbb{E}[Z])_+^{q-1}\right]^{q/(q-1)} \quad + \quad \frac{q}{2}\mathbb{E}\left[V\,(Z - \mathbb{E}[Z])_+^{q-2}\right],$$

Scheme of Proof

In order to establish Burkholder-like inequality, we proceed by induction on q. At each step we take $\alpha = q - 1$, that is $\phi(x) = x^{q/(q-1)}$.

☞ when dealing with moments of high order, we consider functions ϕ that get closer and closer to $x \log x$.

$$\mathbb{E}\left[(Z - \mathbb{E}[Z])_+^q\right] \quad \leq \quad \mathbb{E}\left[(Z - \mathbb{E}[Z])_+^{q-1}\right]^{q/(q-1)} \quad + \quad \frac{q}{2}\mathbb{E}\left[V\,(Z - \mathbb{E}[Z])_+^{q-2}\right],$$

Hölder

$$\mathbb{E}\left[(Z - \mathbb{E}[Z])_+^q\right] \leq \mathbb{E}\left[(Z - \mathbb{E}[Z])_+^{q-1}\right]^{q/(q-1)} + \frac{q}{2}\mathbb{E}\left[V^{q/2}\right]^{2/q}\mathbb{E}\left[(Z - \mathbb{E}[Z])_+^q\right]^{(q-2)/q},$$

Scheme of Proof

Hölder

$$\mathbb{E}\left[(Z-\mathbb{E}[Z])_+^q\right] \leq \mathbb{E}\left[(Z-\mathbb{E}[Z])_+^{q-1}\right]^{q/(q-1)} + \frac{q}{2}\mathbb{E}\left[V^{q/2}\right]^{2/q}\mathbb{E}\left[(Z-\mathbb{E}[Z])_+^q\right]^{(q-2)/q},$$

$m_q \triangleq \|(Z-\mathbb{E}[Z])_+\|_q$ and $v_q \triangleq \left\|\sqrt{V}\right\|_q$.

- Base case : Efron-Stein: $m_2 \leq v_2$.
- Assume $m_{q-1} \leq \sqrt{3(q-1)}v_{q-1}$,

$$\begin{aligned} m_q^q &\leq m_{q-1}^q + \frac{q}{2}v_q^2 m_q^{q-2} \\ &\leq \sqrt{3(q-1)}^q v_{q-1}^q + \frac{q}{2}v_q^2 m_q^{q-2} \\ &\cdots \\ &\leq \sqrt{3q}v_q\,. \end{aligned}$$

Modified ϕ-Sobolev inequalities

Φ : functions ϕ on $\mathbb{R}^+$ (Latala and Oleskiewicz)

- convex and continuous,
- twice differentiable on $(0, +\infty)$,
- ϕ'' is positive ,
- $1/\phi''$ is concave.

ϕ-entropy. $H_\phi(Z) = \mathbb{E}\big[\phi(Z)\big] - \phi\big(\mathbb{E}[Z]\big)\,.$

Examples :

$$\begin{aligned} H_{x\log x}(Z) &= \mathbb{E}[Z\log Z] - \mathbb{E}[Z]\log\mathbb{E}[Z] \\ H_{x^p}(Z) &= \mathbb{E}[Z^p] - (\mathbb{E}[Z])^p \quad p\in(1,2]\,. \end{aligned}$$

Modified ϕ-Sobolev inequalities

Tensorization of ϕ-entropy.

$$H_\phi(Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\phi(Z) \mid X_1^{i-1}, X_{i+1}^n\right] - \phi\left(\mathbb{E}\left[Z \mid X_1^{i-1}, X_{i+1}^n\right]\right)\right] .$$

$$H_\phi(Z) \leq \sum_{i=1}^{n} H_\phi(Z \mid X_1^{i-1}, X_{i+1}^n) .$$

Modified ϕ-Sobolev inequalities

Tensorization of ϕ-entropy.

$$H_\phi(Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\phi(Z) \mid X_1^{i-1}, X_{i+1}^n\right] - \phi\left(\mathbb{E}\left[Z \mid X_1^{i-1}, X_{i+1}^n\right]\right)\right] .$$

Duality and ϕ-entropy.

H_ϕ is convex and continuous in $\mathbb{L}_1^+$.
Obvious when the base space is discrete.

$$H_\phi(Z) = \sup_{T \in \mathbb{L}_1^+, T \neq 0} \left\{\mathbb{E}\left[\left(\phi'(T) - \phi'(\mathbb{E}[T])\right)(Z - T) + \phi(T)\right] - \phi(\mathbb{E}[T])\right\} .$$

The duality representation implies that H_ϕ is convex and lower-semi-continuous.
This is the basis of a Jensen-like property:

$$H(\mathbb{E}[Z \mid X_1]) \leq H(Z \mid X_2)$$

Modified ϕ-Sobolev inequalities

Duality and ϕ-entropy.

H_ϕ is convex and continuous in $\mathbb{L}_1^+$.
Obvious when the base space is discrete.

$$H_\phi(Z) = \sup_{T\in\mathbb{L}_1^+, T\neq 0} \left\{ \mathbb{E}\left[\left(\phi'(T) - \phi'(\mathbb{E}[T])\right)(Z-T) + \phi(T) \right] - \phi(\mathbb{E}[T]) \right\} .$$

The duality representation implies that H_ϕ is convex and lower-semi-continuous.
This is the basis of a Jensen-like property:

$$H(\mathbb{E}[Z \mid X_1]) \leq H(Z \mid X_2)$$

Examples.

$$H_\phi(Z) = \sup_T \{\mathbb{E}[(\log(T) - \log(\mathbb{E}[T]))Z]\} \quad \text{for } \phi(x) = x\log x$$

$$H_\phi(Z) = \sup_T \left\{ p\mathbb{E}\left[Z\left(T^{p-1} - (\mathbb{E}[T])^{p-1} \right) \right] - (p-1)H_\phi(T) \right\}, \quad \text{for } \phi(x) = x^p$$

Modified ϕ-Sobolev inequalities

Optimization.

If $\phi \in \Phi$, then both ϕ' and $x \to (\phi(x) - \phi(0))/x$ are concave functions on $\mathbb{R}_+^*$.

Modified ϕ-Sobolev inequalities

Optimization.

If $\phi \in \Phi$, then both ϕ' and $x \to (\phi(x) - \phi(0))/x$ are concave functions on $\mathbb{R}_+^*$.

Modified ϕ-Sobolev inequalities

Optimization.

If $\phi \in \Phi$, then both ϕ' and $x \to (\phi(x) - \phi(0))/x$ are concave functions on $\mathbb{R}_+^*$.
Let ψ denote the function $x \to (\phi(x) - \phi(0))/x$.

$$H_\phi(f(Z)) \leq \mathbb{E}\left[V^+ f'^2(Z)\, \psi'(f(Z))\right] \quad \text{if} \psi \circ f \text{ is convex.}$$

Modified ϕ-Sobolev inequalities

Optimization.

If $\phi \in \Phi$, then both ϕ' and $x \to (\phi(x) - \phi(0))/x$ are concave functions on $\mathbb{R}_+^*$.
Application : $f(z) = \exp(\lambda z)$ and $\phi(x) = x \log(x)$:

$$H_{x \log x}\left(e^{\lambda Z}\right) \leq \lambda^2 \mathbb{E}\left[V^+ \exp(\lambda Z)\right],$$

Let $q \geq 2$ and let α satisfy $q/2 \leq \alpha \leq q-1$. Then

$$\mathbb{E}\left[(Z - \mathbb{E}[Z])_+^q\right] \leq$$
$$\mathbb{E}\left[(Z - \mathbb{E}[Z])_+^\alpha\right]^{q/\alpha} + \frac{q(q-\alpha)}{2} \mathbb{E}\left[V (Z - \mathbb{E}[Z])_+^{q-2}\right],$$

Khinchine inequalities (warm-up)

$a_1, \ldots, a_n$: non-negative constants,
$X_1, \ldots, X_n$ independent Rademacher variables.
If $Z = \sum_{i=1}^n a_i X_i$
then for any integer $q \geq 2$,

$$\|(Z)_+\|_q = \|(Z)_-\|_q \leq \sqrt{2q}\sqrt{\sum_{i=1}^n a_i^2}$$

☞ Optimal constants can be obtained using hypercontractivity arguments.

$$\text{If } V^+ \leq c, \quad \|(Z - \mathbb{E}[Z])_+\|_q \leq \sqrt{qc}\,.$$

$$V^+ = \sum_{i=1}^n \mathbb{E}\left[(a_i(X_i - X_i'))_+^2 \mid X_i\right] = 2\sum_{i=1}^n a_i^2 \mathbb{1}_{a_i X_i > 0} \leq 2\sum_{i=1}^n a_i^2\,,$$

Rosenthal-Pinelis inequalities

Rosenthal-Pinelis inequalities deal with suprema of empirical processes in the absence of uniform boundedness assumption.
Let $\mathcal{F}$: a countable class of measurable functions from $\mathcal{X} \to \mathbb{R}$.
$X_1, \ldots, X_n$: independent $\mathcal{X}$-valued random variables such that for all $f \in \mathcal{F}$, $\mathbb{E}f(X_i) = 0$.

$$\begin{aligned} \Sigma^2 &= \mathbb{E}\Big[\sup_f \sum_i f^2(X_i)\Big] \\ \sigma^2 &= \sup_f \mathbb{E}\Big[\sum_i f^2(X_i)\Big] \\ M &= \sup_{f,i} |f(X_i)| \,. \end{aligned}$$

In Talagrand's inequality, M is assumed to be bounded.
Here, integrability of M is related wit integrability of Z.

$$Z = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n f(X_i) \right| .$$

Rosenthal-Pinelis inequalities

$$Z = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} f(X_i) \right| .$$

The application of Burkholder like inequality leads to:

$$\left\| (Z - \mathbb{E}[Z])_+ \right\|_q \leq \sqrt{6q} \Big(\Sigma + \sigma \Big) + 6q \Big(\|M\|_q + \sup_{i, f \in \mathcal{F}} \|f(X_i)\|_2 \Big) ,$$

$$\|Z\|_q \quad \leq \quad 2\mathbb{E}Z \quad + \quad 2\sigma\sqrt{6q} \quad + \quad 60q \|M\|_q \quad + \quad 4\sqrt{3q} \|M\|_2 \quad .$$

Rosenthal-Pinelis inequalities

$$Z = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} f(X_i) \right| .$$

$$\left\| (Z - \mathbb{E}[Z])_+ \right\|_q \leq \sqrt{6q} \Big(\Sigma + \sigma \Big) + 6q \Big(\|M\|_q + \sup_{i, f \in \mathcal{F}} \|f(X_i)\|_2 \Big) ,$$

As $\mathbb{E}[f] = 0$ for all $f \in \mathcal{F}$:

$$V^+ \leq \sup_{f \in \mathcal{F}} \sum_i f^2(X_i) + \sup_{f \in \mathcal{F}} \mathbb{E}\Big[f^2(X_i) \Big] = \sup_{f \in \mathcal{F}} \sum_i f^2(X_i) + \sigma^2$$

$$\sqrt{V^+} \leq \sup_{f \in \mathcal{F}} \sup_{\alpha : \sum_i \alpha_i^2 \leq 1} \sum_i \alpha_i f(X_i) + \sigma \, .$$

Applying the second Burkholder-like inequality again:

$$\left\| \sup_{f \in \mathcal{F}} \sup_{\alpha : \sum_i \alpha_i^2 \leq 1} \sum_i \alpha_i f(X_i) \right\|_q \leq \Sigma + \sqrt{3q} \, \|M\|_q$$

Rosenthal-Pinelis inequalities

$$Z = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} f(X_i) \right| .$$

$$\|Z\|_q \quad \leq \quad 2\mathbb{E}Z \quad + \quad 2\sigma\sqrt{6q} \quad + \quad 60q\,\|M\|_q \quad + \quad 4\sqrt{3q}\,\|M\|_2 \quad .$$

To get this inequality, need to relate Σ, $\mathbb{E}[Z]$ and σ
In the uniformly bounded case, symmetrization and the contraction principle were enough.
We need a new ingredient.

Hoffman-Jorgensen inequality.
Let $\epsilon_1, \ldots, \epsilon_n$ denote independent Rademacher variables.
Let $\lambda > 4$ and define $t_0 = \sqrt{\lambda \mathbb{E}\left[M^2\right]}$.

$$\mathbb{E}\left[\sup_f \left| \sum_i \epsilon_i f^2(X_i) \mathbb{1}_{\sup_f |f(X_i)| > t_0} \right| \right] \leq \frac{1}{(1 - 2/\sqrt{\lambda})^2} \mathbb{E}\left[M^2\right] .$$

Rosenthal-Pinelis inequalities

$$Z = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} f(X_i) \right| .$$

$$\begin{aligned}
\Sigma^2 &\leq \mathbb{E}\left[\sup_f \left|\sum_i f^2(X_i) - \mathbb{E}[f^2(X_i)]\right|\right] + \sup_f \mathbb{E}\left[\sum_i f^2(X_i)\right] \\
&\leq \sigma^2 + 2\mathbb{E}\left[\sup_f \left|\sum_i \epsilon_i f^2(X_i)\right|\right] \\
&\leq \sigma^2 + 2\mathbb{E}\left[\sup_f \left|\sum_i \epsilon_i f^2(X_i) \mathbb{1}_{\sup_f |f(X_i)| \leq t_0}\right|\right] \\
&\quad + 2\mathbb{E}\left[\sup_f \left|\sum_i \epsilon_i f^2(X_i) \mathbb{1}_{\sup_f |f(X_i)| > t_0}\right|\right] \\
&\leq \sigma^2 + 4t_0 \mathbb{E}\left[\sup_f \left|\sum_i \epsilon_i f(X_i)\right|\right] + 2\mathbb{E}\left[\sup_f \left|\sum_i \epsilon_i f^2(X_i) \mathbb{1}_{\sup_f |f(X_i)| > t_0}\right|\right] \\
&\leq \sigma^2 + 4t_0 \mathbb{E}\left[\sup_f \left|\sum_i \epsilon_i f(X_i)\right|\right] + \frac{2}{(1-\lambda/2)^2} \mathbb{E}[M^2] \\
&\leq \sigma^2 + 8\sqrt{\lambda \mathbb{E}\left[M^2\right]} \|Z\|_1 + \frac{2}{(1-2/\sqrt{\lambda})^2} \mathbb{E}\left[M^2\right]
\end{aligned}$$

Rademacher complexity

$\mathcal{F}$: countable class of measurable centered real-valued functions.

$$Z = \mathbb{E}\Big[\sup_{f \in \mathcal{F}} \Big| \sum_i \epsilon_i f(X_i) \Big| \mid X_1^n \Big]$$

$$M = \sup_{i,f} |f(X_i)| \, .$$

Tools derived from Burkholder-like inequalities:

$$\text{If } V \leq WZ, \quad \|(Z - \mathbb{E}[Z])_+\|_q \leq 2\sqrt{q \, \|W\|_q \, \mathbb{E}[Z]} + 2q \, \|W\|_q \ .$$

Rademacher complexity

Tools derived from Burkholder-like inequalities:

$$\text{If } V \leq WZ, \quad \|(Z - \mathbb{E}[Z])_+\|_q \leq 2\sqrt{q\,\|W\|_q\,\mathbb{E}[Z]} + 2q\,\|W\|_q \ .$$

$$Z_i = \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \Big|\sum_{j\neq i} \epsilon_j f(X_j)\Big| \mid X^{(i)}\Big]$$

$$\forall i, \qquad 0 \quad \leq \quad Z \ - \ Z_i \quad \leq \quad M \& \sum_i Z \ - \ Z_i \quad \leq \quad Z \text{➔} V \leq MZ \, .$$

Rademacher complexity

Tools derived from Burkholder-like inequalities:

$$\text{If } V \leq WZ, \quad \|(Z - \mathbb{E}[Z])_+\|_q \leq 2\sqrt{q \|W\|_q \mathbb{E}[Z]} + 2q \|W\|_q \ .$$

Hence:

$$\|(Z - \mathbb{E}[Z])_+\|_q \leq 2\sqrt{q \|M\|_q \mathbb{E}[Z]} + 2q \|M\|_q \ .$$

If $\mathcal{F}$ is uniformly-bounded, conditional Rademacher averages have sub-Poissonian tails.
If $\mathcal{F}$ is a VC-class of VC-dimension d, $\mathbb{E}[Z] \leq C\sqrt{dn}$, and

$$\mathbb{P}\{Z \geq \mathbb{E}[Z] + t\} \leq e^{-\frac{t^2}{2(\mathbb{E}[Z]+t/3)}} \ .$$

If the maximum of n envelops of $\mathcal{F}$ is q-integrable, so is Z.

References

References

- Concentration inequalities
  Massart.
  About the constants in Talagrand's concentration inequalities for empirical processes,
  Ann. Probab. **28**, 863-884, 2000.
  Boucheron, Lugosi and Massart.
  A sharp concentration inequality and applications,
  Rand. Struct. Alg., **16**, 277-292, 2000.
  Boucheron, Lugosi and Massart.
  Concentration inequalities by the entropy method,
  Ann. Probab. **31**, No. 3, 2003.
  Bousquet.
  Concentration inequalities for sub-additive funnctions. . 2002.
- Moment inequalities.
  Boucheron, Bousquet, Lugosi and Massart.
  Moment inequalities for functions of independent random variables,
  Manuscript , 2003.