Concentration inequalities & applications

S. Boucheron[1]

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

Belgium-Luxemburg Probability Days 20015
Liege, 30/01/2015

Motivations

Concentration with applications

Concentration of measure

- became a topic of interest in the 1970's,
- probability in Banach spaces (high dimensional probability) as a driving force
- a general attempt to handle smooth random variables in high dimensional ($n > 2$) product spaces
- an attempt to split investigation on functions of independent random variables in two parts
  . analysis of expectation
  . analysis of fluctuations around mean or expectation

Applications

- random combinatorics (J.M. Steele. *Probability theory and combinatorial optimization*. SIAM. 1997, C. McDiarmid Concentration in *Probabilistic Methods for Algorithmic Discrete Mathematics*. Springer. 1998)
- probability and analysis (M. Ledoux. *Concentration of Measure Phenomenon*. AMS. 2001)
- statistics (P. Massart. *Saint-Flour Lecture Notes*. 2003, V. Koltchinskii *Saint-Flour Lecture Notes*. 2008)

Context

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

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

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

Goal : upper-bounds on

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

$$\mathbb{P}\{Z \geq \mathbb{E}Z + t\} \quad \text{and} \quad \mathbb{P}\{Z \leq \mathbb{E}Z - t\} \quad \text{for} t > 0$$

Context ...

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

- high dimensional geometry ;
- random combinatorics ;
- statistics ;

A variety of methods

- Martingales
- Talagrand's induction method
- Transportation method
- Entropy method
- Chatterjee-Stein method (exchangeable pairs, size-biased couplings, ...)

Inspiration: Gaussian concentration

Theorem: Tsirelson, Borell, Gross,..., 1975

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

$$\left|F(x_1, \ldots, x_n) - F(y_1, \ldots, y_n)\right|^2 \leq L^2 \sum_{i=1}^{n} (x_i - y_i)^2$$

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

$$\mathrm{var}[Z] \leq L^2 \qquad \text{(Poincaré's inequality)}$$

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

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

▸ More ...

Outline

- → Motivations
- → The Entropy Method
- → Concentration for self-bounding functionals
- → Suprema of empirical processes
- → Emprical excess risk, Wilks phenomenon
- → Order statistics
- → Tail Index Estimation

The Entropy Method

Efron-Stein-Steele inequalities (1981)

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

$X'_1, ..., X'_n \sim X_1, ..., X_n$ but $\perp\!\!\!\perp$ from $X_1, ..., X_n$.
For each $i \in \{1, ..., n\}$

- $Z'_i = F(X_1, ..., X_{i-1}, X'_i, X_{i+1}, ...X_n)$.
- $X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$
- F_i: a function of $n-1$ arguments
- $Z_i = F_i(X_1, ..., X_{i-1}, X_{i+1}, X_n) = F_i(X^{(i)})$.

Theorem (Jackknife estimates of variance are biased)

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

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

ESS inequalities are a key-ingredient in pedestrian proofs of the Gaussian Poincaré inequalities

▶ Proof ...

Exponential Efron-Stein-Steele inequalities

In many circumstances, the search for concentration inequalities consists in extending Efron-Stein-Steele inequalities to higher, possibly exponential moments

Theorem (Sub-Gaussian behavior)

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

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

Uniformly bounded ESS estimates lead to the so-called bounded-differences inequality

Exponential Efron-Stein-Steele inequalities

Integrability of ESS estimates of variance may provide us with upper bounds on logarithmic moment generating functions

The next result has this flavor

Theorem (B., Lugosi and Massart, 2003)

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

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

This bound may be useful (easy path to concentration inequalities for suprema of bounded centered empirical processes. Massart. Saint-Flour Notes. 2003) but it is not always tight (order statistics. B. and Thomas. ECP. 2012).

Entropy method

Entropy

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

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

Why ?

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

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

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

Bounds on entropy

The entropy, like the variance, has the subadditivity property

Subadditivity

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

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

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

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

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

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

Entropy method in a nutshell

Combining subadditivity of entropy and the minimization of Bregman divergence by expected value:

Theorem (a modified logarithmic sobolev inequality.)

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

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

Summary

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

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

Concentration for self-bounding functionals

Self-bounding property

Definition

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

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

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

-

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

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

A popular choice

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

Self-bounding property: examples (i)

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

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

  This comprises binomial random variables

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

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

Self-bounding property: examples (ii)

Definition (Conditional Rademacher averages)

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

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

Symmetrization inequalities

(Giné and Zinn, 1984)

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

Binomial-Poisson tails

For suprema of bounded positive empirical processes, if $|\mathcal{T}| = 1$, Bennett inequality holds:

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

and

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

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

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

Self-bounding property and concentrations inequalities

Usual suspects

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

and

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

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

If Z satisfies the self-bounding property,

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

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

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

Corollary

Proposition (B., Lugosi and Massart 2003)

Conditional Rademacher averages are self-bounding.

Consequences

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

Conditional Rademacher averages : kind of weighted Bootstrap. In the early 2000's, Rademacher averages became popular empirical complexity measures in Statistical Learning Theory. At least theoretically, they are a key ingredient in the construction and analysis of model selection methods (Koltchinskii, Saint-Flour Notes. 2008)

Patterns of proof (Bis)

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

Starting point

Theorem (a modified logarithmic sobolev inequality.)

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

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

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

Proofs (...)

The proof proceeds in two steps

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

For self-bounding functions $\mathrm{Ent}\left[e^{\lambda Z}\right] \leq \phi(-\lambda)\mathbb{E}Z$ reads as

$$[\lambda - \phi(-\lambda)]\, G'(\lambda) - G(\lambda) \leq v\phi(-\lambda), \tag{2}$$

where $v = \mathbb{E}Z$.

Solving the differential inequality

In the self-bounding case, an explicit solution is easily derived.
In other cases, we may content ourselves with upper bounds on solutions

▸ More ...

Outline

- → Motivations
- → The Entropy Method
- → Concentration for self-bounding functionals
- → Suprema of empirical processes
- → Emprical excess risk, Wilks phenomenon
- → Order statistics
- → Tail Index Estimation

Suprema of empirical processes

Suprema of non-centered empirical processes

Well-understood scenarios

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

In the second (centered) setting, the self-bounding property is not satisfied, but careful computations starting from the modified logarithmic Sobolev inequality lead to the same differential inequality as in the self-bouding case (with a different variance upper bound).

In both cases, the entropy method delivers a Bennett inequality with variance factor coinciding with the Efron-Stein estimate of variance

Talagrand-...-Bousquet inequality

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

Efron-Stein estimate of variance

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

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

For $x > 0$

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

And now ...

The Talagrand-. . .-Bousquet inequality represents the most sophisticated and probably the most useful concentration inequality used in statistics and statistical learning theory.

Combined with old techniques such as peeling and chaining (refined union bounds), it allows us to derive sharp non-asymptotic risk bounds when loss functions are bounded but lack the usual smoothness properties, as in learning theory.

Emprical excess risk, Wilks phenomenon

Emprical risk minimization (Lerning theory)

Binary classification

P an unknown probability distribution over $\mathcal{X} \times \{0,1\}$. $(X_i, Y_i)_{i\in\mathbb{N}} \sim_{i.i.d.} P$.

$((X_i, Y_i)_{i\le n})$: training set (the sample)

Problem : find a classifier $f : \mathcal{X} \to \{0,1\}$ so as to minimize $P\{Y_{i+1} \neq f(X_{i+1})\}$

Bayes classifier: predict 1 if $\mathbb{E}[Y \mid X = X_{i+1}] > 1/2$ (P-dependent)

Empirical risk minimization

Use a dictionary of classifiers $\mathcal{F}$, predict using $\widehat{f}$ that minimizes $R_n(f) = \sum_{i=1}^n \mathbb{I}_{f(X_i)\neq Y_i}$ as a proxy for minimizing $R(f) = \mathbb{E}\mathbb{I}_{f(X)\neq Y}$ over $\mathcal{F}$

Fits into the contrast minimization setting

Contrast is not smooth

Statistical learning theory $\to$ avoid parametric assumptions on P

Concentration inequalities and Empirical Risk Minimization

If $\bar{f}$ minimizes $R(f)$ over $\mathcal{F}$,

$$\underbrace{R(\widehat{f}) - R(\bar{f})}_{\text{excess risk}} \leq R(\widehat{f}) - R_n(\widehat{f}) + R_n(\bar{f}) - R(\bar{f})$$

The excess risk is dominated by the increment of a centered bounded empirical process between $\bar{f}$ and $\widehat{f}$ (a random position).

Talagrand's inequality, allows us to control the fluctuation of reweighted versions of the empirical process

$$\sup_{f \in \mathcal{F}} \frac{R(f) - R_n(f) + R_n(\bar{f}) - R(\bar{f})}{r^2 + R(f)}$$

where $r > 0$ is a tunable parameter.

This leads to sharp deviation inequalities for excess risk (see Massart, Koltchinskii, Saint-Flour Notes)

Another scenario : Excess Empirical Risk

Mind the notational gap !!!

$$X_{i,s} := \mathbb{I}_{f(X_i)\neq Y_i} \qquad f \leftrightarrow s, \quad X_i \leftrightarrow (X_i, Y_i)$$

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

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

Excess risk and empirical counterpart

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

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

Consequences of Talagrand's inequalities (and peeling)

Assumptions

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

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

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

Definition

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

Consequences, cont'd

With probability larger than $1-\delta$

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

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

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

Combining ...

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

Variance bounds for EER

- Consequences of Efron-Stein inequalities :

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

Bernstein inequality for EER

Using

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

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

and tail bounds for excess risk (Massart, Koltchinskii Saint-Flour Notes) leads to

Bernstein like inequality (B., Massart 2011)

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

Variance factor $n\omega(r_*)^2$
Scale factor $\sqrt{n}\omega\left(\frac{\omega(r_*)}{\sqrt{n}r_*}\right)$

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

Outline

- → Motivations
- → The Entropy Method
- → Concentration for self-bounding functionals
- → Suprema of empirical processes
- → Emprical excess risk, Wilks phenomenon
- → Order statistics
- → Tail Index Estimation

Order statistics

Central, intermediate and extreme order statistics

$X_1, \ldots, X_n \sim_{\text{i.i.d.}} F$

Order statistics

$X_{1,n} \geq \ldots \geq X_{n,n}$ non-increasing rearrengement of $X_1, \ldots, X_n$.
If n clear from context, $X_{1,n}, \ldots, X_{n,n}$ denoted by $X_{(1)}, \ldots, X_{(n)}$.

$(X_{k,n})$ is a sequence of

extreme order statistics,	if k fixed, $n \to \infty$;
central order statistics,	if $k/n \to p \in (0,1)$ while, $n \to \infty$;
intermediate order statistics,	if $k/n \to 0$, $k \to \infty$.

Different asymptotics (if any)

Central and intermediate order statistics (often): Gaussian
Extreme order statistics (sometimes): Generalized Extreme Value

Variance bounds, order statistics and spacings

A connection

The variance (and more generally the higher moments) of the k^{th} order statistics can be upper-bounded by moments of the k^{th} *spacing* $X_{(k)} - X_{(k+1)}$.

Lemma (Jackknife bounds)

$$\mathrm{Var}[X_{(k)}] \leq k\mathbb{E}\left[\left(X_{(k)} - X_{(k+1)}\right)^2\right].$$

Convention

$$\Delta_k = X_{(k)} - X_{(k+1)}$$

Beyond variance

The connection between the variance of order statistics (which may be infinite) and the expectation of spacings can be exploited under special conditions

It may lead to concentration inequalities for extreme, intermediate and central order statistics in a seamless way

Special conditions boil down to asserting non-decreasing hazard rate (a condition which is implied by log-concavity)

In order to use the entropy method, we resort to a change of representation trick

Rényi's representation

The order statistics of an exponential sample ...

are partial sums of independent exponentially distributed random variables.

If $F(x) = 1 - \exp(-x)$ for $x > 0$, letting $X_{n+1,n} = 0$,

$$X_{k,n} = \sum_{i=k}^{n} (X_{i,n} - X_{i+1,n})$$

where the spacings $\Delta_i = (X_{i,n} - X_{i+1,n})_{i=1,\ldots,n}$ form an independent family of random variables and $i \times (X_{i,n} - X_{i+1,n}) \sim F$

Quantile transformation

Definition (Quantile function (bis))

$$F^{\leftarrow}(p) = \inf\{x : F(x) \geq p\}, p \in (0,1) \qquad U(t) = F^{\leftarrow}(1 - 1/t), t \in (1, \infty)$$

Representation for order statistics

If $Y_{(1)}, \ldots, Y_{(n)}$ are the order statistics of an exponential sample, then

$$U(e^{Y_{(1)}}) \geq U(e^{Y_{(2)}}) \geq \ldots \geq U(e^{Y_{(n)}})$$

is distributed as the order statistics of a sample drawn according to F.

Hazard rate, spacings and order statistics

Definition (Hazard rate)

The hazard rate of a differentiable distribution function F is $F'/\overline{F} = F'/(1-F)$.

Observation

The distribution function F has non-decreasing hazard rate, iff $U \circ \exp$ is concave.

Lemma

If the distribution function F has non-decreasing hazard rate, then $X_{(k+1)}$ and $\Delta_k = X_{(k)} - X_{(k+1)}$ are negatively associated.

Negative association

For increasing functions f, g

$$\mathbb{E}\left[f(X_{(k+1)}g(\Delta_k)\right] \leq \mathbb{E}\left[f(X_{(k+1)}\right]\mathbb{E}\left[g(\Delta_k)\right]$$

Application to order statistics

Notation

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

Lemma

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

$$\begin{aligned} \mathrm{Ent}\left[e^{\lambda X_{(k)}}\right] &\leq k\mathbb{E}\left[e^{\lambda X_{(k+1)}} \psi\left(\lambda(X_{(k)} - X_{(k+1)})\right)\right] \\ &= k\mathbb{E}\left[e^{\lambda X_{(k+1)}} \psi(\lambda \Delta_k)\right] \end{aligned}$$

Proof parallels the variance bounds derived from Efron-Stein inequalities

Exponential Efron-Stein inequality for order statistics

$V_k = k\Delta_k^2$: the Efron-Stein estimate of the variance of $X_{(k)}$.

Theorem (B. and Thomas, 2012)

If F has non-decreasing hazard rate, then for $\lambda \geq 0$, and $1 \leq k \leq n/2$,

$$\begin{aligned} \log \mathbb{E} e^{\lambda(X_{(k)} - \mathbb{E}X_{(k)})} &\leq \lambda \frac{k}{2} \mathbb{E}\left[\Delta_k \left(e^{\lambda \Delta_k} - 1\right)\right] \\ &= \lambda \frac{k}{2} \mathbb{E}\left[\sqrt{\frac{V_k}{k}} \left(e^{\lambda \sqrt{V_k/k}} - 1\right)\right]. \end{aligned}$$

This ad hoc exponential Efron-Stein inequality exhibits what should be the correct dependence between the exponential integrability the Efron-Estimate of variance and the exponential inequality of the function

Comments

Massaging the previous inequality provide us with correct concentration bounds for order statistics of Gaussian samples

Maxima are 1-Lipschitz functions of Gaussian samples. But Gaussian Poincaré and logarithmic Sobolev inequalities do not deliver tight bounds in that case.

They need to be complemented by stronger ad hoc arguments ($L_1 - L_2$ inequality, see Chatterjee on Superconcentration, or careful semi-group analysis, I. Nourdin)

The exponential Efron-Stein inequality for order statistics also provides correct concentration inequalities for beta-distributed random variables (order statistics of uniform samples)

Tail Index Estimation

Background : Tail index estimation

Extreme value analysis

Whereas learning theory has been a playground and a driving force for the development of concentration inequalities, for a long time, Extreme Value Analysis (EVT) seemed to be immune to the concentration of measure phenomenon

EVT has traditionally been regarded as an asymptotic theory

but the simplest (and most basic) statistical problems in EVT raise model selection issues that somehow could benefit from the availability of non-asymptotic tail bounds

Background : EVT

Stability of conditional excess distribution

$$\lim_{t \to F^{\leftarrow}(\infty)} \frac{\overline{F}(t + \sigma_t x)}{\overline{F}(t)} = \overline{G}(x)$$

for some non-degenerate distribution function G ($F \in \mathrm{MDA}(G)$) and some positive measurable scale function $t \mapsto \sigma_t$.

POT-stable distributions

Limit distributions are in the scale family generated by generalized Pareto distributions : $\overline{G}_\gamma(x) = (1 + \gamma x)^{-1/\gamma}$ $\gamma \in \mathbb{R}, 1 + \gamma x > 0$
γ is called the tail index.

F belongs to the domain of attraction of G_γ iff $U = (\overline{F})^{\leftarrow}$ has the extended regular variation property

$$\lim_{t \to \infty} \frac{U(tx) - U(t)}{U(ty) - U(t)} \text{ exists for all } x, y$$

the limit is $\int_1^x v^{\gamma-1} \mathrm{d}v / \int_1^y v^{\gamma-1} \mathrm{d}v$

Tail index estimation

Assuming $F \in \mathrm{MDA}(\gamma)$, and given $X_1, \ldots, X_n \sim_{i.i.d.} F$, a basic inferential problem consists of estimating γ from the data

The fact that $F \in \mathrm{MDA}(\gamma)$ only impacts the tail of F.
Belonging to a domain of attraction is an asymptotic property of F or U

Inference on the tail index has to be based on the tail of the empirical distribution or equivalently on the tail empirical quantile function

Peaks over thresholds

Functions of largest order statistics

$$\widehat{\gamma}(k) = F(X_{(1)}, \ldots, X_{(k+1)})$$

Hill estimator

$$\widehat{\gamma}(k) = \frac{1}{k} \sum_{i=1}^{k} \ln \frac{X_{(i)}}{X_{(k+1)}} = \frac{1}{k} \sum_{i=1}^{k} i \ln \frac{X_{(i)}}{X_{(i+1)}}$$

Works for $\gamma > 0$ (Fréchet domain)

Pickands estimator

$$\widehat{\gamma}(k) = \log_2 \frac{X_{(k)} - X_{(2k)}}{X_{(2k)} - X_{(4k)}}$$

Tail index estimators as smooth tail functionals

Asymptotic analysis establishes that under midl assumptions, tail index estimators recentered around their expectation and suitably rescaled are asymptotically Gaussian

Using the same representational trick as for order statistics, we describe a non-asymptotic view at this smoothness property

This provides a (relatively) transparent analysis of an adaptive strategy for choosing the number of order statistics used in tail index estimation

Exponential representation (bis)

Quantile transform

$U(t) := \overline{F}(1 - 1/t)$ for $t > 1$

$$X \sim U(\exp(Y)) : \qquad \text{with } Y \sim \text{Exponential}$$

Karamata's representation

$$U(t) = c(t)t^{\gamma} \exp \int_1^t \frac{\eta(s)}{s} \mathrm{d}s \qquad \lim_t c(t) = c \qquad \lim_s \eta(s) = 0$$

Von Mises conditions: $c(t) = c$

Rényi's representation

Order statistics of exponential samples are partial sums

$$(Y_{(1)} \geq \ldots \geq Y_{(n)}) \sim \left(\sum_{i=k}^{n} \frac{E_i}{i} \right)_{k \in \{1,\ldots,n\}} \qquad \text{where} \quad E_i \sim_{\text{i.i.d.}} \text{Exponential}$$

Hill estimators as functions of independent exponential random variables

Combining Rényi's representation, Karamata's representation, the quantile transform and the Von Mises condition ...

Hill estimators as functionals of an exponential sample

$$\left(\widehat{\gamma}(k)\right)_{2\le k\le n} \sim \left(\frac{1}{k}\sum_{i=1}^{k}\int_0^{E_i}\left(\gamma+\eta(\mathrm{e}^{\frac{u}{i}+\Upsilon_{(i+1)}})\right)\mathrm{d}u\right)_{k<n}$$

where $E_1,\ldots,E_n \sim_{i..i.d.}$ standard exponentials, and $\Upsilon_{(k)} = \sum_{j=k}^{n} E_j/j$

Hill estimators are *approximately* distributed like partial sums of independent exponential random variables

Bias

Bias & conditional bias

The bias of the Hill estimator admits a simple integral representation

$$\mathbb{E}\left[\widehat{\gamma}(k) - \gamma\right] = \mathbb{E}\left[b\left(\mathrm{e}^{Y_{(k+1)}}\right)\right]$$

where

$$b(t) = t \int_t^\infty \frac{\eta(v)}{v^2} \mathrm{d}v$$

is a smooth function of $Y_{(k+1)}$ ($k+1$-th order statistic of an exponantial samble, concentrated around $\ln n/k$)

Lower bounds: adaptivity has a price (Carpentier & Kim (2014))

Let $\rho < -1$. Let v belong to $(0, \mathrm{e}/((1+2\mathrm{e}))$. For any tail index estimator $\widehat{\gamma}$, for any sample size n such that $\lfloor \ln n \rfloor > \mathrm{e}/v$, let $M = \lfloor \ln n \rfloor$, then there exists a probability distribution $P \in \mathrm{MDA}(\gamma), \gamma > 0$ satisfying the von Mises condition with von Mises function η satisfying

$$\overline{\eta}(t) \leq t^{\rho}$$

where $\rho = \rho_0$ and such that

$$P^{\otimes n}\left\{|\widehat{\gamma} - \gamma| \geq \frac{\kappa_\rho}{4}\gamma\left(\frac{v \ln\ln n}{n}\right)^{|\rho|/(1+2|\rho|)}\right\} \geq \frac{1}{1+2\mathrm{e}}$$

and

$$\mathbb{E}_P\left[\frac{|\widehat{\gamma} - \gamma|}{\gamma}\right] \geq \frac{\kappa_\rho}{4(1+2\mathrm{e})}\left(\frac{v \ln\ln n}{n}\right)^{|\rho|/(1+2|\rho|)},$$

with $\kappa_\rho = \exp(-1/(1+2|\rho|)^2)$.

If ρ were known (insider deal), we could get rid of $(\ln\ln n)^{|\rho|/(1+2|\rho|)}$.

Poincaré inequality for function of exponential samples

Poincaré inequality for exponential samples

If

- f is a differentiable function over $\mathbb{R}^n$,
- $Z = f(E_1, \ldots, E_n)$
  where $E_1, \ldots, E_n$ are independent standard exponential random variables,

then

$$\mathrm{Var}(Z) \leq 4\mathbb{E}\left[\|\nabla f\|^2\right].$$

- In dimension 1, the proof boils down to Cauchy-Schwarz inequality
- Higher-dimensional statements follow from Efron-Stein inequalities
- The constant 4 cannot be improved

Variance bounds

Variance

$$-\frac{2\gamma}{k}\mathbb{E}\left[\overline{\eta}\left(e^{Y_{(k+1)}}\right)\right] \leq \mathrm{Var}[\widehat{\gamma}(k)] - \frac{\gamma^2}{k} \leq \frac{2\gamma}{k}\mathbb{E}\left[\overline{\eta}\left(e^{Y_{(k+1)}}\right)\right] + \frac{5}{k}\mathbb{E}\left[\overline{\eta}\left(e^{Y_{(k+1)}}\right)^2\right].$$

Assume von Mises function $\eta \in \mathrm{RV}_\rho, \rho \leq 0$, then for any intermediate sequence $(k_n)_n$

$$\lim_{n\to\infty} \frac{k_n \,\mathrm{Var}(\widehat{\gamma}(k_n)) - \gamma^2}{\eta(n/k_n)} = \frac{2\gamma}{(1-\rho)^2}.$$

The variance of Hill estimators is well understood, and well approximated by γ^2/k

Bias-variance tradeoff

Risk of Hill estimator

$$\mathbb{E}\left[(\gamma - \widehat{\gamma}(k))^2\right] = \mathrm{var}(\widehat{\gamma}(k)) + (\gamma - \mathbb{E}\widehat{\gamma}(k))^2$$

Best choice of k depends

on unknown η, or rather b ...

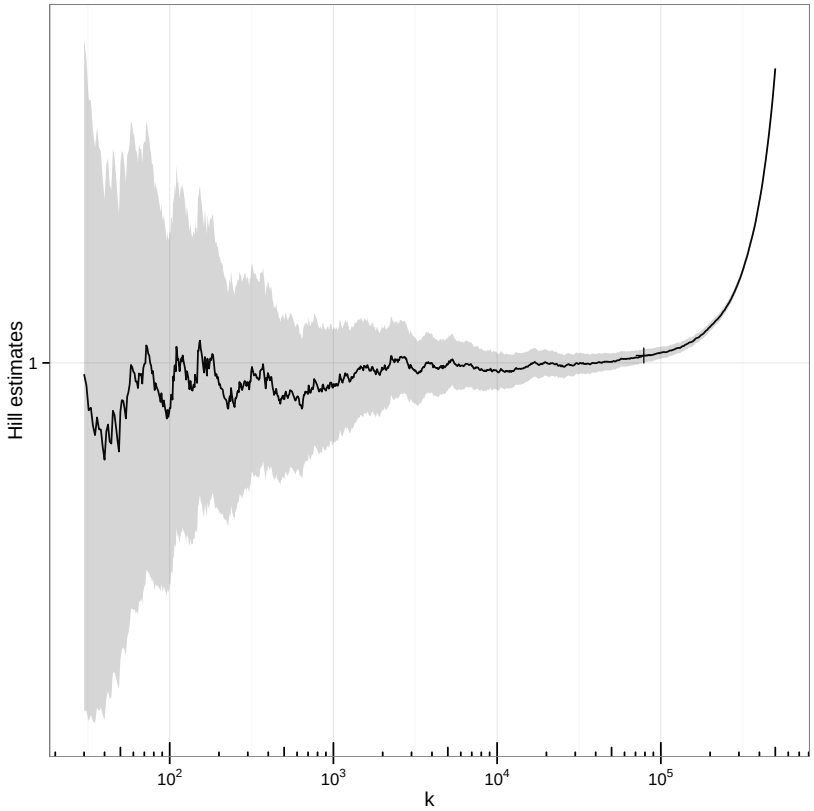
Since 1980,

Many attempts to select asymptotically best possible k under a second order restriction (assuming the bias b is regularly varying with index $\rho < 0$).
Starts with Hall-Welsch (1985), Danielson et al. (1998), Drees and Kaufmann (1998) and too many to name ...

Recently

Attempts to derive risk bounds only assuming that η or b decays sufficiently fast (without assuming second order regularity): see Carpentier and Kim (2014).

Hill plot



- Hill estimators for a Cauchy sample ($n = 10^6$)
- Extreme value index is 1
- Bias (function b) decays like $1/t$
- The gray area is an approximate 95% confidence region $(\widehat{\gamma}(k)(1 + \pm z_{.95}/\sqrt{k}))$

Data driven choice

Lepski's method

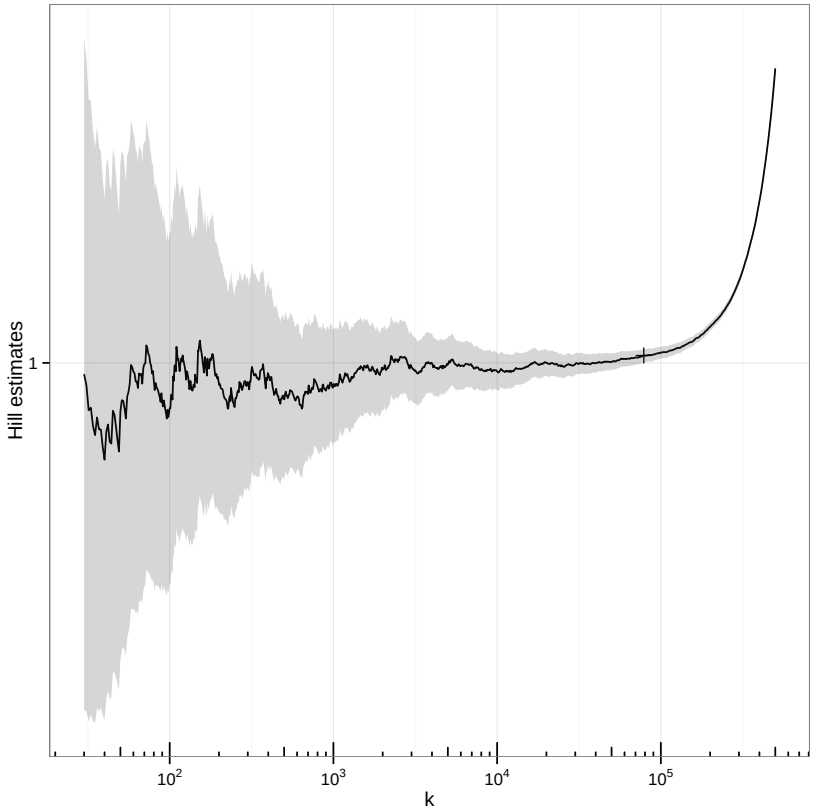
$$\widehat{k}(r_n) = \min\left\{k \in \{2,\dots,n\} : \max_{2\le i\le k} \sqrt{i}\frac{|\widehat{\gamma}(i) - \widehat{\gamma}(k)|}{\widehat{\gamma}(i)} > r_n\right\}$$

with $\sqrt{\ln\ln n} = O(r_n)$ and $r_n = O(\sqrt{n})$

Suggestions for tuning $r_n(\delta)$

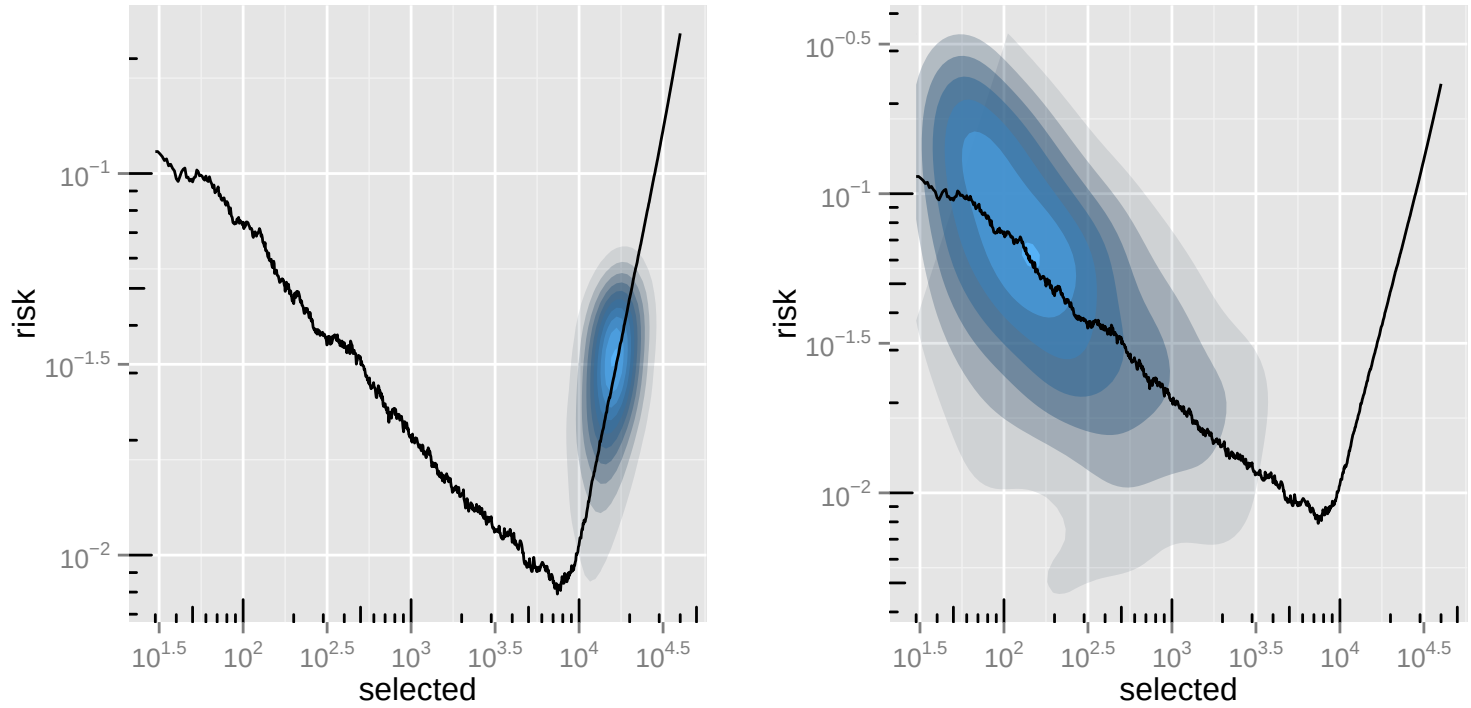
- $r_n(\delta) = 1$, and the ability to replace $\widehat{\gamma}(i)$ by $\mathbb{E}\widehat{\gamma}(i)$ would balance bias and standard deviation.
- $r_n(\delta) \lll \sqrt{\ln\ln n}$ is too conservative
- $r_n(\delta) \propto \sqrt{\ln\ln n}$?

Hill plot



- Hill estimators for a Cauchy sample ($n = 10^6$)
- Extreme value index is 1
- Bias (function b) decays like $1/t$
- The gray area is an approximate 95% confidence region $(\widehat{\gamma}(k)(1 + \pm z_{.95}/\sqrt{k})\ )$
- Point $+$ represents the index selected by Lepski's rule with $r_n = \sqrt{2 \ln \ln n}$.

A calibration problem: Hill estimators on samples from Cauchy distribution ($\gamma = 1$).



The risk of each Hill estimator is estimated from 500 simulations.
Each simulation is performed on a sample of size 10^5.

The selection procedure is carried out with $r_n = \begin{cases} \sqrt{3 \ln \ln n} & \text{(left)} \\ \sqrt{\ln \ln n} & \text{(right)} \end{cases}$

A typical situation

Selecting a reasonable index k amounts to betting that the target distribution fits into a model defined by an envelope condition on the bias b ($\sqrt{k}b(n/k) \approx \gamma$ or rather $\sqrt{k}b(n/k) \approx r_n\gamma$) .

In order to provide risk bounds for this selection procedure, we need to be able to control the whole sequence of random variables $(\widehat{\gamma}(k) - \gamma)_k$.

Good concentration inequalities can be piped into union bounds.

Talagrand's inequality (early 90's)

Concentration for smooth functional of exponential samples

If f is a differentiable function on $\mathbb{R}^n$ with $\max_i |\partial_i f| < \infty$, and $Z = f(E_1, \ldots, E_n)$ where $E_1, \ldots, E_n$ are independent standard exponential random variables.

- Let $c < 1$, then for all λ such that $0 \leq \lambda \max_i |\partial_i f| \leq c$,

$$\mathrm{Ent}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] \leq \frac{2\lambda^2}{1-c}\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\|\nabla f\|^2\right]$$

- Let v be the essential supremum of $\|\nabla f\|^2$. Then, Z is sub-gamma on both tails with variance factor $4v$ and scale factor $\max_i |\partial_i f|$

Maurey (1992) described an alternative proof using infimum -convolution arguments.
Bobkov and Ledoux (1997) gave a proof based on the entropy method. The result generalizes to log-concave distributions

Concentration/deviation inequalities for Hill process

For k such that $\ln(n) \le \ell \le k \le n$,
For $0 < \delta < 1/2$,
with probability larger than $1 - 3\delta$,

$$\begin{aligned}
&\max_{\ell \le i \le k} \sqrt{i} \left| \hat{\gamma}(i) - \mathbb{E}\hat{\gamma}(i) \right| \\
&\quad \le \quad \left(c_1 \sqrt{\ln \log_2 n} + c_1' + 2 \right) (\gamma + 2\overline{\eta}(n/k')) \\
&\qquad\quad + 4 \sqrt{\mathbb{E}\left[\overline{\eta}(\mathrm{e}^{Y_{(k+1)}})^2 \right]} + \ln \frac{2}{\delta} \, (2 + 4\overline{\eta}(1))
\end{aligned}$$

where $k' = k / \left(1 - \frac{\ln(\mathrm{e}/\delta)}{\sqrt{k}}\right)$.

$c_1 \le 4, c_1' \le 16$ are universal constants

Oracle inequalities: a generic result

A pivotal index

$$\widetilde{k}_n = \min\left\{k : 2 \le k \le n \text{ and } \max_{2\le i<k} \sqrt{i}\frac{\left|\mathbb{E}[\widehat{\gamma}(i) - \gamma]\right|}{\gamma} \ge r_n\right\} - 1$$

with $r_n = O(\sqrt{\ln\ln n})$.

B. and Thomas (2015)

$r_n(\delta) = 8c_4\sqrt{2\ln(2\log_2(n)/\delta)}$ $\qquad$ $\xi_n = c_1\sqrt{\ln\log_2(n)} + c_1'$ $\qquad$ $\widehat{k}_n = \widehat{k}(r_n(\delta))$

Assume that n is large enough so that

i) $\overline{\eta}(1) < c_5\xi_n\gamma$, with c_5 an universal constant,

ii) $\overline{\eta}\left(\frac{n}{\widetilde{k}_n + \frac{\ln(1/\delta)}{2}}\right) < \gamma/4$.

With probability larger than $1 - 5\delta$,

$$\left|\gamma - \widehat{\gamma}(\widehat{k}_n)\right| \le \left|\gamma - \widehat{\gamma}(\widetilde{k}_n)\right|\left(1 + \frac{r_n(\delta)}{\sqrt{\widetilde{k}_n}}\right) + \frac{r_n(\delta)\gamma}{\sqrt{\widetilde{k}_n}}\,.$$

Oracle inequalities: Enveloppe conditions for the bias

If, for some $C > 0$ and $\rho < 0$, $\left|\gamma - \mathbb{E}\hat{\gamma}(k)\right| \leq C\left(\frac{n}{k}\right)^{\rho}$ then

$$\sqrt{\tilde{k}_n + 1} \geq \left(\frac{\gamma r_n}{C}\right)^{1/(1+2|\rho|)} n^{|\rho|/(1+2|\rho|)}.$$

and, with probability larger than $1 - 5\delta$,

$$\begin{aligned}
&\left|\widehat{\gamma}(\widehat{k}_n) - \gamma\right| \\
&\quad \leq \quad \gamma\left(\frac{\gamma r_n}{C}\right)^{-1/(1+2|\rho|)} n^{-|\rho|/(1+2|\rho|)} \left(r_n(\delta) + \left(1 + \frac{r_n(\delta)}{\sqrt{\tilde{k}_n}}\right)\left(r_n + \left(\sqrt{8\ln\frac{2}{\delta}} + \frac{\ln\frac{2}{\delta}}{\sqrt{\tilde{k}_n}}\right)\right)\right) \\
&\qquad + \left(1 + \frac{r_n(\delta)}{\sqrt{\tilde{k}_n}}\right) \frac{8\overline{\eta}(\mathrm{e}^{Y_{(\tilde{k}_n+1)}})}{\sqrt{\tilde{k}_n}}
\end{aligned}$$

which matches the lower bound

Appendix

Proof of ESS inequalities

$\mathbb{E}_i Z := \mathbb{E}[Z \mid X_1, \ldots, X_i]$ and $\Delta_i := \mathbb{E}_i Z - \mathbb{E}_{i-1} Z$ for $i = 1, \ldots, n$.

$$Z - \mathbb{E}Z = \sum_{i=1}^{n} \Delta_i \qquad \text{(Doob's martingale)} \qquad \Rightarrow \mathrm{Var}(Z) = \sum_{i=1}^{n} \mathbb{E}\Delta_i^2$$

Note

$$\mathbb{E}_i \left[\mathbb{E}^{(i)} Z \right] = \mathbb{E}_{i-1} Z \quad \Rightarrow \quad \Delta_i = \mathbb{E}_i \left[Z - \mathbb{E}^{(i)} Z \right]$$

By Conditional Jensen's inequality

$$\Delta_i^2 \leq \mathbb{E}_i \left[\left(Z - \mathbb{E}^{(i)} Z \right)^2 \right]$$

From $\mathrm{Var}(Z) = \sum_{i=1}^{n} \mathbb{E}\Delta_i^2$,

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

Now

$$\mathrm{Var}^{(i)}(Z) = \frac{1}{2} \mathbb{E}^{(i)} \left[(Z - Z_i')^2 \right] = \mathbb{E}^{(i)} \left[(Z - Z_i')_+^2 \right]$$

Back to ESS inequalities ...

Inspiration: Gaussian concentration

The Gaussian concentration inequality is

- tight
- dimension-free

It can be proved in many ways

- Semi-group methods (many flavors)
- Correlation inequalities
- Bobkov's method
- Entropy method

Back to Gaussian concentration ...

Gross logarithmic Sobolev inequality

The entropy approach to Gaussian concentration proceeds through functional inequalities

Theorem (Gross,. . . , 1975)

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

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

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

Pick $F(X_1, \ldots, X_n) = \exp(\lambda f(X_1, \ldots, X_n)/2)$ where f is differentiable, Gross's inequality reads as

$$\mathrm{Ent}\left(e^{\lambda f(X_1,\ldots,X_n)}\right) \leq \frac{\lambda^2}{2}\mathbb{E}\left[e^{\lambda f(X_1,\ldots,X_n)}\left\|\nabla f\right\|^2\right]$$

◂ Back to Gaussian concentration ...

Proofs (...)

Start from the modified logarithmic Sobolev inequality

$$\begin{aligned}
\mathrm{Ent}\left[e^{\lambda Z}\right] &\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z}\phi\left(-\lambda(Z-Z_i)\right)\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z}\phi\left(-\lambda\right)\left(Z-Z_i\right)\right] \\
&\quad \text{as } \phi \text{ is convex, } \phi(0)=0 \text{ and } Z-Z_i \in [0,1] \\
&\leq \phi(-\lambda)\mathbb{E}\left[e^{\lambda Z} Z\right] \\
&\quad \text{as } Z \text{ is self-bounding}
\end{aligned}$$

Elementary computations lead to

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

Back ...

Proofs (...)

Letting $G(\lambda) = \log \mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}$, multiplying by e^{λ}, the preceding inequality reads as

$$(e^{\lambda} - 1)G'(\lambda) - e^{\lambda}G(\lambda) \leq (e^{\lambda}(\lambda - 1) + 1)\mathbb{E}Z$$

which is readily solved

$$G(\lambda) \leq \mathbb{E}Z\left(e^{\lambda} - \lambda - 1\right)$$

for $\lambda \in \mathbb{R}$

This approach is called the Herbst argument. In many circumstances it is not as easy: the differential inequality may not be solved explicitly. Then we may use comparison theorems to upper bound solutions of (difficult) differential inequalities by solutions of easier ones.

◂ Back ...