Concentration inequalities, the entropy method, search for *super*-concentration

Concentration, ...

S. Boucheron[1]

[1] LPMA CNRS & Université Paris-Diderot

Stein's Method Colloquium, Borchard Foundation, June 29th - July 2nd 2014

Part I

Introduction

Outline

Concentration inequalities ...

Concentration inequalities extend exponential inequalities for sums of independent random variables (Hoeffding, Bennett, Bernstein, ...)

Example: Hoeffding inequality

$X_1, \ldots, X_n$ independent r.v. with $a_i \leq X_i \leq b_i$ for each $i \leq n$, $Z = \sum_{i=1}^n X_i$

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \frac{(b_i - a_i)^2}{4} =: v\,.$$

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

Concentration inequalities ...

There is nothing special about sums

Concentration in product spaces

Any *smooth* function of many independent random variables that does not depend too much on any of them is concentrated around its mean value

But ...

the right notion(s) of smoothness are not obvious

Gaussian setting

Cirelson inequality (1975)

$X_1, \ldots, X_n \sim_{\text{i.i;d}} \mathcal{N}(0,1)$ standard Gaussian vector

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

$$f \quad L-\text{Lipschitz} \quad \Rightarrow \quad \mathbb{P}\left\{Z \geq \mathbb{E}Z + t\right\} \leq \exp\left(-\frac{t^2}{2L^2}\right)$$

Lipschitz functions of standard Gaussian vectors are sub-Gaussian

This inequality is dimension-free.

Concentration inequalities and beyond

Concentration inequalities are just a component of a more general *concentration of measure phenomenon* which stems from Geometric Functional Analysis (See Ledoux, 2001).

There are many ways to derive concentration inequalities:

- martingales (McDiarmid, 1998).
- transportation (Martin 1996).
- induction and ingenuity (Talagrand 1996, 2014),
- tailorings of Stein's method (Chatterjee 2006, Chen, Goldstein and Shao 2010, Ross 2011).

The so-called entropy method starts from functional inequalities satisfied by Gaussian, Product, ... measures and builds on those functional inequalities to derive concentration inequalities. The roots of the entropy method go back to advances in Functional Analysis during the 1970's. It become increasing popular during the last two decades thanks to M. Ledoux modular derivation of Talagrand's functional Bennett inequality (1996).

Gaussian concentration and function inequalities

Gaussian concentration may be characterized by functional inequalities

$X = (X_1, \ldots, X_n)$ a standard Gaussian vector
f a differentiable function

Poincaré	$\operatorname{Var} f(X) \leq \mathbb{E}\|\nabla f\|^2$
Logarithmic Sobolev	$\operatorname{Ent}(f(X)^2) \leq 2\mathbb{E}\|\nabla f\|^2$
Modified Logarithmic Sobolev	$\operatorname{Ent}(f(X)) \leq 2\mathbb{E}\frac{\|\nabla f\|^2}{f}$

where $\operatorname{Ent}(f(X)) = \mathbb{E}f(X) \log f(X) - \mathbb{E}f(X) \log \mathbb{E}f(X)$.

Those inequalities are dimension-free.

From logarithmic Sobolev inequality to Gaussian concentration: Herbst's argument

$g : \mathbb{R}^n \to \mathbb{R}$ differentiable, $Z = g(X_1, \ldots, X_n)$ with $\|\nabla g\| \leq L$.
Apply the Logarithmic Sobolev Inequality to $f(X_1, \ldots, X_n) = \exp\left(\frac{\lambda}{2} g(X_1, \ldots, X_n)\right)$

$$\text{Ent}\left[\mathrm{e}^{\lambda g}\right] \leq \frac{\lambda^2}{2}\mathbb{E}\left[\|\nabla g\|^2 \mathrm{e}^{\lambda g}\right] \leq \frac{\lambda^2 L^2}{2}\mathbb{E}\left[\mathrm{e}^{\lambda g}\right]$$

Solving

$$\frac{\mathrm{d}\frac{1}{\lambda}\log \mathbb{E}\mathrm{e}^{\lambda(g-\mathbb{E}g)}}{\mathrm{d}\lambda} \leq \frac{L^2}{2}$$

leads to

$$\log \mathbb{E}\mathrm{e}^{\lambda(g-\mathbb{E}g)} \leq \frac{L^2\lambda^2}{2}$$

which leads to Cirelson's inequality by Markov inequality.

Concentration in product spaces

How can we connect the fluctuations of a function of many independent random variables with the smoothness of the function?

A first step consists in bounding the variance

A second step consists in deriving bounds on the logarithmic moment generating function which reflect the variance upper bounds

Smoothness

Smoothness in product spaces may be defined with respect to ...

▷ Hamming distance: there exists $c_1, \ldots, c_n$

$$|f(x_1, \ldots, x_n) - f(y_1, \ldots, y_n)| \leq \sum_{i=1}^{n} c_i \mathbb{I}_{x_i \neq y_i} \quad \forall y_1, \ldots, y_n$$

▷ Suprema of weighted Hamming distances: $\forall x_1 \ldots, x_n \quad \exists c_i(x_1, \ldots, x_n)$,

$$f(x_1, \ldots, x_n) - f(y_1, \ldots, y_n) \leq \sum_{i=1}^{n} c_i(x_1, \ldots, x_n) \mathbb{I}_{x_i \neq y_i} \quad \forall y_1, \ldots, y_n$$

▷ Euclidean distance: $\exists L, \forall, x_1, \ldots, x_n \quad y_1, \ldots, y_n$

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

Self-bounding functions

A example of smoothness

$f : \mathcal{X}^n \to \mathbb{R}$ is self-bounding if for all $i \leq n$,

$$\exists f_i \, : \, \mathcal{X}^{n-1} \to \mathbb{R}, \quad 0 \leq f(x_1, \ldots, x_n) - f_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \leq 1$$

$$\sum_{i=1}^{n} f(x_1, x_2, \ldots, x_n) - f_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \leq f(x_1, x_2, \ldots, x_n)$$

Examples

Longest increasing subesequence, Empirical VC-dimension, Empirical VC-entropy, Conditional Rademacher complexity, ...

Self-boundedness and concentration

Starting from a modified logarithmic Sobolev inequality, using a variation of Herbst's argument leads to

Sub-Poisson concentration

If f is self-bounding and $Z = f(X_1, \ldots, X_n)$

$$\log \mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} \leq \mathbb{E}Z\left(e^{\lambda} - \lambda - 1\right) \quad \lambda \in \mathbb{R}$$

B., Lugosi and Massart, 2000-3

The tails of self-bounding functions are not heavier than those of a Poisson distribution with the same expectation.

Smoothness may not be enough

Off the shelf inequalities (like the concentration inequality for self-bounding functions)

may fail to capture some aspects of the concentration phenomenon.

Longest increasing subesequence

$X_1, \ldots, X_n \sim$ uniform on $[0,1]$

$$Z = \max\left\{k : \exists 1 \leq i_1 < i_2 < \ldots < i_k \leq n \text{ with } X_{i_1} < \ldots < X_{i_k}\right\}$$

$$\mathbb{E}Z = (1+o(1))2\sqrt{n} \quad \text{but} \quad \mathrm{Var}(Z) = O\left(n^{1/3}\right)!!!!$$

The Longest Increasing Subsequence in a sequence of independent random real (LIS in a random permutation) is an example of self-bounding random variable that concentrates more than predicted.
This is an example of superconcentration.

Beyond sub-Gaussian, sub-Poissonian scenarii

Traditionally

Methods dedicated to establishing concentration inequalities (Martingales, Transportation, Exchangeable pairs, ...) usually attempt to compare tails for smooth functionals with Gaussian or Poissonian tails.

But ...

Gaussian and Poisson random variables are not the only possible limits.

Variations of the entropy method

may be able to capture such behaviors ...

Part II

Order statistics

A simple example: order statistics

Order statistics (empirical quantiles) provide examples of simple random variables that enjoy non-trivial concentration properties

Order statistics have been used and studied intensively in different branches of statistics: robust statistics, extreme value theory, ...

Order statistics provide a playground for the entropy method.

Notation

Order statistics

Sample :

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

$$X_{1,n} \geq \ldots \geq X_{n,n} \quad \text{non-increasing rearrengement of} \quad X_1, \ldots, X_n$$

If n clear from context,

$$X_{1,n}, \ldots, X_{n,n} \text{ denoted by } X_{(1)}, \ldots, X_{(n)}$$

Examples

$X_{(1)}$ extreme $\qquad X_{(k_n)}, k_n \nearrow \infty, \frac{k_n}{n} \searrow 0$ (intermediate) $\qquad X_{(n/2)}$ central

Goal

simple, non-asymptotic variance/tail bounds

Off-the shelf concentration inequalities and order statistics

$f(X_1, \ldots, X_n) = X_{(i)}$

An order statistics is a simple function of many independent random variables that does not depend *too much* on any of them.

Gaussian order statistics

Almost surely, $\|\nabla f\| = 1$.
Poincaré's inequality $\Rightarrow$ $\mathrm{Var}(f(X_1, \ldots, X_n)) \leq 1$ But:

$$\mathrm{Var}(X_{(1)}) = O(1/\log n)$$
$$\mathrm{Var}(X_{(n/2)}) = O(1/n)$$

We do not understand (clearly)

in which way the maximum is a smooth function of the sample.

Part III

Variance bounds

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*

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

Lemma (Jackknife bounds)

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

Jackknife estimate of variance

EFRON-STEIN-STEELE inequalities (1981)

$Z = f(X_1, \ldots, X_n)$ ... a function of independent random variables

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

where $\mathrm{Var}^{(i)}(Z)$ is the variance of Z conditionally on $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$

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

... may be chosen as any measurable function of $X_1, \ldots, X_{i-1}, , X_{i+1}, \ldots, X_n$

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

... $\sum_{i=1}^{n} (Z - Z_i)^2$ is a jackknife (leave one out) estimate of variance

Proof : application of Efron-Stein-Steele inequality

- $Z = X_{(k)}$
- Z_i as the rank k statistic from subsample $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$:

$$Z_i = \begin{cases} Z_i = X_{(k+1)} & \text{if } X_i \geq X_{(k)} \\ Z_i = Z & \text{otherwise.} \end{cases}$$

- Jackknife estimate of variance of $X_{(k)}$:

$$\sum_{i=1}^{n} (Z - Z_i)^2 = \sum_{i: X_i \geq X_{(k)}} (X_{(k)} - X_{(k+1)})^2 = k\Delta_k^2$$

□

How tight is $\mathrm{Var}(X_{(k)}) \leq k\mathbb{E}\Delta_k^2$?

Partial assessment of the tightness of the variance upper bound may be performed without heavy computations by resorting to asymptotic comparisons

For central order statistics (the median), if the density at the median is not null, the expected value of the squared spacing is $O(1/n^2)$.

For extreme order statistics, Extreme Value Theory provides a framework for benchmarking

Asymptotic assessment for extreme order statistics

Maximum Domain of Attraction $\mathrm{MDA}(\gamma), \gamma \in \mathbb{R}$

$F \in \mathrm{MDA}(\gamma)$ if there exists a function $a : \mathbb{R}_+ \to \mathbb{R}_+$, such that

$$\mathbb{P}\left\{\frac{X_{1,n} - U(n)}{a(n)} \leq x\right\} \to \exp\left(-(1+\gamma x)^{-1/\gamma}\right)$$

according to the sign of extreme value index $\gamma \begin{cases} > 0 & \text{Frechet domain} \\ = 0 & \text{Gumbel domain} \\ < 0 & \text{Weibull domain} \end{cases}$

Asymptotic assessment for extreme order statistics (ii)

If $F \in \mathrm{MDA}(\gamma)$ with $\gamma < 1/2$,

the ratio between the jackknife estimate and the variance converges toward a limit that depends on k and γ, for $k = 1$:

$$\lim_{n\to\infty} \frac{\mathbb{E}\left[\left(X_{(1)} - X_{(2)}\right)^2\right]}{\mathrm{Var}[X_{(1)}]} = \frac{\frac{2\Gamma(2(1-\gamma))}{(1-\gamma)(1-2\gamma)}}{\frac{\Gamma(1-2\gamma)-\Gamma(1-\gamma)^2}{\gamma^2}}$$

In the Guembel domain ($\gamma = 0$),

for $k = 1$, the limit is $12/\pi^2 \approx 1.2159$.

Explicit variance bounds and beyond

Variance bounds are to be complemented by bounds on the logarithmic moment generating function in order to derive exponentiel tail bounds (Chernoff-bounding)

$X_{(1)}$ is exponentially integrable only if X_1 is.
We also need a handy way to bound moments of spacings

Rényi's representation and apropriate assumption on the hazard function of the distribution of X_i do the job

Part IV

Rényi's representation

Rényi's representation

The order statistics of an exponential sample ...

are partial sums of independent exponentially distributed random variables.

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

$$X_{k,n} = \sum_{i=k}^{n} \Delta_i$$

where

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

ii) spacings are rescaled exponentials, $i \times \Delta_i \sim 1 - \mathrm{e}^{-x}$

Quantile transformation

Quantile transformation

- If V is uniformly distributed over $[0, 1]$, then $F^{\leftarrow}(V)$ is distributed according to F.
- If E is exponentially distributed, then $U(\mathrm{e}^{E}) = F^{\leftarrow}(1 - \exp(-E))$ is distributed according to F.

This observation has found uncountably many applications in random simulation, statistics, coupling, etc.
When combined with the fact that order statistics of an exponential sample are partial sums of independent random variables, it leads to a very convenient distributional representation for any sample of order statistics.

Distributional 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

$-\log \overline{F}$ is the hazard function associated to the distribution function F
If F is differentiable, the hazard rate is defined as the derivative of the hazard function $F'/\overline{F}$.

The exponential distribution has constant hazard rate (memoryless property).
Distributions with non-decreasing hazard rate have lighter right-tails than the exponential distribution.

The function $U \circ \exp$ is the generalized inverse of the hazard function

$$U \circ \exp = (-\log \overline{F})^{\leftarrow}$$

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

Negative association

Negative association

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.
For increasing functions f, g

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

Taking advantage of increasing hazard rate

If F has non-decreasing hazard rate h,

The variance of the k^{th} order statistics is simply related to the hazard rate.
For $1 \leq k \leq n/2$,

$$\mathrm{Var}\left[X_{(k)}\right] \leq \mathbb{E} V_k \leq \frac{2}{k}\mathbb{E}\left[\left(\frac{1}{h(X_{(k+1)})}\right)^2\right] ,$$

Some more calculus leads to:

for $n \geq 3$, for $1 \leq k \leq n/2$,

$$\mathrm{Var}[X_{(k)}] \leq \frac{1}{k \log 2} \frac{8}{\log \frac{2n}{k} - \log(1 + \frac{4}{k} \log\log \frac{2n}{k})}$$

where $X_{(k)}$ is an order statistic of a sample of absolute values of Gaussians.

Part V

Exponential Efron-Stein inequalities

Outline

Goal

Beyond variance

Sticking to Efron-Stein inequalities, relying on arguments geared toward order statistics, allows to go beyond variance bounds

Context

If F has increasing hazard rate (more concentrated than exponential), extreme and intermediate order statistics have exponentiel moments.

Log-concavity of F

implies non-decreasing hazard rate.
It also implies log-concavity of the joint distribution of order statistics (and by Borell's argument, sub-exponential tails).

Next

- Exponential Efron-Stein inequalities and Bernstein-like exponential inequalities
- Using the entropy method

Bernstein bounds, sub-Gamma distributions

What we are looking for ?

- Maxima of independent Gaussians are asymptotically Gumbel (sub-exponential on the right tail)
- Central and intermediate order statistics are asymptotically Gaussian (Smirnov)

We expect sub-Gamma behavior (on the right-tail)

Sub-gamma on the right tail with variance factor v and scale parameter c

$$\log \mathbb{E} e^{\lambda(X - \mathbb{E}X)} \leq \frac{\lambda^2 v}{2(1 - c\lambda)} \text{ for every } \lambda \quad \text{such that} \quad 0 < \lambda < 1/c\,.$$

Bernstein's inequality

$$\text{for } t > 0, \mathbb{P}\left\{X \geq \mathbb{E}X + \sqrt{2vt} + ct\right\} \leq \exp\left(-t\right).$$

Entropy method

Ledoux's entropy method

has been inspired by derivations of Gaussian concentration inequalities starting from Gross logarithmic Sobolev inequality

Applications

- ▷ Suprema of bounded empirical processes (Talagrand,...,Bousquet)
- ▷ Self-bounded functions (configuration functions, VC-entropy, conditional Rademacher averages...)

Revisiting the proof of Hoeffding inequality

By independence

$$\log \mathbb{E}\mathrm{e}^{\lambda Z} \log \mathbb{E}\mathrm{e}^{\lambda \sum_i (X_i - \mathbb{E}X_i)} = \sum_i \log \mathbb{E}\mathrm{e}^{\lambda (X_i - \mathbb{E}X_i)}$$

For each i,

$$\frac{\mathrm{d}^2 \log \mathbb{E}\mathrm{e}^{\lambda (X_i - \mathbb{E}X_i)}}{\mathrm{d}\lambda^2} = \frac{\mathbb{E}\left[X_i^2 \mathrm{e}^{\lambda (X_i - \mathbb{E}X_i)}\right]}{\mathbb{E}\mathrm{e}^{\lambda (X_i - \mathbb{E}X_i)}} - \left(\frac{\mathbb{E}\left[X_i \mathrm{e}^{\lambda (X_i - \mathbb{E}X_i)}\right]}{\mathbb{E}\mathrm{e}^{\lambda (X_i - \mathbb{E}X_i)}}\right)^2$$

The variance of a random variable with support in $[a_i, b_i]$ is not larger than $(b_i - a_i)^2/4$

$$\frac{\mathrm{d}^2 \log \mathbb{E}\mathrm{e}^{\lambda Z}}{\mathrm{d}\lambda^2} \leq \sum_i \frac{(b_i - a_i)^2}{4}$$

Integration of the differential inequality leads to Hoeffding inequality

The entropy method

For more general functions of $X_1, \ldots, X_n$

the logarithmic moment generating function is not usually a sum

But ...

$$\frac{\mathrm{d} \frac{1}{\lambda} \log \mathbb{E} \mathrm{e}^{\lambda Z}}{\mathrm{d}\lambda} = \frac{\mathbb{E}\left[\lambda Z \mathrm{e}^{\lambda Z}\right] - \mathbb{E}\mathrm{e}^{\lambda Z} \log \mathbb{E}\mathrm{e}^{\lambda Z}}{\mathbb{E}\mathrm{e}^{\lambda Z}} =: \frac{\mathrm{Ent}\left[\mathrm{e}^{\lambda Z}\right]}{\mathbb{E}\mathrm{e}^{\lambda Z}}$$

Subadditivity property of entropy (just like for the variance)

$$\mathrm{Ent}\left[\mathrm{e}^{\lambda Z}\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Ent}^{(i)}\left[\mathrm{e}^{\lambda Z}\right]\right]$$

The *entropy method* takes advantage of this subadditivity to derive differential inequalities for logarithmic moment generating functions of functions of many independent random variables

Modified logarithmic Sobolev inequalities

As usual

Z is a function of n independent random variables $X_1, \ldots, X_n$
For $i \leq n$, Z_i is a function of $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$

Modified logarithmic Sobolev inequality (L. Wu, P. Massart, 2000)

$$\begin{aligned} \text{Ent}\left[e^{\lambda Z}\right] &\leq \sum_{i=1}^{n} \mathbb{E}\left[\text{Ent}^{(i)}\left[e^{\lambda Z}\right]\right] \\ &\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z}\tau\left(-\lambda(Z - Z_i)\right)\right] \quad \text{for } \lambda \in \mathbb{R} \end{aligned}$$

where $\tau(x) = e^x - x - 1$

This inequality holds in any product space, it is the starting point of the derivation of the tail bounds for self-bounding functions, for suprema of empirical processes (Talagrand 1996, Ledoux 1996, Massart 2000, Rio 2001, Bousquet 2002)

Application to order statistics

Notation

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

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

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

Proof parallels the variance bounds derived from Efron-Stein inequalities.

Exponential Efron-Stein inequality for order statistics

Bernstein inequality for order statistics, (B. and Thomas, 2012)

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

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

where $V_k = k(X_{(k)} - X_{(k+1)})^2$ is the Efreon-Stein estimate of variance for $X_{(k)}$.

Assessment

- Does not follow from previous exponential Efron-Stein inequality

$$\log \mathbb{E} e^{\lambda(X_{(k)} - \mathbb{E} X_{(k)})} \quad \leq \quad \frac{\lambda\theta}{1-\lambda\theta} \log \mathbb{E} e^{\lambda V_k/\theta}$$

for $\theta > 0, 0 \leq \lambda \leq 1/\theta$

(B., Lugosi and Massart. Ann. Probab. 2003)

- V_k may not have exponential moments while $\sqrt{V_k}$ has!
- Going beyond B., Lugosi and Massart (2003) critically depends on taking advantage of negative association rather than on

$$\mathbb{E}\left[W e^{\lambda Z}\right] \leq \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{W}\right] + \text{Ent}(e^{\lambda Z})$$

- Sharp (up to constants) for exponential samples.
- Works both for central, intermediate and extreme order statistics.

Proof (i)

- $\psi(x) = x(e^x - 1)$ is non-decreasing over $\mathbb{R}_+$,
- $X_{(k+1)}$ and Δ_k are negatively associated:

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

- Multiplying both sides by $\exp(-\lambda\mathbb{E}X_{(k)})$, leads to

$$\mathrm{Ent}\left[e^{\lambda(X_{(k)}-\mathbb{E}X_{(k)})}\right] \leq k\mathbb{E}\left[e^{\lambda(X_{(k)}-\mathbb{E}X_{(k)})}\right] \times \mathbb{E}\left[\psi(\lambda\Delta_k)\right] .$$

Proof (ii) Herbst's argument

Let $G(\lambda) = \mathbb{E}e^{\lambda\Delta_k}$.
Obviously, $G(0) = 1$, and as $\Delta_k \geq 0$, G and its derivatives are increasing on $[0, \infty)$,

$$\mathbb{E}\left[\psi(\lambda\Delta_k)\right] = 1 - G(\lambda) + \lambda G'(\lambda) = \int_0^\lambda sG''(s)\mathrm{d}s \leq G''(\lambda)\frac{\lambda^2}{2} .$$

Hence, for $\lambda \geq 0$,

$$\frac{\text{Ent}\left[e^{\lambda(X_{(k)}-\mathbb{E}X_{(k)})}\right]}{\lambda^2\mathbb{E}\left[e^{\lambda(X_{(k)}-\mathbb{E}X_{(k)})}\right]} = \frac{\mathrm{d}\frac{1}{\lambda}\log\mathbb{E}e^{\lambda(X_{(k)}-\mathbb{E}X_{(k)})}}{\mathrm{d}\lambda} \leq \frac{k}{2}\frac{\mathrm{d}G'}{\mathrm{d}\lambda} .$$

Proof (iii) solving the differential inequality

Integrating both sides, using the fact that

$$\lim_{\lambda \to 0} \frac{1}{\lambda} \log \mathbb{E} e^{\lambda (X_{(k)} - \mathbb{E} X_{(k)})} = 0,$$

leads to

$$\begin{aligned}\frac{1}{\lambda} \log \mathbb{E} e^{\lambda (X_{(k)} - \mathbb{E} X_{(k)})} &\leq \frac{k}{2} (G'(\lambda) - G'(0)) \\ &= \frac{k}{2} \mathbb{E} \left[\Delta_k \left(e^{\lambda \Delta_k} - 1 \right) \right] .\end{aligned}$$

□

Maxima of Gaussians

For n such that the solution v_n of equation

$$16/x + \log(1 + 2/x + 4\log(4/x)) = \log(2n)$$

is smaller than 1,
for all $0 \leq \lambda < \frac{1}{\sqrt{v_n}}$,

$$\log \mathbb{E} e^{\lambda(X_{(1)} - \mathbb{E}X_{(1)})} \leq \frac{v_n \lambda^2}{2(1 - \sqrt{v_n}\lambda)} \ .$$

For all $t > 0$,

$$\mathbb{P}\left\{X_{(1)} - \mathbb{E}X_{(1)} > \sqrt{v_n}(t + \sqrt{2t})\right\} \leq e^{-t} \ .$$

Median of Gaussians

...

The same approach works for extreme, intermediate and central order statistics

Let $v_n = 8/(n \log 2)$.
For all $0 \leq \lambda < n/(2\sqrt{v_n})$,

$$\log \mathbb{E} e^{\lambda (X_{(n/2)} - \mathbb{E} X_{(n/2)})} \leq \frac{v_n \lambda^2}{2(1 - 2\lambda\sqrt{v_n/n})} \ .$$

For all $t > 0$,

$$\mathbb{P}\left\{ X_{(n/2)} - \mathbb{E} X_{(n/2)} > \sqrt{2 v_n t} + 2\sqrt{v_n/n} t \right\} \leq e^{-t} \ .$$

Part VI

Assessment

Assessment (i)

Rényi's representation

Order statistics are functions of sums of independent random variables (spacings of exponential samples).

If the function $U \circ \exp$ is concave, concavity may be used in several ways.

What about plugging tail bounds for order statistics of exponential samples ?

Ad hoc tail bounds

What can be obtained from Rényi's representation and exponential inequalities for sums of Gamma-distributed random variables ?

Lemma

Let $X_{(1)}$ be the maximum of the absolute values of n independent standard Gaussian random variables, and let $\widetilde{U}(s) = \Phi^{\leftarrow}(1 - 1/(2s))$ for $s \geq 1$. For $t > 0$,

$$\mathbb{P}\left\{X_{(1)} - \mathbb{E}X_{(1)} \geq t/(3\widetilde{U}(n)) + \sqrt{t}/\widetilde{U}(n) + \delta_n\right\} \leq \exp\left(-t\right) ,$$

where $\delta_n > 0$ and $\lim_n(\widetilde{U}(n))^3\delta_n = \frac{\pi^2}{12}$.

This is a deviation inequality, not a concentration inequality. Assessing its quality requires a good understanding of the second-order regular variation property the Gaussian quantile function.

Alternative approach: revisiting smoothness

A refinement of the Poincaré inequality may be used to prove tight bounds for variance of maxima of Gaussian vectors

$L_1 - L_2$ method (Talagrand-...-Chatterjee)

$$\mathrm{Var}(f) \leq C \sum_{i=1}^{n} \frac{\mathbb{E}|\partial_i f|^2}{1 + \log \frac{(\mathbb{E}|\partial_i f|^2)^{1/2}}{\mathbb{E}|\partial_i f|}}$$

C is a universal constant related to the Poincaré and logarithmic Sobolev constants

The $L_1 - L_2$ approach provides a simple derivation of a tight variance bound for the maximum of a standard Gaussian vector

$$\mathrm{Var}\left(\max(X_1, \ldots, X_n)\right) \leq \frac{C}{1 + \log n}$$

The $L_1 - L_2$ approach

Applications

- First and last passage percolation (Benjamini-Kalai-Schramm, Benaim-Rossignol, Graham, Chatterjee)
- Criterion for super-concentration of monotone functions (Chatterjee)

$$\text{Is } \frac{\sum_i (\mathbb{E}|\partial_i f|)^2}{\sum_i (\mathbb{E}|\partial_i f|_2)^2} \text{ small ?}$$

- Harmonic analysis of Boolean functions
- Local concentration Devroye-Lugosi

Relies on

hyper-contractivity of a Markov semi-group whose stationary distribution should be the sampling distribution.

Poincaré's inequality for products of exponentials

If $Z = f(E_1, \ldots, E_n)$ is a differentiable function of independent exponential R.V.

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

The constant 4 can not be improved

Proof

Combine Efron-Stein and Cauchy-Schwarz inequalities :

$$\begin{aligned}\int_0^\infty \mathrm{e}^{-x}(f(x) - f(0))^2 \mathrm{d}x &= 2\int_0^x \mathrm{e}^{-x} f'(x)(f(x) - f(0))\mathrm{d}x \\ &\leq 2\left(\int_0^x \mathrm{e}^{-x}(f'(x))^2 \mathrm{d}x\right)^{1/2} \left(\int_0^x \mathrm{e}^{-x}(f(x) - f(0))^2 \mathrm{d}x\right)^{1/2}\end{aligned}$$

Exponential Poincaré inequality and Rényi's representation

Assuming $F^{\leftarrow}$ is differentiable

$$\operatorname{Var}(X_{(k)}) \leq 4 \sum_{i=k}^{n} \frac{1}{i^2} \mathbb{E}\left[\frac{1}{h(X_{(k)})^2}\right]$$

This upper-bound is valid whatever the shape of the hazard rate function. Under the non-decreasing hazard rate assumption, the simple Efron-Stein inequality allows to improve the constant 4 to 2.

Proof

$$X_{(k)} \sim U \circ \exp\left(\sum_{i=k}^{n} \frac{E_i}{i}\right) =: f(E_1, \ldots, E_n)$$

$$\partial_i f = \frac{1}{i} \frac{1}{h\left(f(E_1, \ldots, E_n)\right)} \quad \text{for} \quad i \geq k$$

Talagrand's inequality for products of exponentials

Talagrand (1991), Maurey (1991), Bobkov and Ledoux (1997) showed that smooth functions of independent exponential random variables satisfy concentration inequalities of Bernstein type.

The next result is extracted from the derivation of Talagrand's concentration phenomenon for product of exponentials (Bobkov, Ledoux, PTRF 1997).

If $Z = f(E_1, \ldots, E_n)$ is a differentiable function of independent exponential R.V.

If $\max_i |\partial_i f| \leq c'$, for all $0 \leq \lambda < c < c'$,

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

Consequences ...

If $f(E_1, \ldots, E_n) = U \circ \exp\left(\sum_{i=k}^n E_i/i\right)$, dans $U \circ \exp$ is C_1 and concave,

$$|\partial_i f| \leq \frac{1}{i} \sup_x \frac{1}{h(x)} \text{ for } i \geq k$$

$$\|\nabla f\|^2 = \sum_{i=k}^n \frac{1}{i^2} \frac{1}{(h \circ f)^2}$$

The function $1/(h(z))^2$ is a non-increasing function of z, by negative association,

$$\frac{\text{Ent}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}Z)}\right]}{\mathbb{E}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}Z)}\right]} \leq \frac{\lambda^2}{2(1-c)} 4\mathbb{E}\left[\|\nabla f\|^2\right]$$

This implies that $X_{(k)}$ is sub-Gamma with variance factor $4\mathbb{E}\left[1/(h(Z))^2\right]$ and scale factor larger than $1/\inf_x h(x)$.

Comparison

Combining the Talagrand-Bobkov-Ledoux inequality and Rényi's representation

leads to another Bernstein-type inequality for order statistics when the sampling distribution has non-decreasing hazard rate.

For non-decreasing hazard rate

the Poincaré estimate of variance is an upper bound on the Efron-Stein estimate of variance.

Scale factors are of the same order of magnitude

Thanks to the change of representation, off the shelf arguments provide sharp bounds for fluctuations of order statistics: order statistics are genuine smooth functions of exponential spacings.

Further readings

S.B., G. Lugosi, and P. Massart.
Concentration Inequalities.
Oxford University Press. Feb. 2013.

S. B. and M. Thomas.
Concentration inequalities for order statistics.
Electronic Communications in Probability. 17 (2012). 1-12
http://arxiv.org/abs/1207.7209

S. Chatterjee.
Superconcentration and Related Topics.
Springer, 2014.

M. Ledoux.
The concentration of measure phenomenon
(Vol. 89). American Mathematical Soc. 2001.

M. Talagrand.
New concentration inequalities in product spaces.
Inventiones Mathematicae, 126:505–563, 1996.