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

Concentration, super-concentration, ...

S. Boucheron[1]

[1] LPMA CNRS & Université Paris-Diderot

Conference on Numerical Analysis and Scientific Computing, Lepizig, January 2014

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

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

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

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

Gaussian setting

Cirelson inequality

$X_1, \ldots, X_n$ standard Gaussian vector, $Z = f(X_1, \ldots, f(X_n)$, f L-Lipschitz

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

Gaussian concentration

may be characterized by functional inequalities
$X = (X_1, \ldots, X_n)$ a standard Gaussian vector

$$\begin{aligned}
\text{Poincaré} \quad & \operatorname{Var} f(X) \leq \mathbb{E}\|\nabla f\|^2 \\
\text{logarithmic Sobolev} \quad & \operatorname{Ent}(f(X)^2) \leq 2\mathbb{E}\|\nabla f\|^2 \\
\text{modified logarithmic Sobolev} \quad & \operatorname{Ent}(f(X)) \leq 2\mathbb{E}\frac{\|\nabla f\|^2}{f}
\end{aligned}$$

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

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

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

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

Examples

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

B., Lugosi and Massart, 2000-3: sub-Poisson concentration

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

Smoothness may not be enough

Off the shelf inequalities

may fail to capture some aspects of the concentration phenomenon.

Longest increasing subesequence

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

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

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

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

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 ...

i) Order statistics
ii) Empirical excess risk

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 :
$\mathrm{Var}(\max(X_1, \ldots, X_n)) = O(1/\log n)$
$\mathrm{Var}(X_{(n/2)}) = O(1/n)$

We do not understand (clearly)

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

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)

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

Proof (i)

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 (ii) : 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 $X_{(k)}$:

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

□

Asymptotic assessment for extreme order statistics

Quantile function

$$F^{\leftarrow}(p) = \inf\{x \colon F(x) \geq p\} \quad U(t) = F^{\leftarrow}\left(1 - \frac{1}{t}\right)$$

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

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

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

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 $k = 1$:

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

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

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

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

Hazard rate of a differentiable distribution function F

$-\log \overline{F}$ is the hazard function associated to F

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

$F'/\overline{F}$ is the associated hazard rate
The distribution function F has non-decreasing hazard rate, iff $U \circ \exp$ is concave

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 $1 \leq k \leq n/2$,

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

Some more calculus leads to:

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

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

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

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.

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.

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

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

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)

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

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

Holds in any product space

Application to order statistics

Notation

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

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

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

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

Exponential Efron-Stein inequality for order statistics

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

B. and Thomas (2012)

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

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

Assessment

- Does not follow from previous exponential Efron-Stein inequality

$$\log \mathbb{E} e^{\lambda(X_{(k)} - \mathbb{E} X_{(k)})} \quad \leq \quad \frac{\lambda\theta}{1-\lambda\theta} \log \mathbb{E} e^{\lambda V_k/\theta}$$
$$\text{for } \theta > 0, 0 \leq \lambda \leq 1/\theta$$

  (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

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

- 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 $\lambda \geq 0$,

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

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 $0 \leq \lambda < \frac{1}{\sqrt{v_n}}$,

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

For all $t > 0$,

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

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 $t > 0$,

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

Application(s) to statistical learning

Statistical learning theory, as initiated by Vapnik and Chervonenkis, served as a playground for empirical process theory.

During the 1990's and 2000's, the availability of sharp concentration inequalities for suprema of empirical processes and so-called empirical complexities simplified and sometimes made possible the derivation of sharp performance bounds

Caveat

Statistical learning theory started before concentration inequalities became available. Vapnik-Chervonenkis inequalities (deviation inequalities) were sufficient to achieve a lot of results. As of today, concentration inequalities are not general enough to deal with many problems.

Statistical Learning setting (i)

- $\overbrace{\mathcal{X}}^{\text{examples}} \times \overbrace{\mathcal{Y}}^{\text{labels}}$ endowed with unknown P,
- Sample of labelled examples $(X_i, Y_i)_{i \leq n}$ picked independently according to some unknown probability distribution P on $\mathcal{X} \times \mathcal{Y}$
- Binary classification: $\mathcal{Y} = \{-1, 1\}$
  Bounded regression: $\mathcal{Y} = [-b, b]$
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$
  - Hard loss: $\ell(f(X), Y) = \mathbf{1}_{f(X) \neq Y}$
  - Hinge loss (convex): $\ell(f(X), Y) = (1 - f(X)Y)_+$
  - ...
- Risk of $f \in \mathcal{Y}^{\mathcal{X}}$

$$R(f) = P\ell(f(X), Y) = \mathbb{E}_P \ell(f(X), Y)$$

Statistical learning setting (ii)

- Assumption/notation: f^* minimizes $R(f) \in \mathcal{Y}^{\mathcal{X}}$
- Example: *Bayes classifier* in binary classification

$$f^*(x) = 2\mathbf{1}_{\mathbb{E}[Y|X]>0} - 1 = \mathrm{sign}(\mathbb{E}[\mathrm{Y} \mid \mathrm{X}])$$

- Goal: given a model $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$
  find $\bar{f} \in \mathcal{F}$ that minimizes risk $R(.)$ over $\mathcal{F}$
- Recipes: minimize empirical risk

$$R_n(f) = P_n \ell(f(X), Y) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

- Assumption/notation: $\hat{f}$ minimizes empirical risk over $\mathcal{F}$

Statistical learning setting(iii)

- Model bias $L(\bar{f}) = R(\bar{f}) - R(f^*)$
- Excess risk $R(\hat{f}) - R(\bar{f})$
- *Excess empirical risk* $R_n(\bar{f}) - R_n(\hat{f})$
- Notation: $\bar{R}_n(f) = R_n(f) - R(f)$

Fundamental relation

$$\bar{R}_n(\bar{f}) - \bar{R}_n(\hat{f}) = \overbrace{R(\hat{f}) - R(\bar{f})}^{\text{Excess risk}} + \overbrace{R_n(\bar{f}) - R_n(\hat{f})}^{\text{Excess empirical risk}}$$

Control of excess risk/empirical excess risk

In order to bound excess risk and/or empirical excess risk, it is enough to get bounds on increments of the centered empirical process .

Control of increments of centered empirical process

- If random function ϕ_n satisfies

$$\forall f \in \mathcal{F} \quad \left|\bar{R}_n(f) - \bar{R}_n(\bar{f})\right| \leq \phi_n(R(f))$$

- looking for largest value r of $R(f)$ that satisfies

$$R(f) - R(\bar{f}) \leq \phi_n(R(f))$$

- leads to

$$\max(R_n(\bar{f}) - R_n(\widehat{f}), R(\widehat{f}) - R(\bar{f})) \quad \leq \quad r$$

Controlling modulus of continuity of $\bar{R}_n(\cdot) - \bar{R}_n(\bar{f})$

- Loss class $\mathcal{H} = \{\ell(f(\cdot), \cdot), f \in \mathcal{F}\}$
  - $h(X, Y) = \ell(f(X), Y)$
  - $h^*(X, Y) = \ell(f^*(X), Y)$
  - $\bar{h}(X, Y) = \ell(\bar{f}(X), Y)$
- Complexity of the L_2 neighborhood of $\bar{h}$ in $\mathcal{H}$:

$$\sqrt{n}\mathbb{E}\left[\sup_{h \in \mathcal{F}, P(h-\bar{h})^2 \leq r^2} \left|(P_n - P)(h - \bar{h})\right|\right] \leq \psi(r)$$

- Noise conditions

$$\sup\left\{\left(P(h - h^*)^2\right)^{1/2} : P(h - h^*) \leq r^2\right\} \leq \omega(r)\,.$$

- Assumptions: $\psi, \omega \nearrow$, continuous ≥ 0, $\psi(x)/x, \omega(x)/x \searrow$ and $\psi(1), \omega(1) \geq 1$

Talagrand's inequality (Talagrand, 1996,..., Bousquet 2002)

Bennett's inequality for suprema of bounded centered empirical processes

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

$$\begin{aligned} \sigma^2 &= \sup_{h \in \mathcal{H}} \mathrm{Var}[h(X))] \\ b &= \sup_{h \in \mathcal{H}} \|h(X) - \mathbb{E}[h(X)]\|_\infty \\ Z &= \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} (h(X_i) - \mathbb{E}[h(X)]) = n \sup_{h \in \mathcal{H}} (P_n - P)h \end{aligned}$$

Let $v = 2b\mathbb{E}[Z] + n\sigma^2$.

$$\forall \delta > 0, \quad \mathbb{P}\left\{ Z \geqslant \mathbb{E}[Z] + \sqrt{2v \log \frac{1}{\delta}} + \frac{b}{3} \log \frac{1}{\delta} \right\} \leqslant \delta.$$

Deviation for excess risk

Excess risk and Excess empirical risk satisfy deviation inequalities

Deviation inequalities provide information on tail but may fail to describe typical fluctuations around mean or median.
Investigating separately expectations and concentration has proved to clarify things

- $(P, \ell, \mathcal{F})$ learning task
- r_*, solution of

$$\sqrt{n}r^2 = \psi(2\omega(r)) .$$

- $\exists \kappa_1, \kappa_2, \kappa_3 \geq 1$, w.p. $\geq 1 - 2\delta$:

$$\begin{aligned}
\max\left(R(\widehat{f}) - R(\bar{f}), R_n(\bar{f}) - R_n(\widehat{f})\right) &\leq \kappa_1 L(\bar{f}) + \kappa_2 r_*^2 + \kappa_3 r_*^2 \log\frac{1}{\delta} \\
\max\left(\mathbb{E}[R(\widehat{f}) - R(\bar{f})], \mathbb{E}[R_n(\bar{f}) - R_n(\widehat{f})]\right) &\leq \kappa_1 L(\bar{f}) + (\kappa_2 + \kappa_3) r_*^2
\end{aligned}$$

Benchmark: VC classes under gentle noise

- Examples: half-spaces in $\mathbb{R}^d$
  $\hookrightarrow$ vc-dimension: $V = d + 1$
- Classification.
  Hard loss. $\ell(y, y') = \mathbf{1}_{y \neq y'}$
- Random classification noise $|\mathbb{E}[Y \mid X]| = \beta$
  $Y = \mathrm{sign}(\eta(X))$ with probability β
- $\omega(r) = \frac{r}{\sqrt{\beta}}$
  $\hookrightarrow$ If $P(h - h^*) \leq r^2$ then $P(h - h^*)^2 \leq \frac{r^2}{\beta}$

$$\psi(r) = Cr\sqrt{V(1 + \log(1 \vee r^{-1}))}$$

Goal: concentration inequalities for Empirical Excess Risk

Empirical Excess Risk is a supremum of empirical process!

$$Z = nP_n(\bar{h} - \hat{h}) = n \sup_{h \in \mathcal{H}} P_n(\bar{h} - h)$$

This empirical process

has non-centered components

$$\forall h \in \mathcal{H}, \quad \mathbb{E}\left[nP_n(\bar{h} - h)\right] \leq 0$$

But it is non-negative

$$\mathbb{E}\left[n \sup_{h \in \mathcal{H}} P_n(\bar{h} - h)\right] \geq 0$$

Excess Empirical Risk does not fit in the framework of suprema of bounded centered empirical processes handled using Talagrand's inequality.

Variance bounds for empirical excess risk

- Let $\hat{h}_n$ minimize $P_n h$
- The variance of the empirical excess risk is intimately related to the L_2 distances between $\hat{h}_n$ and $\bar{h}$

$$
\begin{aligned}
&\mathrm{Var}\left[nP_n(\bar{h}-\widehat{h}_n)\right] \\
&\leq \quad 2n\left(\mathbb{E}\left[P_n(\bar{h}-\widehat{h}_n)^2\right]+\mathbb{E}\left[P(\bar{h}-\widehat{h}_n)^2\right]\right)
\end{aligned}
$$

- It can also be related with the increment of a jaccknifed empirical process between $\bar{h}$ and $\widehat{h}_n$

$$
\begin{aligned}
&\mathrm{Var}\left[nP_n(\bar{h}-\widehat{h}_n)\right] \\
&\leq \quad 2n\mathbb{E}\left[\left((P_{n-1}-P)(\bar{h}-\widehat{h}_{n-1})\right)\right]+2n\mathbb{E}\left[P\left(\bar{h}-\widehat{h}_n\right)^2\right]
\end{aligned}
$$

Combining with risk bounds and properties of ω

- $(P, \ell, \mathcal{F})$: learning task.
- $\psi, \omega \in \mathcal{C}_1$ complexity and the noise
- Let r_* denote the positive solution of

$$\sqrt{n} r^2 = \psi(2\omega(r)) \, .$$

$\exists \kappa_4$ such that

$$\mathrm{Var}\left[n(R_n(\bar{f}) - R_n(\widehat{f}))\right] \leq n\kappa_4 \left(\omega^2\left(r_*\right) + \omega^2\left(\sqrt{L(\bar{f})}\right)\right)$$

VC Classes

- VC classes under random classification noise ($L(\bar{f}) = 0$)
- $\hookrightarrow r_*^2 \leq C^2 \left(\left(\frac{V(1+\log(n\beta^2/V))}{n\beta} \right) \wedge \sqrt{\frac{V}{n}} \right)$
- $\omega^2(r_*) \leq C^2 \left(\left(\frac{V(1+\log(n\beta^2/V))}{n\beta^2} \right) \wedge \sqrt{\frac{V}{n\beta^2}} \right)$
- $$\begin{aligned} & \mathbb{E}\left[n(R_n(\bar{f}) - R_n(\widehat{f})) \right] \\ & \quad \leq \quad (\kappa_2 + \kappa_3) \left(C^2 \left(\left(\frac{V(1+\log(n\beta^2/V))}{\beta} \right) \wedge \sqrt{nV} \right) \right) \\ & \mathrm{Var}\left[n(R_n(\bar{f}) - R_n(\widehat{f})) \right] \\ & \quad \leq \quad \kappa_4 C^2 \left(\left(\frac{V(1+\log(n\beta^2/V))}{\beta^2} \right) \wedge \sqrt{\frac{nV}{\beta^2}} \right) \end{aligned}$$

Another look at Bernstein inequalities

- Z satisfies a Bernstein inequality with paramters V and c

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

- Recentered $\Gamma(p, c)$ random variable satisfy Berstein inequalities
- If

$$\|Z - \mathbb{E}Z\|_q \leq \sqrt{vq} + cq$$

for $q \geq 2$ then Z satisfies a Bernstein inequality.

General moment bounds

- $Z = F(X_1, \ldots, X_n)$ with $(X_1, \ldots, X_n)$ independent random variables
- $X_1', \ldots, X_n'$, independent copies of $X_1, \ldots, X_n$ and $Z_i' = F(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$.
- $V_+ = \sum_{i=1}^n \mathbb{E}' \left[(Z - Z_i')_+^2 \right]$.

B., Bousquet, Lugosi & Massart, AoP, 2005

For any $q \geq 2$:

$$\|(Z - \mathbb{E}[Z])_+\|_q \leq \sqrt{3q \, \|V_+\|_{q/2}} = \sqrt{3q} \left\| \sqrt{V_+} \right\|_q .$$

Assuming $\exists M$ r.v. with $(Z - Z_i')_+ \leq M \; \forall i \leq n$, for all $q \geq 2$

$$\|(Z - \mathbb{E}[Z])_-\|_q \leq \sqrt{5q} \left(\left\| \sqrt{V_+} \right\|_q \vee \|M\|_q \right) .$$

Main statement

A Bernstein-like inequality for excess empirical risk

Let $Z = nP_n\left(\bar{h} - \widehat{h}_n\right)$. For $q \geqslant 2$.

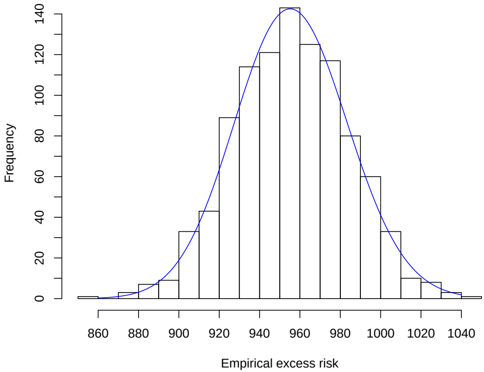
$$\|Z - \mathbb{E}[Z]\|_q$$

$$\leq \quad \sqrt{n\kappa_5'}\left(\underbrace{\omega\left(\sqrt{L(\bar{f})}\right)}_{\text{bias}} + \underbrace{\omega(r_*)}_{\text{variance}}\right) q^{1/2} + \sqrt{n\kappa_6'}\omega(r_*)q\,.$$

High dimensional Wilks phenomenon

Variance proxy and scale proxy depend on model complexity and noise conditions but not directly on sample size.

Learning VC classes



Toy problem from Kearns et al., *Machine Learning*, 1997

- VC-dimension of $\mathcal{F}$: 1600
- $R(f^*) = .2$
- $\omega(r) = \frac{r}{\sqrt{\beta}}$
- $n = 20000$
- 1000 trials, , $\frac{\beta}{2} = .3$,
- $\mathbb{E}[n(R_n(\hat{f}) - R_n(f^*))] \approx 956.$
- Sample variance: $784.$
- Blue line: Gamma$(1165, 1.21)$

Proof (i): Deviation inequalities for L_2 distances

- $\exists \kappa_5$ and κ_6 such that for $q \geq 2$

$$\left\| P\left(\widehat{h}-\bar{h}\right)^2 \right\|_q \vee \left\| P_n\left(\widehat{h}-\bar{h}\right)^2 \right\|_q \leq \kappa_5 \left(\omega^2\left(\sqrt{L(\bar{f})}\right) + \omega^2\left(r_*\right)\right) + \kappa_6 \omega^2(r_*) q \,.$$

- Argument: the same as for deriving deviation inequalities for excess risk.
- Work on $\left\{(\bar{h}-h)^2 : h \in \mathcal{H}\right\}$
  - Risk: expectation !
  - Bounded process ...
  - $P\left((h-\bar{h})^2\right)^2 \leq \omega^2\left(\sqrt{L(h)}\right)$
  - Use contraction principle to get a convenient complexity function

Proof (ii)

- Back to variance bounds:

$$V_+ \leq 2n \left(P_n(\bar{h} - \widehat{h}_n)^2 + P(\bar{h} - \widehat{h}_n)^2 \right) .$$

- For $q \geq 2$:

$$\begin{aligned}
&\|(Z - \mathbb{E}[Z])_+\|_q \\
&\leq \quad \sqrt{3q} \left\| \sqrt{2n \left(P_n(\bar{h} - \widehat{h}_n)^2 + P(\bar{h} - \widehat{h}_n)^2 \right)} \right\|_q \\
&\leq \quad \sqrt{6nq} \left(\sqrt{\left\| P_n(\bar{h} - \widehat{h}_n)^2 \right\|_{q/2}} + \sqrt{\left\| P(\bar{h} - \widehat{h}_n)^2 \right\|_{q/2}} \right)
\end{aligned}$$

Plugging bounds on L_2 distances

$$\leq \quad 2\sqrt{6n\kappa_5} \left(\omega \left(\sqrt{L(\bar{f})} \right) + \omega(r_*) \right) q^{1/2} + 2\sqrt{3n\kappa_6}\, \omega(r_*) q$$

Take home messages

- For the end-user, concentration inequalities provide tail bounds that are good enough to be combined with union bounds
- Separate characterization of expected value and investigation of fluctuations
- ...

What's next ? (hopefully)

- Getting rid of the boundedness/Gaussian assumptions for suprema of empirical processes
- Understanding aspects of super-concentration without resorting to hypercontractivity arguments

Further readings

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

S. B. and P. Massart.
A high dimensional Wilks phenomenon.
Probability Theory and Related Fields. **150** (2011) 405-433.

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

P. Massart.
Concentration inequalities and model selection
Springer. 2006. Lecture Notes in Mathematics 1896.

S. B., O. Bousquet, and G. Lugosi.
Theory of classification: some recent advances.
ESAIM Probability & Statistics, pages 329–375, 2006.

E. Giné and V. Koltchinskii.
Concentration inequalities and asymptotic results for ratio type empirical processes.
Annals of Probability, 34(3):1143–1216, 2006.

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