Concentration

for functions of bounded

interaction

Andreas Maurer

.

Which properties of a bounded function

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

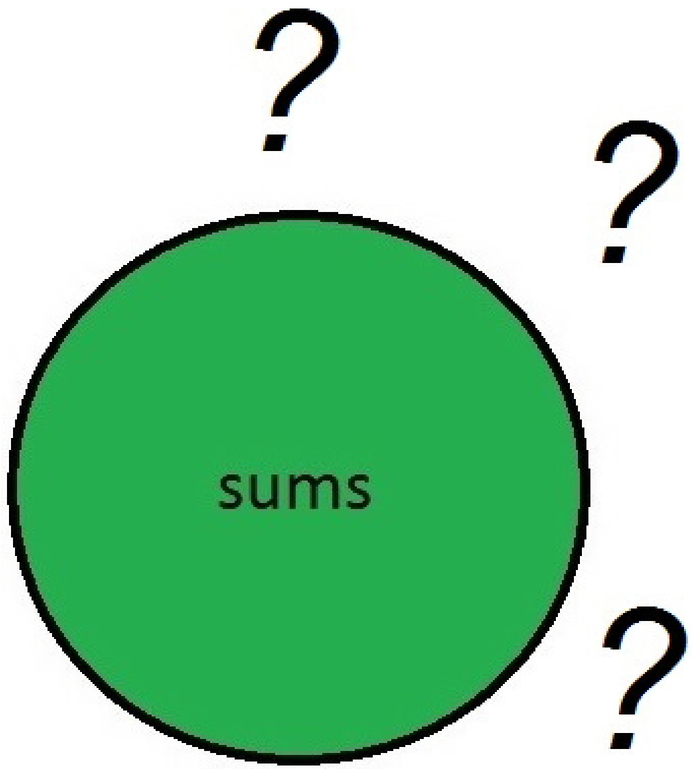
guarantee nice behaviour of the random variable $f(\mathbf{X})$,
with $X = (X_1, ..., X_n)$ a vector of independent random variables?

Nice behaviour of sums

$$f(\mathbf{x}) = \sum_{i=1}^{n} g_i(x_i) \text{ with } g_i : \mathcal{X} \to [a, b].$$

$$\begin{array}{lrcl}
\text{Additive variance} & \sigma^2(f) & = & \sum_{i=1}^{n} \sigma^2(g_i) \\
\text{normal approximation} & \dfrac{f - Ef}{\sigma(f)} & \approx & \mathcal{N}(0, 1) \text{ for large } n \\
\text{Hoeffding inequality} & \Pr\{f - Ef > t\} & \leq & \exp\left(\dfrac{-2t^2}{n(b-a)^2}\right) \\
\text{Bernstein inequality} & \Pr\{f - Ef > t\} & \leq & \exp\left(\dfrac{-t^2}{2\sigma^2(f) + 2(b-a)t/3}\right)
\end{array}$$

What about functions which are not sums?



The bounded difference inequality

Partial difference operator

$$D^k_{y,y'} f(\mathbf{x}) := f(\ldots, x_{k-1}, y, x_{k+1}, \ldots) - f\left(\ldots, x_{k-1}, y', x_{k+1}, \ldots\right).$$

Define maximal variation in any argument

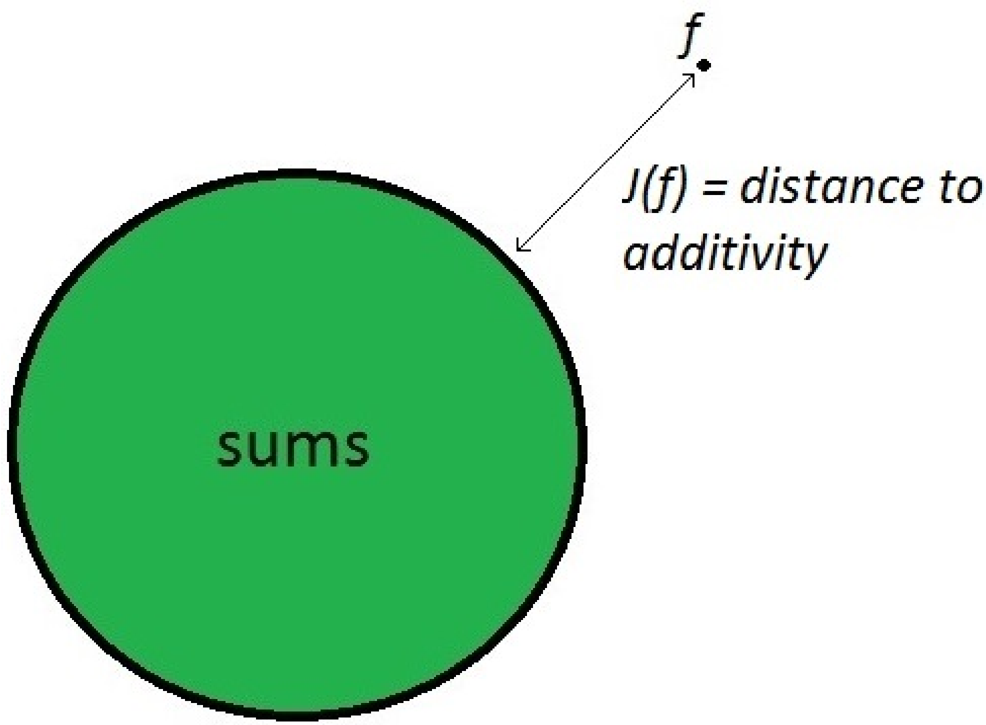
$$M(f) := \max_k \sup_{\mathbf{x}, y, y'} D^k_{y,y'} f(\mathbf{x}).$$

Theorem (Hoeffding, Azuma, McDiarmid):

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-2t^2}{nM(f)^2}\right), \text{ for all } f : \mathcal{X}^n \to \mathbb{R}$$

Extends Hoeffding's inequality to general functions.

What about functions which are close to being sums?



Interaction functional

For bounded $f : \mathcal{X}^n \to \mathbb{R}$ define

$$J(f) := n \max_{k,l:k\neq l} \sup_{\mathbf{x},y,y',z,z'} D^l_{z,z'} D^k_{y,y'} f(\mathbf{x}).$$

- J is a seminorm which vanishes for sums

- Key property for "sum-like" behaviour:
  $J(f)$ is at most of the same order in n as $M(f)$
  or, put differently:
  maximal mixed second difference $\leq O(1/n) \times$ maximal first difference

Weak interactions

Definition:

$f : \mathcal{X}^n \to \mathbb{R}$ has (a, b)-*weak interactions*,
if $M(f) \leq a/n$ and $J(f) \leq b/n$.

A sequence $(f_n)_{n \geq 2}$ of functions $f_n : \mathcal{X}^n \to \mathbb{R}$ has (a, b)-*weak interactions* if every f_n has (a, b)-weak interactions.

Outline

Examples of weak interactions:
- U- and V-statistics
- Lipschitz L-statistics
- Generalization error of the Gibbs algorithm
- Generalization error of ℓ_2-regularized classification

Properties of weak interactions:
- Small bias in the Efron-Stein inequality
- Bernstein's inequality
- Variance estimation
- Normal approximation

Proof of Bernstein's inequality

V- and U-statistics

Fix $1 \leq m < n$,
for $\mathbf{j} = (j_1, ..., j_m) \in \{1, ..., n\}^m$ let

$$\kappa_{\mathbf{j}} : \mathcal{X}^m \to \mathbb{R}, \left|\kappa_{\mathbf{j}}\right| \leq 1$$

and define $V, U : \mathcal{X}^m \to \mathbb{R}$,

$$\begin{aligned} V(\mathbf{x}) &= n^{-m} \sum_{\mathbf{j} \in \{1,...,n\}^m} \kappa_{\mathbf{j}} \left(x_{j_1}, ..., x_{j_m} \right) \\ U(\mathbf{x}) &= \binom{n}{m}^{-1} \sum_{1 \leq j_1 < ... < j_m \leq m} \kappa_{\mathbf{j}} \left(x_{j_1}, ..., x_{j_m} \right) \end{aligned}$$

$V =$ **V**on Mises statistic (1947)
$U =$ **U**nbiased statistic (Hoeffding, 1948)

V- and U-statistics have weak interactions

$$
\begin{aligned}
V(\mathbf{x}) &= n^{-m} \sum_{\mathbf{j}\in\{1,\dots,n\}^m} \kappa_{\mathbf{j}}\left(x_{j_1},\dots,x_{j_m}\right) \\
D^k_{y,y'}V(\mathbf{x}) &\leq \frac{2}{n^m}\left|\{\mathbf{j} : k\in\mathbf{j}\}\right| = \frac{2}{n^m}\left|\bigcup_{r=1}^{m}\left\{\mathbf{j} : r = \min_{j_i=k} i\right\}\right| \\
&= \frac{2mn^{m-1}}{n^m} = \frac{2m}{n} \\
D^l_{z,z'}D^{k:k\neq l}_{y,y'}V(\mathbf{x}) &\leq \frac{4}{n^m}\left|\{\mathbf{j} : k,l\in\mathbf{j}\}\right| = \frac{4}{n^m}\left|\bigcup_{r,s:r\neq s}\left\{\mathbf{j} : r = \min_{j_i=k} i \wedge s = \min_{j_i=l} i\right\}\right| \\
&= \frac{4m(m-1)n^{m-2}}{n^m} = \frac{4m(m-1)}{n^2}.
\end{aligned}
$$

So V has $(2m, 4m(m-1))$-weak interactions!

Similar argument and result for U (A.M,17a)

Lipschitz L-statistics

$\mathcal{X} = [a, b]$ and $\left(x_{(1)}, ..., x_{(n)}\right)$ = order statistic of $\mathbf{x} \in \mathcal{X}^n$

$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} F(i/n)\, x_{(i)}$$

where $F : [0, 1] \to \mathbb{R}$ has Lipschitz constant L.

Examples: mean, smoothly trimmed mean, smoothed quantiles, etc

Then f as $(L(b-a), L(b-a))$-weak interactions
(bound on $M(f)$ easy, bound on $J(f)$ cumbersome - many cases)

A chain rule

Extend definition of M and J to Banach space-valued functions $f : \mathcal{X}^n \to B$

$M(f) = \max_k \sup_{x,y,y'} \left\| D^k_{yy'} f(x) \right\|$ and $J(f) = n \max_{k \neq l} \sup_{x,y,y',z,z'} \left\| D^l_{zz'} D^k_{yy'} f(x) \right\|$.

Lemma: B be a Banach space, $U \subseteq B$ convex, $f : \mathcal{X}^n \to U$, $F : U \to \mathbb{R}$ be twice Fréchet-differentiable. Then

$$\begin{aligned} M(F \circ f) &\leq \sup_{v \in U} \left\| F'(v) \right\| M(f) \text{ and} \\ J(F \circ f) &\leq n \sup_{v \in U} \left\| F''(v) \right\| M(f)^2 + \sup_{v \in U} \left\| F'(v) \right\| J(f). \end{aligned}$$

If f has weak interactions and $\|F''(v)\|$ and $\|F'(v)\|$ are bounded on U, then $F \circ f$ also has weak interactions.

Free energy and Gibbs distributions

Ω a mble space with finite, positive measure ρ.
$H : \Omega \times \mathcal{X}^n \to [-1, 1]$ a "Hamiltonian",
$\beta > 0$ an "inverse temperature".

For $x \in \mathcal{X}^n$ define $$A_H(\mathbf{x}) = \ln Z_H(\mathbf{x}) := \ln \int_\Omega e^{-\beta H(\omega,\mathbf{x})} d\rho(\omega)$$

The chain rule with

$$\begin{aligned} f &: \mathbf{x} \in \mathcal{X}^n \mapsto H(., \mathbf{x}) \in L_\infty(\Omega) \text{ and} \\ F &: G(.) \in L_\infty(\Omega) \mapsto \ln \int_\Omega e^{-\beta G(\omega)} d\rho(\omega) \end{aligned}$$

gives $$M(A_H) \leq \beta M(H) \text{ and } J(A_H) \leq \beta J(H) + 2n\beta^2 M(H)^2$$

Also define Gibbs distribution on Ω

$$d\pi_H(\mathbf{x}) = Z_H^{-1}(x) e^{-\beta H(\omega,\mathbf{x})} d\rho(\omega)$$

"Generalization" of the Gibbs algorithm

$$
\begin{aligned}
\text{loss of model } \omega \text{ on datum } x &: \quad \ell(\omega, x) \;\text{ where } \ell : \Omega \times \mathcal{X} \to [0,1] \\
\text{empirical loss on sample } \mathbf{x} &: \quad H(\omega, \mathbf{x}) = \frac{1}{n} \sum_{n=1}^{n} \ell(\omega, x_i) \\
\text{true loss on r.v. } X &: \quad \bar{H}(\omega) = E_X[\ell(\omega, X)]. \\
\text{Gibbs measure for empirical loss} &: \quad d\pi_H(\mathbf{x}) \\
\text{Gibbs measure for true loss} &: \quad d\pi_{\bar{H}} \\
\text{a measure of "generalization"} &: \quad KL(d\pi_H(\mathbf{x}), d\pi_{\bar{H}}) =: f(\mathbf{x})
\end{aligned}
$$

By the chain rule

$$M(f) \leq \frac{6\beta(1+2\beta)}{n} \text{ and } J(f) \leq \frac{24\beta^2(1+2\beta)}{n},$$

So f has $\left(6\beta(1+2\beta), 24\beta^2(1+2\beta)\right)$-weak interactions!

Generalization of ℓ_2-regularized algorithms

$(H, \langle ., . \rangle, \|.\|)$ a real Hilbert space with unit ball $\mathcal{X}$
define $g : \mathcal{X}^n \to H$ by

$$
\begin{aligned}
\text{returned weight vector } g(\mathbf{x}) &= \arg\min_{w \in H} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle x_i, w \rangle) + \lambda \|w\|^2 \\
\text{empirical loss} \quad \hat{L}(\mathbf{x}) &= \frac{1}{n} \sum_i \ell(\langle x_i, g(\mathbf{x}) \rangle), \\
\text{true expected loss} \quad L(\mathbf{x}) &= E[\ell(\langle X, g(\mathbf{x}) \rangle)], \\
\text{generalization error} \quad \Delta(\mathbf{x}) &= L(\mathbf{x}) - \hat{L}(\mathbf{x})
\end{aligned}
$$

Then Δ has $\left(O\left(\lambda^{-3/2}\right)\|\ell''\|_\infty, O\left(\lambda^{-3}\right)\|\ell'''\|_\infty\right)$-weak interactions!
(A.M.17b, by implicit differentiation)

Properties of functions with weak interactions

- Small bias in the Efron-Stein inequality
- Bernstein's inequality
- Variance estimation
- Normal approximation

The bias of the Efron-Stein inequality

$$
\begin{aligned}
k\text{-th conditional variance} &: \quad \sigma_k^2(f)(\mathbf{x}) = \frac{1}{2} E_{(y,y')\sim\mu_k\times\mu_k}\left[\left(D^k_{y,y'} f(\mathbf{x})\right)^2\right] \\
\text{sum of conditional variances} &: \quad \Sigma^2(f)(\mathbf{x}) = \sum_{k=1}^{n} \sigma_k^2(f)(\mathbf{x}) \\
\text{Efron-Stein inequality} &: \quad \sigma^2(f) \le E\left[\Sigma^2(f)\right]
\end{aligned}
$$

Theorem (Houdré, 1998):

$$
E\left[\Sigma^2(f)\right] \le \sigma^2(f) + \frac{1}{4}\sum_{k,l:k\neq l} E_{\mathbf{x},z,z',y,y'}\left[\left(D^l_{zz'} D^k_{yy'} f(\mathbf{x})\right)^2\right] \le \sigma^2(f) + \frac{J(f)^2}{4}.
$$

If f has weak interactions then $\sigma^2(f) = E\left[\Sigma^2(f)\right] + O\left(1/n^2\right)$.

Bernstein's inequality

Theorem (A.M.17a): For bounded mble $f : \mathcal{X}^n \to \mathbb{R}$

$$\Pr\left\{f - E\left[f\right] > t\right\} \leq \exp\left(\frac{-t^2}{2E\left[\Sigma^2\left(f\right)\right] + \left(2M\left(f\right)/3 + J\left(f\right)\right)\ t}\right)$$

extends Bernstein's inequality from sums to general functions.

Corollary: If f has (a, b)-weak interactions then
(using $E\left[\Sigma^2\left(f\right)\right] \leq \sigma^2\left(f\right) + J\left(f\right)/4$)
$\forall \delta \in (0, 1/e)$ with probability at least $1 - \delta$

$$f \leq E\left[f\right] + \sqrt{2\sigma^2\left(f\right)\ln\left(1/\delta\right)} + \frac{\left(2a/3 + 2b\right)\ln\left(1/\delta\right)}{n}.$$

Estimating variance

Theorem: For any bounded $f : \mathcal{X}^n \to \mathbb{R}$
there exists $g : \mathcal{X}^{n+1} \to \mathbb{R}$ such that
for any iid sequence $X_1, ..., X_n, ...$ with values in $\mathcal{X}$
and for $0 < \delta \leq 1/e$ with probability at least $1 - \delta$

$$\sqrt{g(\mathbf{X})} - K_1(f)\sqrt{\ln(2/\delta)} \leq \sqrt{\sigma^2(f)} \leq \sqrt{g(\mathbf{X})} + K_2(f)\sqrt{\ln(2/\delta)}$$

$$\begin{aligned} \text{with } K_1(f) &= J(f)/2 + \sqrt{M(f)^2 + 8J(f)^2} \\ \text{and } K_2(f) &= \sqrt{\max\left\{M(f)^2 + 8J(f)^2, M(f)(M(f) + 2J(f))\right\}} \end{aligned}$$

Also: g is an unbiased estimator for the Efron Stein bound $E\left[\Sigma^2(f)\right]$.

The variance estimator

For any n and $\mathbf{x} \in \mathcal{X}^n$ define

$$\begin{array}{lrcl} \text{replacement operator} & S_y^k \mathbf{x} & = & (x_1, ..., x_{k-1}, y, x_{k+1}, ..., x_n) \in \mathcal{X}^n \\ \text{deletion operator} & S_-^k \mathbf{x} & = & (x_1, ..., x_{k-1}, x_{k+1}, ..., x_n) \in \mathcal{X}^{n-1}. \end{array}$$

The variance estimator $g : \mathcal{X}^{n+1} \to \mathbb{R}$ is

$$g(\mathbf{x}) = \frac{1}{2(n+1)} \sum_{i=1}^{n+1} \sum_{j:j\neq i} \left(f\left(S_-^j \mathbf{x}\right) - f\left(S_-^j S_{x_j}^i \mathbf{x}\right)\right)^2.$$

Needs $O\left(n^2\right)$ computations of f, but only a sample of $O(n)$!

So for weak interactions with high probability

$$\sqrt{\sigma^2(f)} = \sqrt{g(X)} + O\left(\frac{1}{n}\right).$$

Normal approximation

$\delta(W, V) = \{\sup |E[h(W)] - E[h(V)]| : h \text{ a Lipschitz-1 function}\}$

Theorem : Let $E[f] = 0$ and $Z \sim \mathcal{N}(0, 1)$. Then

$$\delta(f(\mathbf{X}), Z) \leq \frac{\sqrt{8n}M(f)(J(f) + M(f))}{\sigma^2(f)} + \frac{nM(f)^3}{2\sigma^3(f)}.$$

If (f_n) has (a, b)-weak interactions and $\sigma(f_n) \geq Cn^{-p}$ for constant C, then

$$\delta(f_n(\mathbf{X}), Z) \leq \frac{\sqrt{8}Ca(a+b) + a^3}{C^3 n^{2-3p}}.$$

$(1/2 \leq p < 2/3) \implies$ asymptotic normality.
$(p = 1/2) \implies$ rate is $n^{-1/2}$.

Intermission

A Bernstein-type inequality

Theorem: Let $f : \mathcal{X}^n \to \mathbb{R}$ and for some b and all $m \geq 2$

$$\sum_{k=1}^{n} E_k \left[(f - E_k f)^m \right] \leq \frac{\Sigma^2 (f)}{2} m! \; b^{m-2}.$$

Then for $t > 0$

$$\Pr \{ f - Ef > t \} \leq \exp \left(\frac{-t^2}{2E \left[\Sigma^2 (f) \right] + (2b + J (f)) \, t} \right).$$

The proof uses the *entropy method* (Boltzmann, Gibbs, Shannon, Nelson, Lieb, Ledoux, Bobkov, Massart, Boucheron, Lugosi, and many others)

We prove this for $b = 1$. The theorem then follows from rescaling

Definitions

$$
\begin{array}{lrcl}
\text{thermal measure} & \mu_{\beta f} & : & = e^{\beta f} d\mu / E\left[e^{\beta f}\right] \\
\text{thermal expectation} & E_{\beta f}[g] & : & = E\left[g e^{\beta f}\right] / E\left[e^{\beta f}\right] \\
\text{thermal variance} & \sigma^2_{\beta f}[g] & : & = E_{\beta f}\left[\left(g - E_{\beta f}[g]\right)^2\right] \\
\text{entropy} & \text{Ent}(\beta f) & : & = KL\left(d\mu_{\beta f}, d\mu\right) = \beta E_{\beta f}[f] - \ln E\left[e^{\beta f}\right] \\
\text{conditional expectation} & E_k[g] & : & = E\left[g | X_1, ..., X_{k-1}, X_{k+1}, ..., X_n\right]
\end{array}
$$

the conditional quantities $E_{k,\beta f}[.]$, $\sigma^2_{k,\beta f}[.]$, $\text{Ent}_k(\beta f)$ are all w.r.t. $E_k[.]$

$$
\begin{array}{lrcl}
\text{replacement operator} & S^k_y x & = & (x_1, ..., x_{k-1}, y, x_{k+1}, ..., x_n) \in \mathcal{X}^n \\
 & \left(S^k_y f\right)(x) & = & f\left(S^k_y x\right) \text{ for } f : \mathcal{X}^n \to \mathcal{Y}
\end{array}
$$

Facts used in the proof

Markov inequality	$\Pr\{f - Ef > t\}$	$\leq$	$\inf_{\beta \geq 0} \exp\left(\ln E\left[e^{\beta(f-Ef)}\right] - \beta t\right)$
representation of mgf	$\ln E\left[e^{\beta(f-Ef)}\right]$	$=$	$\beta \int_0^\beta \frac{\mathrm{Ent}(\gamma f)}{\gamma^2} d\gamma$
subadditivity of entropy	$\mathrm{Ent}(\beta f)$	$\leq$	$E_{\beta f}\left[\sum_{k=1}^n \mathrm{Ent}_k(\beta f)\right]$
fluctuation representation	$\mathrm{Ent}(\beta f)$	$=$	$\int_0^\beta \int_t^\beta \sigma_{sf}^2[f]\, ds\ dt$
	$\mathrm{Ent}(\beta f)$	$\leq$	$E_{\beta f}\left[\sum_{k=1}^n \int_0^\beta \int_t^\beta \sigma_{k,sf}^2[f]\, ds\ dt\right]$
Fenchel-Young inequality	$E_{\beta f}[g]$	$\leq$	$\mathrm{Ent}(\beta f) + \ln E\left[e^g\right]$

The sum of conditional variances, I

Lemma: Let $\beta \in (0,1)$ and suppose that for all $m \geq 2$

$$\sum_{k=1}^{n} E_k\left[(f - E_k f)^m\right] \leq \frac{\Sigma^2(f)}{2} m!.$$

$$\text{Then} \qquad \text{Ent}(\beta f) \leq \frac{\beta^2}{2(1-\beta)^2} E_{\beta f}\left[\Sigma^2(f)\right].$$

Proof: For $s \in (0, \beta)$

$$\begin{aligned}
\sigma_{k,sf}^2(f) &\leq E_{k,sf}\left[(f - E_k(f))^2\right] \qquad \text{(variational property of variance)} \\
&= \frac{E_k\left[(f - E_k(f))^2 e^{s(f - E_k f)}\right]}{E_k\left[e^{s(f - E_k f)}\right]}
\end{aligned}$$

The sum of conditional variances, II

Lemma: Let $\beta \in (0,1)$ and suppose that for all $m \geq 2$

$$\sum_{k=1}^{n} E_k \left[(f - E_k f)^m \right] \leq \frac{\Sigma^2(f)}{2} m!.$$

$$\text{Then} \qquad \text{Ent}(\beta f) \leq \frac{\beta^2}{2(1-\beta)^2} E_{\beta f} \left[\Sigma^2(f) \right].$$

Proof: For $s \in (0, \beta)$

$$\begin{aligned} \sigma^2_{k,sf}(f) &\leq \frac{E_k\left[(f - E_k(f))^2 e^{s(f - E_k f)} \right]}{E_k\left[e^{s(f-E_k f)} \right]} \\ &\leq E_k\left[(f - E_k(f))^2 e^{s(f - E_k f)} \right] \qquad \text{(Jensen's inequality)} \end{aligned}$$

The sum of conditional variances, III

Lemma: Let $\beta \in (0,1)$ and suppose that for all $m \geq 2$

$$\sum_{k=1}^{n} E_k\left[(f - E_k f)^m\right] \leq \frac{\Sigma^2(f)}{2} m!.$$

$$\text{Then} \qquad \mathrm{Ent}(\beta f) \leq \frac{\beta^2}{2(1-\beta)^2} E_{\beta f}\left[\Sigma^2(f)\right].$$

Proof: For $s \in (0, \beta)$

$$\begin{aligned} \sum_{k=1}^{n} \sigma_{k,sf}^2(f) &\leq \sum_{k=1}^{n} E_k\left[(f - E_k f)^2 e^{s(f - E_k f)}\right] \\ &= \sum_{m=0}^{\infty} \frac{s^m}{m!} \sum_{k=1}^{n} E_k\left[(f - E_k f)^{m+2}\right] \end{aligned}$$

The sum of conditional variances, IV

Lemma: Let $\beta \in (0, 1)$ and suppose that for all $m \geq 2$

$$\sum_{k=1}^{n} E_k \left[(f - E_k f)^m \right] \leq \frac{\Sigma^2(f)}{2} m!.$$

$$\text{Then} \qquad \text{Ent}(\beta f) \leq \frac{\beta^2}{2(1-\beta)^2} E_{\beta f} \left[\Sigma^2(f) \right].$$

Proof: For $s \in (0, \beta)$

$$\begin{aligned} \sum_{k=1}^{n} \sigma_{k,sf}^2(f) &\leq \sum_{m=0}^{\infty} \frac{s^m}{m!} \sum_{k=1}^{n} E_k \left[(f - E_k f)^{m+2} \right] \\ &\leq \frac{\Sigma^2(f)}{2} \sum_{m=0}^{\infty} (m+1)(m+2) s^m \text{ (by hypothesis)} \end{aligned}$$

The sum of conditional variances, V

Lemma: Let $\beta \in (0,1)$ and suppose that for all $m \geq 2$

$$\sum_{k=1}^{n} E_k\left[(f - E_k f)^m\right] \leq \frac{\Sigma^2(f)}{2} m!.$$

$$\text{Then} \qquad \operatorname{Ent}(\beta f) \leq \frac{\beta^2}{2(1-\beta)^2} E_{\beta f}\left[\Sigma^2(f)\right].$$

Proof:

$$\begin{aligned} \operatorname{Ent}(\beta f) &\leq E_{\beta f}\left[\int_0^\beta \int_t^\beta \sum_{k=1}^{n} \sigma^2_{k,sf}(f)\; ds\; dt\right] \text{ (subadditivity + fluctuation rep.)} \\ &\leq \frac{E_{\beta f}\left[\Sigma^2(f)\right]}{2} \sum_{m=0}^{\infty} (m+1)(m+2) \int_0^\beta \int_t^\beta s^m ds dt \end{aligned}$$

The sum of conditional variances, VI

Lemma: Let $\beta \in (0,1)$ and suppose that for all $m \geq 2$

$$\sum_{k=1}^{n} E_k\left[(f - E_k f)^m\right] \leq \frac{\Sigma^2(f)}{2} m!.$$

$$\text{Then} \qquad \mathrm{Ent}(\beta f) \leq \frac{\beta^2}{2(1-\beta)^2} E_{\beta f}\left[\Sigma^2(f)\right].$$

Proof:

$$\begin{aligned}
\mathrm{Ent}(\beta f) &\leq \frac{E_{\beta f}\left[\Sigma^2(f)\right]}{2} \sum_{m=0}^{\infty} (m+1)(m+2) \int_0^\beta \int_t^\beta s^m \, ds \, dt \\
&= \frac{E_{\beta f}\left[\Sigma^2(f)\right]}{2} \beta^2 \sum_{m=0}^{\infty} (m+1) \beta^m \qquad \square
\end{aligned}$$

Spin-off: another version of Bernstein's inequality

Theorem (McDiarmid 1998): If f satisfies the conditions of the lemma then

$$\begin{aligned}
\Pr\{f - Ef > t\} &\leq \inf_{\beta\in(0,1)} \exp\left(\beta \int_0^\beta \frac{\mathrm{Ent}\left(\gamma f\right)}{\gamma^2} d\gamma - \beta t\right) \\
&\leq \inf_{\beta\in(0,1)} \exp\left(\beta \int_0^\beta \frac{1}{2\left(1-\gamma\right)^2} E_{\gamma f}\left[\Sigma^2\left(f\right)\right] d\gamma - \beta t\right) \\
&\leq \inf_{\beta\in(0,1)} \exp\left(\frac{\left\|\Sigma^2\left(f\right)\right\|_\infty}{2} \frac{\beta^2}{1-\beta} - \beta t\right) \\
&\leq \exp\left(\frac{-t^2}{2\left(\left\|\Sigma^2\left(f\right)\right\|_\infty + t\right)}\right)
\end{aligned}$$

Another bound on entropy, I

Define
$$D^2 f \quad : \quad = \sum_k \left(f - \inf_{y\in\mathcal{X}} S_y^k f \right)^2$$

Lemma:
$$\mathsf{Ent}\left(\beta f\right) \quad \leq \quad \frac{\beta^2}{2} E_{\beta f}\left[D^2\left(f\right) \right]$$

Proof: For $0 < s \leq \beta$. Let $h := f - \inf_{y\in\mathcal{X}} S_y^k f$.

$$\begin{aligned}
\frac{d}{ds} E_{k,sf}\left[h^2\right] &= \frac{d}{ds} E_{k,sh}\left[h^2\right] \\
&= E_{k,sh}\left[h^3\right] - E_{k,sh}\left[h^2\right] E_{k,sh}\left[h\right] \geq 0 \\
\text{so} \quad \sigma^2_{k,sf}\left(f\right) &\leq E_{k,sf}\left[\left(f - \inf_{y\in\mathcal{X}} S_y^k f \right)^2 \right] \leq E_{k,\beta f}\left[\left(f - \inf_{y\in\mathcal{X}} S_y^k f \right)^2 \right]
\end{aligned}$$

Another bound on entropy, II

Define $$D^2 f \quad : \quad = \sum_k \left(f - \inf_{y \in \mathcal{X}} S_y^k f \right)^2$$

Lemma: $$\text{Ent}\left(\beta f\right) \leq \frac{\beta^2}{2} E_{\beta f}\left[D^2\left(f\right)\right]$$

Proof: For $0 < s \leq \beta$.

$$\begin{aligned}
\text{Ent}\left(\beta f\right) &\leq E_{\beta f}\left[\int_0^\beta \int_t^\beta \sum_{k=1}^n \sigma_{k,sf}^2\left(f\right) \, ds \, dt\right] \\
&\leq E_{\beta f}\left[\int_0^\beta \int_t^\beta \sum_{k=1}^n E_{k,\beta f}\left[\left(f - \inf_{y \in \mathcal{X}} S_y^k f\right)^2\right] ds \, dt\right] \\
&= \frac{\beta^2}{2} E_{\beta f}\left[D^2\left(f\right)\right] \quad \text{(because } E_{\beta f}\left[E_{k,\beta f}\left[.\right]\right] = E_{\beta f}\left[.\right] \text{)} \qquad \square
\end{aligned}$$

Spin-off concentration inequality

Theorem: Let $f : \mathcal{X}^n \to \mathbb{R}$

$$
\begin{aligned}
\Pr\{f - Ef > t\} &\leq \inf_{\beta > 0} \exp\left(\beta \int_0^\beta \frac{\mathrm{Ent}(\gamma f)}{\gamma^2} d\gamma - \beta t\right) \\
&\leq \inf_{\beta > 0} \exp\left(\frac{\beta^2 \left\|D^2(f)\right\|_\infty}{2} - \beta t\right) \\
&\leq \exp\left(\frac{-t^2}{2\left\|D^2(f)\right\|_\infty}\right)
\end{aligned}
$$

Applications: Concentration of convex Lipschitz functions, shortest T.S.P., largest eigenvalue of random symmetric matrix, and many more ...

Self-bounded functions

Lemma: If $D^2 f \leq a^2\ f$ then for $\beta \in \left(0, 2/a^2\right)$

$$\ln E\left[e^{\beta f}\right] \leq \frac{\beta}{1 - a^2\beta/2} E\left[f\right].$$

Proof:

$$\begin{aligned}
\ln E\left[e^{\beta(f - E[f])}\right] &= \beta \int_0^\beta \frac{\mathrm{Ent}\left(\gamma f\right)}{\gamma^2} d\gamma \\
&\leq \frac{\beta}{2} \int_0^\beta E_{\gamma f}\left[D^2 f\right] d\gamma \\
&\leq \frac{a^2\beta}{2} \int_0^\beta E_{\gamma f}\left[f\right] d\gamma = \frac{a^2\beta}{2} \ln E\left[e^{\beta f}\right] \qquad \square
\end{aligned}$$

The sum of conditional variances is self-bounded, I

Proposition:

$$D^2\left(\Sigma^2(f)\right) \leq J(f)^2 \ \Sigma^2(f)$$

Proof: Fix $x \in \mathcal{X}^n$ and let $z \in \mathcal{X}^n$, $z_l := \arg\min_z S_z^l \Sigma^2(f)$.

$$\begin{aligned} D^2\left(\Sigma^2(f)\right) &= \sum_l \left(\Sigma^2(f) - S_{z_l}^l \Sigma^2(f)\right)^2 \\ &= \sum_l \left(\sum_{k:k\neq l} \left(\sigma_k^2(f) - S_{z_l}^l \sigma_k^2(f)\right)\right)^2. \end{aligned}$$

The sum of conditional variances is self-bounded, II

Proposition:

$$D^2\left(\Sigma^2(f)\right) \leq J(f)^2\ \Sigma^2(f)$$

Proof: Fix $x \in \mathcal{X}^n$ and let $z \in \mathcal{X}^n$, $z_l := \arg\min_z S_z^l \Sigma^2(f)$.

$$\begin{aligned}
4D^2\left(\Sigma^2(f)\right) &= 4\sum_l \left(\sum_{k:k\neq l}\left(\sigma_k^2(f) - S_{z_l}^l \sigma_k^2(f)\right)\right)^2 \\
&= \sum_l \left(\sum_{k\neq l} E_{(y,y')\sim\mu_k^2}\left[\left(D_{y,y'}^k f\right)^2 - \left(D_{y,y'}^k S_{z_l}^l f\right)^2\right]\right)^2
\end{aligned}$$

The sum of conditional variances is self-bounded, III

Proposition:

$$D^2\left(\Sigma^2(f)\right) \leq J(f)^2\ \Sigma^2(f)$$

Proof: Fix $x \in \mathcal{X}^n$ and let $z \in \mathcal{X}^n$, $z_l := \arg\min_z S_z^l \Sigma^2(f)$.

$$\begin{aligned}
&4D^2\left(\Sigma^2(f)\right)\\
&= \sum_l \left(\sum_{k \neq l} E_{(y,y') \sim \mu_k^2} \left[\left(D_{y,y'}^k f \right)^2 - \left(D_{y,y'}^k S_{z_l}^l f \right)^2 \right] \right)^2\\
&= \sum_l \left(\sum_{k \neq l} E_{(y,y') \sim \mu_k^2} \left[\left(D_{y,y'}^k f - D_{y,y'}^k S_{z_l}^l f \right) \left(D_{y,y'}^k f + D_{y,y'}^k S_{z_l}^l f \right) \right] \right)^2
\end{aligned}$$

The sum of conditional variances is self-bounded, IV

Proposition:

$$D^2\left(\Sigma^2(f)\right) \leq J(f)^2\ \Sigma^2(f)$$

Proof: Fix $x \in \mathcal{X}^n$ and let $z \in \mathcal{X}^n$, $z_l := \arg\min_z S_z^l \Sigma^2(f)$.

$$4D^2\left(\Sigma^2(f)\right)$$

$$= \sum_l \left(\sum_{k \neq l} E_{(y,y')\sim\mu_k^2}\left[\left(D_{y,y'}^k f - D_{y,y'}^k S_{z_l}^l f\right)\left(D_{y,y'}^k f + D_{y,y'}^k S_{z_l}^l f\right)\right]\right)^2$$

$$\leq \sum_l \sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left[D_{y,y'}^k\left(f - S_{z_l}^l f\right)\right]^2 \sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left[D_{y,y'}^k f + D_{y,y'}^k S_{z_l}^l f\right]^2$$

The sum of conditional variances is self-bounded, V

Proposition:

$$D^2\left(\Sigma^2(f)\right) \leq J(f)^2\ \Sigma^2(f)$$

Proof: Fix $x \in \mathcal{X}^n$ and let $z \in \mathcal{X}^n$, $z_l := \arg\min_z S_z^l \Sigma^2(f)$.

$$\begin{aligned}
&4D^2\left(\Sigma^2(f)\right)\\
&\leq \sum_l \sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left[D_{y,y'}^k\left(f - S_{z_l}^l f\right)\right]^2 \sum_{k:k\neq l} E_{(y,y')\sim\mu_k^2}\left[D_{y,y'}^k f + D_{y,y'}^k S_{z_l}^l f\right]^2\\
&\leq 2\sum_l \sum_{k:k\neq l} \sup_{z,z',y,y'}\left[D_{z,z'}^l D_{y,y'}^k(f)\right]^2\left(\Sigma^2(f) + S_{z_l}^l \Sigma^2(f)\right)\\
&\leq 4J(f)^2\,\Sigma^2(f) \qquad \square
\end{aligned}$$

Proof of Bernstein's inequality, I

Let $0 < \gamma \leq \beta < 1/\left(1 + J/2\right)$,
$\theta := \gamma / \left(J\left(1-\gamma\right)\right) \implies \gamma^2 / \left(2\left(1-\gamma\right)^2\right) < \theta < 2/J^2$.

$$\begin{aligned}
\theta \ \mathrm{Ent}\left(\gamma f\right) &\leq \frac{\gamma^2}{2\left(1-\gamma\right)^2} E_{\gamma f}\left[\theta\ \Sigma^2\left(f\right)\right] && \text{(1st Lemma)} \\
&\leq \frac{\gamma^2}{2\left(1-\gamma\right)^2}\left(\mathrm{Ent}\left(\gamma f\right) + \ln E\left[e^{\theta \Sigma^2(f)}\right]\right) && \text{(Fenchel-Young)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Ent}\left(\gamma f\right)\left(\theta - \frac{\gamma^2}{2\left(1-\gamma\right)^2}\right) &\leq \frac{\gamma^2}{2\left(1-\gamma\right)^2} \ln E\left[e^{\theta \Sigma^2(f)}\right] \\
\mathrm{Ent}\left(\gamma f\right) &\leq \frac{\gamma J}{2\left(1 - \left(1 + \left(J/2\right)\right)\gamma\right)} \ln E\left[e^{\theta \Sigma^2(f)}\right]
\end{aligned}$$

Proof of Bernstein's inequality, II

Let $0 < \gamma \leq \beta < 1/\left(1 + J/2\right)$,
$\theta := \gamma/\left(J\left(1-\gamma\right)\right) \implies \gamma^2/\left(2\left(1-\gamma\right)^2\right) < \theta < 2/J^2$.

$$\begin{aligned}
\text{Ent}\left(\gamma f\right) &\leq \frac{\gamma J}{2\left(1-\left(1+\left(J/2\right)\right)\gamma\right)} \ln E\left[e^{\theta\Sigma^2(f)}\right] \\
\ln E\left[e^{\theta\Sigma^2(f)}\right] &\leq \frac{\theta}{1-J^2\theta/2} E\left[\Sigma^2\left(f\right)\right] \qquad \text{(self bounded } \Sigma^2\left(f\right)\text{)} \\
&= \frac{\gamma/J}{1-\left(1+J/2\right)\gamma} E\left[\Sigma^2\left(f\right)\right]. \\
\text{Ent}\left(\gamma f\right) &\leq \frac{\gamma^2}{2\left(1-\left(1+J/2\right)\gamma\right)^2} E\left[\Sigma^2\left(f\right)\right]
\end{aligned}$$

Proof of Bernstein's inequality, III

Let $0 < \gamma \leq \beta < 1/\left(1 + J/2\right)$,

$$\begin{aligned}
\mathsf{Ent}\left(\gamma f\right) &\leq \frac{\gamma^2}{2\left(1-\left(1+J/2\right)\gamma\right)^2} E\left[\Sigma^2\left(f\right)\right] \\
\Pr\left\{f - Ef > t\right\} &\leq \inf_{\beta\in(0,1/(1+J/2))} \exp\left(\beta\int_0^\beta \frac{\mathsf{Ent}\left(\gamma f\right)}{\gamma^2} d\gamma - \beta t\right) \\
&\leq \inf_{\beta\in(0,1/(1+J/2))} \exp\left(\frac{E\left[\Sigma^2\left(f\right)\right]}{2}\frac{\beta^2}{1-\left(1+J/2\right)\beta} - \beta t\right) \\
&\leq \exp\left(\frac{-t^2}{2\left(E\left[\Sigma^2\left(f\right)\right]+\left(1+J/2\right)t\right)}\right) \qquad \square
\end{aligned}$$

References

[1] S.Bernstein, Theory of Probability, Moscow, 1927.

[2] S.Boucheron,G.Lugosi,P.Massart, Concentration Inequalities using the entropy method, *Annals of Probability* 31, Nr 3, 2003

[3] S.Boucheron, G.Lugosi, P.Massart, On concentration of self-bounding functions, *Electronic Journal of Probability* Vol.14 (2009), Paper no. 64, 1884–1899, 2009

[4] S. Boucheron, G. Lugosi, P. Massart. Concentration Inequalities, Oxford University Press (2013)

[5] Efron, B., & Stein, C. (1981). The jackknife estimate of variance. The Annals of Statistics, 586-596.

[6] M.Ledoux, *The Concentration of Measure Phenomenon*, AMS Surveys and Monographs 89, 2001.

[7] A.Maurer, Thermodynamics and concentration. *Bernoulli* 18.2 (2012): 434-454.

[8] Maurer, A. (2017). A Bernstein-type inequality for functions of bounded interaction. arXiv preprint arXiv:1701.06191.

[9] C.McDiarmid, Concentration, in *Probabilistic Methods of Algorithmic Discrete Mathematics*, p. 195–248. Springer, Berlin, 1998.

[10] J.M.Steele, An Efron-Stein inequality for nonsymmetric statistics, *Annals of Statistics* 14:753–758, 1986