Concentration properties and examples of functions with weak interactions

Andreas Maurer

Setting

$\mathcal{X}$ a space of potential observations

$f : \mathcal{X}^n \rightarrow \mathbb{R}$ a bounded function

$\mathbf{X} = (X_1, ..., X_n)$ a random vector of independent observations

Question

Which properties of f could guarantee,

that observation of $\mathbf{X}$ provides useful information on $W = f(\mathbf{X})$

(that is on $E[f] = E[W]$, $\sigma^2[f] = \sigma^2[W]$, other moments etc)?

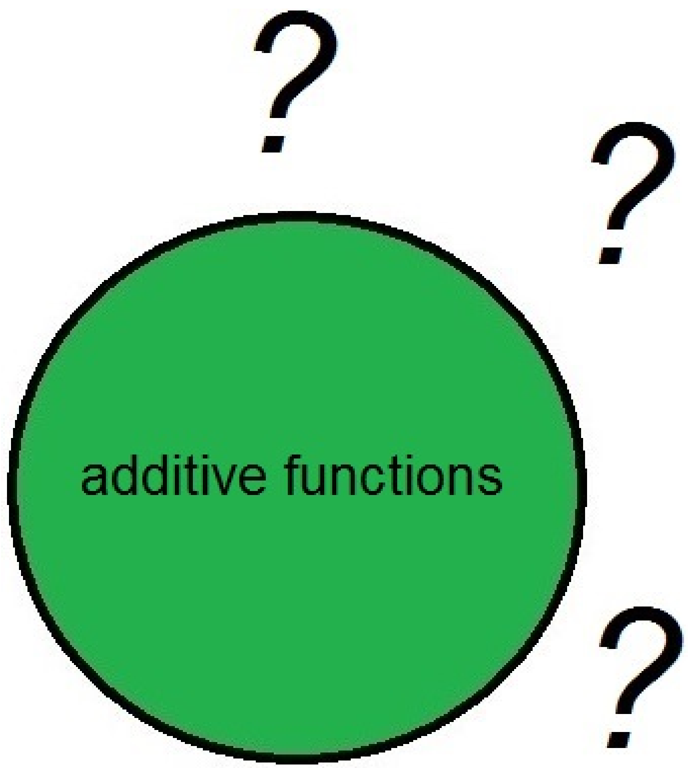
Additive functions work well

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

Then we have

normal approximation	$\dfrac{f(\mathbf{X}) - Ef}{\sigma(f)}$	$\approx$	$\mathcal{N}(0, 1)$ for large n
Hoeffding inequality	$\Pr\{f(\mathbf{X}) - Ef > t\}$	$\leq$	$\exp\left(\dfrac{-2t^2}{n(b-a)^2}\right)$
Bernstein inequality	$\Pr\{f(\mathbf{X}) - Ef > t\}$	$\leq$	$\exp\left(\dfrac{-t^2}{2\sigma^2(f) + 2(b-a)t/3}\right)$

What about functions which are not additive?



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

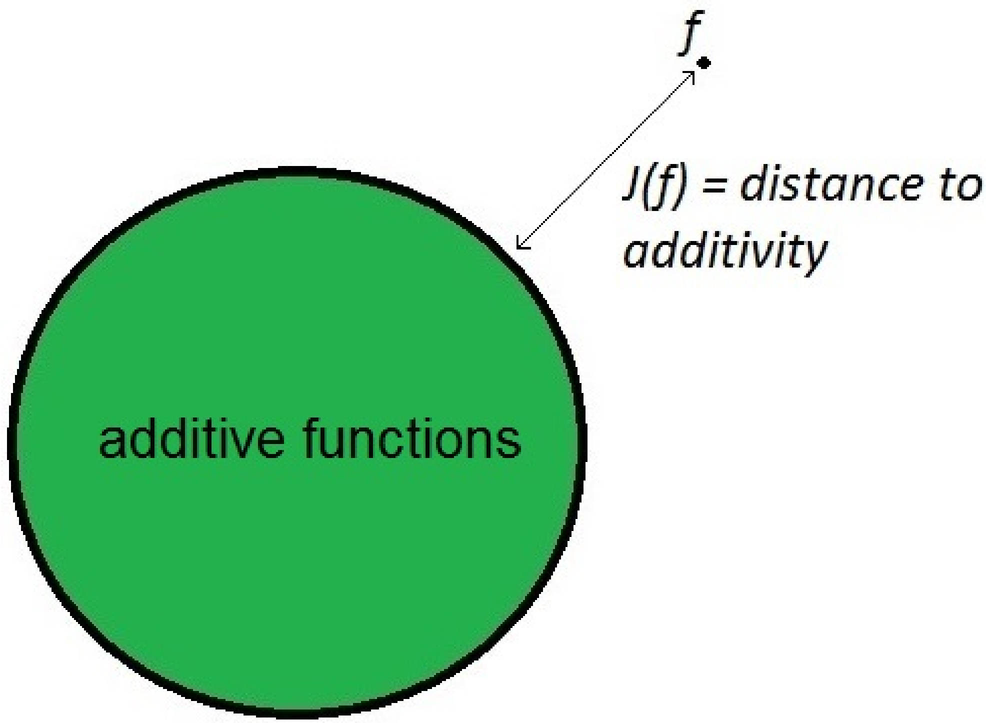
$$M(f) := \max_k \sup_{\mathbf{x},y,y'} D^k_{y,y'} f(\mathbf{x}).$$

Theorem (Hoeffding, Azuma, McDiarmid):

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

Extends Hoeffding's inequality to general functions.

What about functions
which are close to being additive?



Interaction

$$\mathbf{J}(f)_{kl}(\mathbf{x}) = \begin{cases} \sup_{y,y',z,z'} D^l_{z,z'} D^k_{y,y'} f(\mathbf{x}) & \text{if} \quad k \neq l \\ 0 & \text{if} \quad k = l \end{cases}, \text{ for } \mathbf{x} \in \mathcal{X}^n$$

The interaction matrix $\mathbf{J}$ vanishes for additive functions.

A measure of total interaction:

$$\begin{aligned} \sup_{\mathbf{x}\in\mathcal{X}^n} \left\| \mathbf{J}(f)_{kl}(\mathbf{x}) \right\|_{Fr} &= \sup_{\mathbf{x}\in\mathcal{X}^n} \sqrt{\sum_{k\neq l} \left(\sup_{y,y',z,z'} D^l_{z,z'} D^k_{y,y'} f(\mathbf{x}) \right)^2} \\ &\leq n \max_{k,l:k\neq l} \sup_{\mathbf{x},y,y',z,z'} D^l_{z,z'} D^k_{y,y'} f(\mathbf{x}) \\ &= : J(f) = \text{ simplified interaction functional.} \end{aligned}$$

Seminorms

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

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

- M is a seminorm which vanishes on constants
- J is a seminorm which vanishes on additive functions

Weak interactions

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

or equivalently

$\forall k, l \in \{1, ..., n\}, k \neq l, \mathbf{x} \in \mathcal{X}^n$, $y, y', z, z' \in \mathcal{X}$,

$$D^k_{y,y'} f(\mathbf{x}) \leq \frac{a}{n} \text{ and } D^l_{z,z'} D^k_{y,y'} f(\mathbf{x}) \leq \frac{b}{n^2}.$$

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

Concentration and other properties of weak interactions:

- Bernstein's inequality
- Normal approximation
- Variance estimation
- Empirical bounds

Examples of weak interactions:

- U- and V-statistics
- Lipschitz L-statistics
- Generalization error of ℓ_2-regularized classification
- Properties of the Gibbs algorithm

The bias of the Efron-Stein inequality

$$k\text{-th conditional variance} \quad : \quad \sigma_k^2(f)(\mathbf{x}) = \frac{1}{2} E_{(y,y')\sim\mu_k\times\mu_k}\left[\left(D^k_{y,y'} f(\mathbf{x})\right)^2\right]$$

$$\text{sum of conditional variances} \quad : \quad \Sigma^2(f)(\mathbf{x}) = \sum_{k=1}^{n} \sigma_k^2(f)(\mathbf{x})$$

$$\text{Efron-Stein inequality} \quad : \quad \sigma^2(f) \leq E\left[\Sigma^2(f)\right]$$

Theorem (Houdré, 1998):

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

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 (M.2017): For bounded mble $f : \mathcal{X}^n \to \mathbb{R}$

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

extends Bernstein's inequality from sums to general functions.

See also Götze,Sambale 2017 and Bobkov, Götze, Sambale 2017.

Bernstein's inequality

Theorem (M.2017): For bounded mble $f:\mathcal{X}^n\to\mathbb{R}$

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

extends Bernstein's inequality from sums to general functions.

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

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

Normal approximation

Let $Z \sim \mathcal{N}(0,1)$. Define distance to normality of r.v. W:

$$d_{\mathcal{N}}(W) = \sup \left\{ \left| E\left[h\left(\frac{W - E[W]}{\sigma(W)} \right) \right] - E[h(Z)] \right| : h \text{ a real Lipschitz-1 function} \right\}$$

Theorem (M. 2017, nach Chatterjee 2008):

$$d_{\mathcal{N}}\left(f\left(\mathbf{X}'\right)\right) \leq \frac{\sqrt{n}M(f)(J(f)+M(f))}{\sigma^2(f)} + \frac{nM(f)^3}{2\sigma^3(f)}.$$

Normal approximation

Let $Z \sim \mathcal{N}(0,1)$. Define distance to normality of r.v. W:

$$d_{\mathcal{N}}(W) = \sup\left\{\left|E\left[h\left(\frac{W-E[W]}{\sigma(W)}\right)\right] - E[h(Z)]\right| : h \text{ a real Lipschitz-1 function}\right\}$$

Theorem (M. 2017, nach Chatterjee 2008):

$$d_{\mathcal{N}}\left(f\left(\mathbf{X}'\right)\right) \leq \frac{\sqrt{n}M(f)(J(f)+M(f))}{\sigma^2(f)} + \frac{nM(f)^3}{2\sigma^3(f)}.$$

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

$$d_{\mathcal{N}}\left(f\left(\mathbf{X}'\right)\right) \leq \frac{Ca(a+b)+a^3}{C^3 n^{2-3p}}.$$

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

Estimating variance

Theorem (M. 2017): For any bounded $f : \mathcal{X}^n \to \mathbb{R}$
there exists $v_f : \mathcal{X}^{n+1} \to \mathbb{R}$ 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 $1 - \delta$

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

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

Also: v_f 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

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

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

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

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

So for weak interactions with high probability

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

Empirical bounds for weak interactions

Theorem (empirical Bernstein inequality, M., M.Pontil, 2018) :
If f has (a,b)-weak interactions and the X_i are iid, then for $\delta > 0$ with probability at least $1-\delta$

$$f(\mathbf{X}) \leq E[f] + \sqrt{2v_f(\mathbf{X})\ln(2/\delta)} + \frac{(8a/3+5b)\ln(2/\delta)}{n}.$$

Theorem (empirical normal approximation, M., M.Pontil, 2018):
If f has (a,b)-weak interactions and the X_i are iid, then for $\delta > 0$ with probability at least $1-\delta$

$$\text{either} \quad \frac{\sqrt{v_f(\mathbf{X})}}{2} < \frac{\left(b/2+\sqrt{2a^2+8b^2}\right)\sqrt{\ln(1/\delta)}}{n},$$

$$\text{or} \quad d_{\mathcal{N}}\left(f\left(\mathbf{X}'\right)\right) \leq \frac{4\left(a^2+ab\right)}{v_f(\mathbf{X})\,n^{3/2}} + \frac{4a^3}{v_f(\mathbf{X})^{3/2}\,n^2}.$$

Examples of functions with weak interactions

- U- and V-statistics
- Lipschitz L-statistics
- Generalization error of ℓ_2-regularized classification
- Properties of the Gibbs algorithm

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

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

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

Similar argument and result for U (M, 2017)

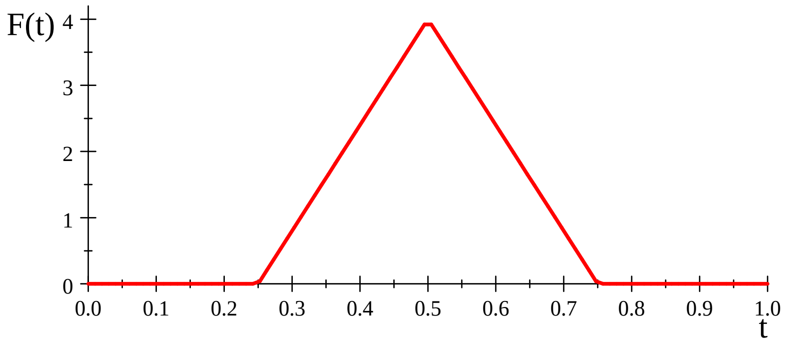
Lipschitz L-statistics

$\mathcal{X} = [a, b]$ 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 $\|F\|_{Lip}$.

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



A "smoothed median"

Lipschitz L-statistics have weak interactions

For $y, y' \in \mathbb{R}$ define

$$\left[\left[y, y'\right]\right] = \left[\text{min}\left\{y, y'\right\}, \text{max}\left\{y, y'\right\}\right].$$

Then (M, M.Pontil, 2018) for $k \neq l$

$$\begin{aligned} D^k_{y,y'} f(x) &\leq \frac{\|F\|_\infty \, diam\,[[y, y']]}{n} \\ D^l_{z,z'} D^k_{y,y'} f(x) &\leq \frac{\|F\|_{Lip} \, diam\,([[z, z']] \cap [[y, y']])}{n^2} \end{aligned}$$

$\Longrightarrow$ f has $\left(\|F\|_\infty (b-a), \|F\|_{Lip} (b-a)\right)$-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!
(M. 2017)

A chain rule

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

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

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.

Gibbs distributions

Ω a mble space of states/models/classifiers with probability measure ρ.
$F : \Omega \to \mathbb{R}$ a "Hamiltonian" (energy or error function),
$\beta > 0$ an "inverse temperature"

$$
\begin{aligned}
\text{Partition function} &: \quad Z_{\beta F} = \int_\Omega e^{-\beta F(\omega)} d\rho(\omega) \\
\text{Free energy} &: \quad A_{\beta F} = \ln Z_{\beta F} \\
\text{Gibbs distribution} &: \quad d\pi_{\beta F}(\omega) = Z_{\beta F}^{-1} e^{-\beta F(\omega)} d\rho(\omega)
\end{aligned}
$$

The Gibbs algorithm

loss of model ω on datum x	:	$\ell(\omega, x)$ where $\ell : \Omega \times \mathcal{X} \longrightarrow [0,1]$
empirical loss on sample $\mathbf{x}$	:	$H(\omega, \mathbf{x}) = \frac{1}{n}\sum_{n=1}^{n} \ell(\omega, x_i)$
Gibbs measure for empirical loss	:	$d\pi_{\beta H(.,\mathbf{x})}$
generic function on Ω	:	$F : \Omega \longrightarrow [0,1]$

By the chain rule

Function on $\mathcal{X}^n$	has weak interactions
$\mathbf{x} \mapsto A_{\beta H(.,\mathbf{x})}$	$\left(\beta, 2\beta^2\right)$
$\mathbf{x} \mapsto \int_\Omega F(\omega)\, d\pi_{\beta H(.,\mathbf{x})}(\omega)$	$\left(2\beta, 6\beta^2\right)$
$\mathbf{x} \mapsto \int_\Omega H(\omega, \mathbf{x})\, d\pi_{\beta H(.,\mathbf{x})}(\omega)$	$\left(2\beta + 1, 6\beta^2 + 4\beta\right)$
$\mathbf{x} \mapsto KL\left(d\pi_{\beta H(.,\mathbf{x})}, d\pi_{\beta F}\right)$	$\left(4\beta^2 + 2\beta, 12\beta^3 + 6\beta^2\right)$

Open problems

- Softer interaction functional for variance estimation
- Weakly dependent variables
- Find more examples of functions with weak interactions

Thank you!

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