Concentration inequalities

Gábor Lugosi

ICREA and Pompeu Fabra University

Barcelona

classical concentration

Law of large numbers: if $X_1, \dots, X_n \in [0,1]$ are independent,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

For example,

$$\operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \leq \frac{1}{4n}$$

Hoeffding's inequality

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]\right| > t\right\} \leq 2e^{-2nt^2}$$

Wassily Hoeffding, born in 1914, Mustamäki, Finland

concentration

The phenomenon extends to general functions of independent random variables.

Various methods:

- martingales (Yurinskii, 1974; Milman and Schechtman, 1986; McDiarmid, 1989,1998);

- information theoretic methods (Alhswede, Gács, and Körner, 1976; Marton 1986, 1996, 1997; Dembo 1997);

- Talagrand's induction method 1996;

- logarithmic Sobolev inequalities (Ledoux 1996, Massart 1998, Boucheron, Lugosi, Massart 1999, 2001).

Efron-Stein inequality (1981)

Let $X_1, \dots, X_n$ be independent random variables taking values in $\mathcal{X}$. Let $f : \mathcal{X}^n \to \mathbb{R}$ and

$$Z = f(X_1, \dots, X_n) \ .$$

If $X'_1, \dots, X'_n$ are independent copies of $X_1, \dots, X_n$, and

$$Z^{(i)} = f(X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_n),$$

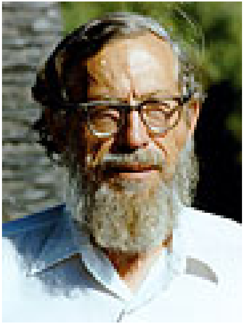
then

$$\mathbf{var}(Z) \leq \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{n}(Z - Z^{(i)})^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}(Z - Z^{(i)})_+^2\right] .$$

Message: Z is concentrated if it doesn't depend too much on any of its variables.

Efron and Stein

Efron and Stein



example: uniform deviations

Let $\mathcal{A}$ be a collection of subsets of $\mathcal{X}$, and let $\mathbf{X_1, \ldots, X_n}$ be $\mathbf{n}$ random points in $\mathcal{X}$, drawn i.i.d.
Let

$$\mathbf{P(A)} = \mathbb{P}\{\mathbf{X_1} \in \mathbf{A}\} \quad \text{and} \quad \mathbf{P_n(A)} = \frac{1}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbb{1}_{\mathbf{X_i} \in \mathbf{A}}$$

If $\mathbf{Z} = \sup_{\mathbf{A} \in \mathcal{A}} |\mathbf{P(A)} - \mathbf{P_n(A)}|$,

$$\mathrm{var}(\mathbf{Z}) \leq \frac{1}{\mathbf{2n}}$$

example: random VC dimension

Let $\mathbf{f(x_1, \ldots, x_n)}$ be the size of the largest subset of $\{\mathbf{x_1, \ldots, x_n}\}$ shattered by $\mathcal{A}$.

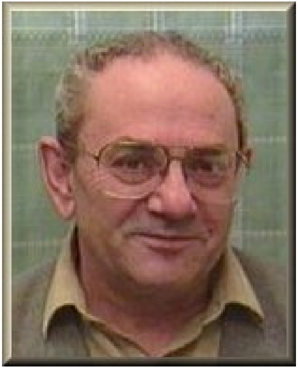
If $\mathbf{Z = f(X_1, \ldots, X_n)}$, then, deterministically,

$$\sum_{i=1}^{n} (Z - Z^{(i)})_+^2 \leq Z$$

so by Efron-Stein,

$$\mathrm{var}(Z) \leq \mathbb{E} Z$$

Vapnik and Chervonenkis



example: conditional Rademacher averages

Let

$$f(x_1, \ldots, x_n) = \mathbb{E}_\epsilon \left[\sup_{A \in \mathcal{A}} \sum_{i=1}^{n} \epsilon_i \mathbb{1}_{x_i \in A} \right]$$

where the ϵ_i are i.i.d. $\mathbb{P}\{\epsilon_i = 1\} = \mathbb{P}\{\epsilon_i = -1\} = 1/2$. If $Z = f(X_1, \ldots, X_n)$, then, again,

$$\sum_{i=1}^{n} (Z - Z^{(i)})_+^2 \leq Z$$

and

$$\mathrm{var}(Z) \leq \mathbb{E}Z$$

Shannon entropy

In many cases much more can be said: exponential inequalities.

If $\mathbf{X}, \mathbf{Y}$ are random variables taking values in a set of size $\mathbf{n}$,

$$\mathbf{H}(\mathbf{X}) = -\sum_{\mathbf{x}} \mathbf{p}(\mathbf{x}) \log \mathbf{p}(\mathbf{x})$$

$$\mathbf{H}(\mathbf{X}|\mathbf{Y}) = \mathbf{H}(\mathbf{X}, \mathbf{Y}) - \mathbf{H}(\mathbf{Y}) = -\sum_{\mathbf{x}} \mathbf{p}(\mathbf{x}, \mathbf{y}) \log \mathbf{p}(\mathbf{x}|\mathbf{y})$$

$$\mathbf{H}(\mathbf{X}) \leq \log \mathbf{n} \quad \text{and} \quad \mathbf{H}(\mathbf{X}|\mathbf{Y}) \leq \mathbf{H}(\mathbf{X})$$

Han's inequality

Te Sun Han

If $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$
and $\mathbf{X}^{(i)} = (\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n)$,
then

$$\sum_{i=1}^{n} \left(\mathbf{H}(\mathbf{X}) - \mathbf{H}(\mathbf{X}^{(i)}) \right) \leq \mathbf{H}(\mathbf{X})$$

an isoperimetric inequality in the hypercube

If $\mathbf{A} \subset \{-1,1\}^{\mathbf{n}}$, let $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_n})$ be uniformly distributed in $\mathbf{A}$.
Then $\mathbf{H(X)} = \log |\mathbf{A}|$ and Han's inequality implies

$$|\partial_{\mathbf{E}}(\mathbf{A})| \geq |\mathbf{A}| \log \frac{2^{\mathbf{n}}}{|\mathbf{A}|}$$

where $\partial_{\mathbf{E}}(\mathbf{A})$ is the set of edges between $\mathbf{A}$ and $\mathbf{A}^{\mathbf{c}}$.

"Edge isoperimetric inequality." Equality for sub-cubes.

a logarithmic Sobolev inequality

If $\mathbf{Z} \geq 0$ is a random variable,

$$\mathbf{Ent}(\mathbf{Z}) = \mathbb{E}[\mathbf{Z}\log\mathbf{Z}] - (\mathbb{E}\mathbf{Z})\log(\mathbb{E}\mathbf{Z})$$

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be uniformly distributed over $\{-1, 1\}^n$. If $\mathbf{f} : \{-1, 1\}^n \to \mathbb{R}$ and $\mathbf{Z} = \mathbf{f}(\mathbf{X})$,

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

Implies Efron-Stein and the isoperimetric inequality.

exponential concentration

If $f : \{-1,1\}^n \to \mathbb{R}$, the log-Sobolev inequality is used with

$$g(x) = e^{\lambda f(x)/2} \quad \text{where} \quad \lambda \in \mathbb{R}$$

If $F(\lambda) = \mathbb{E}\left[e^{\lambda f(X)}\right]$ is the moment generating function of $f(X)$,

$$\begin{aligned} \mathrm{Ent}(g(X)^2) &= \lambda \mathbb{E}\left[f(X)e^{\lambda f(X)}\right] - \mathbb{E}\left[e^{\lambda f(X)}\right] \log \mathbb{E}\left[f(X)e^{\lambda f(X)}\right] \\ &= \lambda F'(\lambda) - F(\lambda)\log F(\lambda). \end{aligned}$$

Differential inequalities are obtained for $F(\lambda)$.
For example, if f such that $\sum_{i=1}^n (Z - Z^{(i)})_+^2 \le v$,

$$\lambda F'(\lambda) - F(\lambda)\log F(\lambda) \le \frac{v\lambda^2}{4} F(\lambda)$$

Solution:

$$F(\lambda) \le e^{\lambda \mathbb{E}Z - \lambda^2 v/4}$$

$$\mathbb{P}\{Z > \mathbb{E}Z + t\} \le e^{-t^2/v}$$

bounded differences inequality

An easy consequence: if $\mathbf{f}$ is such that $|\mathbf{Z} - \mathbf{Z}^{(\mathbf{i})}| \leq \mathbf{1}$,

$$\mathbb{P}\{|\mathbf{Z} - \mathbb{E}\mathbf{Z}| > \mathbf{t}\} \leq \mathbf{2e}^{-\mathbf{2t}^2/\mathbf{n}}$$

"Azuma's inequality."
"McDiarmid's inequality."

variations

The method allows one to derive a variety of exponential concentration inequalities.

For example, if **Z** is either the random VC dimension or a conditional Rademacher average,

$$\mathbb{P}\{\mathbf{Z} > \mathbb{E}\mathbf{Z} + \mathbf{t}\} \leq \mathbf{e}^{-\mathbf{t}^2/(2\mathbb{E}\mathbf{Z}+2\mathbf{t}/3)}$$

and

$$\mathbb{P}\{\mathbf{Z} < \mathbb{E}\mathbf{Z} - \mathbf{t}\} \leq \mathbf{e}^{-\mathbf{t}^2/(2\mathbb{E}\mathbf{Z})}$$

Influences

If $\mathbf{A} \subset \{-1, 1\}^n$ and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is uniform, the influence of the **i**-th variable is

$$\mathbf{I}_i(\mathbf{A}) = \mathbb{P}\left\{\mathbb{1}_{\mathbf{X} \in \mathbf{A}} \neq \mathbb{1}_{\mathbf{X}^{(i)} \in \mathbf{A}}\right\}$$

where $\mathbf{X}^{(i)} = (\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, 1 - \mathbf{X}_i, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n)$.

The total influence is

$$\mathbf{I}(\mathbf{A}) = \sum_{i=1}^{n} \mathbf{I}_i(\mathbf{A}) \quad \left(= \frac{|\partial_E(\mathbf{A})|}{2^{n-1}}\right)$$

Influences: examples

dictatorship: $\mathbf{A = \{x : x_1 = 1\}}$. $\mathbf{I(A) = 1}$.

parity: $\mathbf{A = \{x : \sum_i \mathbb{1}_{x_i=1}}$ is even$\mathbf{\}}$. $\mathbf{I(A) = n}$.

majority: $\mathbf{A = \{x : \sum_i x_i > 0\}}$. $\mathbf{I(A) \approx \sqrt{2n/\pi}}$.

$$\text{by Efron-Stein,} \quad \mathbf{P(A)(1 - P(A)) \leq \frac{I(A)}{4}}$$

so dictatorship has smallest total influence (if $\mathbf{P(A) = 1/2}$).

Falik-Samorodnitsky

If $\mathbf{A} \subset \{-1, 1\}^{\mathbf{n}}$, logarithmic Sobolev inequality implies

$$\mathbf{P(A)(1 - P(A)) \log \frac{4P(A)(1 - P(A))}{\sum_i I_i(A)^2} \leq \frac{I(A)}{4}}$$

Corollary: (Kahn, Kalai, Linial, 1988).

$$\max_{\mathbf{i}} \mathbf{I_i(A)} \geq \frac{\mathbf{P(A)(1 - P(A)) \log n}}{\mathbf{n}}$$

If the influences are equal,

$$\mathbf{I(A) \geq P(A)(1 - P(A)) \log n}$$

Another corollary: (Friedgut, 1998).
If $\mathbf{I(A)} \leq \mathbf{c}$, $\mathbf{A}$ (basically) depends on a bounded number of variables. $\mathbf{A}$ is a "junta."

threshold phenomena

Let $A \subset \{-1,1\}^n$ be a monotone set and let $X = (X_1, \ldots, X_n)$ be such that

$$\mathbb{P}\{X_i = 1\} = p \qquad \mathbb{P}\{X_i = -1\} = 1 - p$$

$$P_p(A) = \sum_{x \in A} p^{\|x\|}(1-p)^{n-\|x\|}$$

is an increasing function of $p \in [0,1]$.

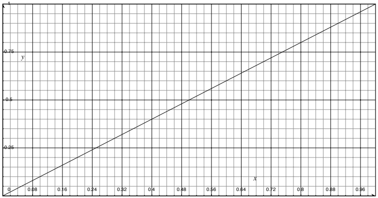
Let p_a be such that $P_{p_a}(A) = a$.

Critical value $= p_{1/2}$

Threshold width: $p_{1-\epsilon} - p_\epsilon$

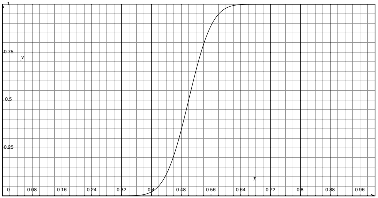
Two (extreme) examples

dictatorship



threshold width $= \mathbf{1 - 2\epsilon}$

majority (with $\mathbf{n = 101}$)



$\leq \sqrt{\mathbf{\log(1/\epsilon)/(2n)}}$

In what cases do we have a quick transition?

Russo's lemma

If $\mathbf{A}$ is monotone,

$$\frac{dP_p(A)}{dp} = I^{(p)}(A)$$

The Kahn, Kalai, Linial result, generalized for $\mathbf{p \neq 1/2}$, implies that
if $\mathbf{A}$ is such that $I_1^{(p)} = I_2^{(p)} = \cdots = I_n^{(p)}$, then

$$p_{1-\epsilon} - p_\epsilon = O\left(\frac{\log\frac{1}{\epsilon}}{\log n}\right)$$

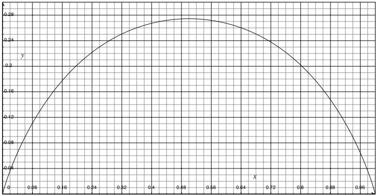
On the other hand, if $p_{3/4} - p_{1/4} \geq c$ then $\mathbf{A}$ is (basically) a junta.

Bobkov's inequality (1997)

Let $\mathbf{f} : \{-1, 1\}^{\mathbf{n}} \to [\mathbf{0}, \mathbf{1}]$ and $\mathbf{X}$ be uniformly distributed.

$$\gamma(\mathbf{p}) = \phi(\mathbf{\Psi}^{-1}(\mathbf{p}))$$

where $\mathbf{\Psi}$ is the standard normal distribution function and $\phi = \mathbf{\Psi}'$.



$$\gamma(\mathbb{E}\mathbf{f}(\mathbf{X})) \leq \mathbb{E}\sqrt{\gamma(\mathbf{f}(\mathbf{X}))^2 + \frac{1}{4}\sum_{\mathbf{i}}\left(\mathbf{f}(\mathbf{X}) - \mathbf{f}(\mathbf{X}^{(\mathbf{i})})^2\right)}$$

Bobkov proves the gaussian isoperimetric theorem based on this inequality.

aplication for influences

If $\mathbf{A} \subset \{-1,1\}^{\mathbf{n}}$, define $\mathbf{f}(\mathbf{x}) = \mathbb{1}_{\mathbf{x}\in\mathbf{A}}$.

Then $\gamma(\mathbf{f}(\mathbf{x})) = \mathbf{0}$ and by Bobkov's inequality,

$$\gamma(\mathbf{P}(\mathbf{A})) \leq \mathbb{E}\sqrt{\mathbf{h}_{\mathbf{A}}(\mathbf{X})}$$

where $\mathbf{h}_{\mathbf{A}}(\mathbf{x}) = \#\{\mathbf{i} : \mathbf{x}^{(\mathbf{i})} \notin \mathbf{A}, \mathbf{x} \in \mathbf{A}\}$.

By Cauchy-Schwarz,

$$\mathbb{E}\sqrt{\mathbf{h}_{\mathbf{A}}(\mathbf{X})} \leq \sqrt{\mathbb{E}\mathbf{h}_{\mathbf{A}}(\mathbf{X})}\sqrt{\mathbf{P}(\partial_{\mathbf{V}}(\mathbf{A}))}$$

where $\partial_{\mathbf{V}}(\mathbf{A}) \subset \mathbf{A}$ is the boundary of $\mathbf{A}$.

But $\mathbb{E}\mathbf{h}_{\mathbf{A}}(\mathbf{X}) = \mathbf{I}(\mathbf{A})$ so if $\mathbf{P}(\mathbf{A}) \approx 1/2$,

$$\mathbf{I}(\mathbf{A})\mathbb{P}(\partial_{\mathbf{V}}(\mathbf{A})) \geq \text{constant}$$