# Lectures on learning theory

Gábor Lugosi

ICREA and Pompeu Fabra University

Barcelona

# what is learning theory?

A mathematical theory to understand the behavior of learning algorithms and assist their design.

# what is learning theory?

A mathematical theory to understand the behavior of learning algorithms and assist their design.

Ingredients:

- Probability;
- (Linear) algebra;
- Optimization;
- Complexity of algorithms;
- High-dimensional geometry;
- Statistics–hypothesis testing, regression, Bayesian methots, etc.
- ...

# learning theory

Statistical learning

supervised–classification, regression, ranking, ...

unsupervised–clustering, density estimation, ...

semi-unsupervised learning

active learning

online learning

# statistical learning

How is it different from "classical" statistics?

- Focus is on prediction rather than inference;
- Distribution-free approach;
- Non-asymptotic results are preferred;
- High-dimensional problems;
- Algorithmic aspects play a central role.

# statistical learning

How is it different from "classical" statistics?

- Focus is on prediction rather than inference;
- Distribution-free approach;
- Non-asymptotic results are preferred;
- High-dimensional problems;
- Algorithmic aspects play a central role.

Here we focus on concentration inequalities.

# a binary classification problem

$(\mathbf{X}, \mathbf{Y})$ is a pair of random variables.

$\mathbf{X} \in \mathbf{X}$ represents the observation

$\mathbf{Y} \in \{-1, 1\}$ is the (binary) label.

A classifier is a function $\mathcal{X} \to \{-1, 1\}$ whose risk is

$$\mathbf{R(g)} = \mathbb{P}\{\mathbf{g(X)} \neq \mathbf{Y}\} \ .$$

# a binary classification problem

$(\mathbf{X}, \mathbf{Y})$ is a pair of random variables.

$\mathbf{X} \in \mathbf{X}$ represents the observation

$\mathbf{Y} \in \{-1, 1\}$ is the (binary) label.

A classifier is a function $\mathcal{X} \to \{-1, 1\}$ whose risk is

$$\mathbf{R(g)} = \mathbb{P}\{\mathbf{g(X)} \neq \mathbf{Y}\} \ .$$

training data: $\mathbf{n}$ i.i.d. observation/label pairs:

$$\mathbf{D_n} = ((\mathbf{X_1}, \mathbf{Y_1}), \ldots, (\mathbf{X_n}, \mathbf{Y_n}))$$

The risk of a data-based classifier $\mathbf{g_n}$ is

$$\mathbf{R(g_n)} = \mathbb{P}\{\mathbf{g_n(X)} \neq \mathbf{Y} | \mathbf{D_n}\} \ .$$

# empirical risk minimization

Given a class $\mathcal{C}$ of classifiers, choose one that minimizes the empirical risk:

$$g_n = \operatorname*{argmin}_{g \in \mathcal{C}} R_n(g) = \operatorname*{argmin}_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{g(X_i) \neq Y_i}$$

# empirical risk minimization

Given a class $\mathcal{C}$ of classifiers, choose one that minimizes the empirical risk:

$$g_n = \operatorname*{argmin}_{g \in \mathcal{C}} R_n(g) = \operatorname*{argmin}_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{g(X_i) \neq Y_i}$$

Fundamental questions:

- How close is $R_n(g)$ to $R(g)$?
- How close is $R(g_n)$ to $\min_{g \in \mathcal{C}} R(g)$?
- How close is $R(g_n)$ to $R_n(g_n)$?

# empirical risk minimization

To understand $|R_n(g) - R(g)|$, we need to study deviations of empirical averages from their means.

For the other two, note that

$$|R(g_n) - R_n(g_n)| \leq \sup_{g \in \mathcal{C}} |R(g) - R_n(g)|$$

and

$$\begin{aligned} R(g_n) - \min_{g \in \mathcal{C}} R(g) &= (R(g_n) - R_n(g_n)) + \left( R_n(g_n) - \min_{g \in \mathcal{C}} R(g) \right) \\ &\leq 2 \sup_{g \in \mathcal{C}} |R(g) - R_n(g)| \end{aligned}$$

# empirical risk minimization

To understand $|R_n(g) - R(g)|$, we need to study deviations of empirical averages from their means.

For the other two, note that

$$|R(g_n) - R_n(g_n)| \leq \sup_{g \in \mathcal{C}} |R(g) - R_n(g)|$$

and

$$\begin{aligned} R(g_n) - \min_{g \in \mathcal{C}} R(g) &= (R(g_n) - R_n(g_n)) + \left( R_n(g_n) - \min_{g \in \mathcal{C}} R(g) \right) \\ &\leq 2 \sup_{g \in \mathcal{C}} |R(g) - R_n(g)| \end{aligned}$$

We need to understand uniform deviations of empirical averages from their means.

## markov's inequality

If $Z \geq 0$, then

$$\mathbb{P}\{Z > t\} \leq \frac{\mathbb{E}Z}{t} .$$

# markov's inequality

If $Z \geq 0$, then

$$\mathbb{P}\{Z > t\} \leq \frac{\mathbb{E}Z}{t} .$$

This implies Chebyshev's inequality: if $Z$ has a finite variance $\mathbf{Var}(Z) = \mathbb{E}(Z - \mathbb{E}Z)^2$, then

$$\mathbb{P}\{|Z - \mathbb{E}Z| > t\} = \mathbb{P}\{(Z - \mathbb{E}Z)^2 > t^2\} \leq \frac{\mathbf{Var}(Z)}{t^2} .$$

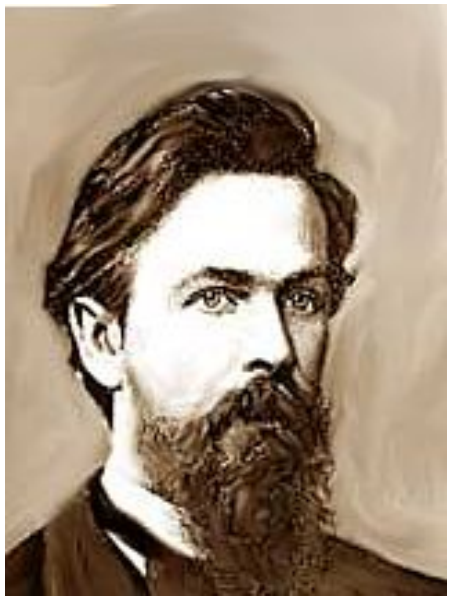
# markov's inequality

If $Z \geq 0$, then

$$\mathbb{P}\{Z > t\} \leq \frac{\mathbb{E}Z}{t} .$$

This implies Chebyshev's inequality: if $Z$ has a finite variance $\mathbf{Var}(Z) = \mathbb{E}(Z - \mathbb{E}Z)^2$, then

$$\mathbb{P}\{|Z - \mathbb{E}Z| > t\} = \mathbb{P}\{(Z - \mathbb{E}Z)^2 > t^2\} \leq \frac{\mathbf{Var}(Z)}{t^2} .$$



Andrey Markov (1856–1922)

# sums of independent random variables

Let $X_1, \ldots, X_n$ be independent real-valued and let $Z = \sum_{i=1}^n X_i$. By independence, $\mathbf{Var}(Z) = \sum_{i=1}^n \mathbf{Var}(X_i)$. If they are identically distributed, $\mathbf{Var}(Z) = n\mathbf{Var}(X_1)$, so

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i - n\mathbb{E}X_1\right| > t\right\} \leq \frac{n\mathbf{Var}(X_1)}{t^2} .$$

Equivalently,

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i - n\mathbb{E}X_1\right| > t\sqrt{n}\right\} \leq \frac{\mathbf{Var}(X_1)}{t^2} .$$

Typical deviations are at most of the order $\sqrt{n}$.

# sums of independent random variables

Let $X_1, \ldots, X_n$ be independent real-valued and let $Z = \sum_{i=1}^n X_i$. By independence, $\mathbf{Var}(Z) = \sum_{i=1}^n \mathbf{Var}(X_i)$. If they are identically distributed, $\mathbf{Var}(Z) = n\mathbf{Var}(X_1)$, so

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i - n\mathbb{E}X_1\right| > t\right\} \leq \frac{n\mathbf{Var}(X_1)}{t^2} .$$

Equivalently,

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i - n\mathbb{E}X_1\right| > t\sqrt{n}\right\} \leq \frac{\mathbf{Var}(X_1)}{t^2} .$$

Typical deviations are at most of the order $\sqrt{n}$.

Pafnuty Chebyshev (1821–1894)

# chernoff bounds

By the central limit theorem,

$$\lim_{n\to\infty} \mathbb{P}\left\{\sum_{i=1}^{n} X_i - n\mathbb{E}X_1 > t\sqrt{n}\right\} = 1 - \Psi(t/\sqrt{\mathrm{Var}(X_1)})$$

$$\leq e^{-t^2/(2\mathrm{Var}(X_1))}$$

so we expect an exponential decrease in $t^2/\mathrm{Var}(X_1)$.

# chernoff bounds

By the central limit theorem,

$$\lim_{n\to\infty} \mathbb{P}\left\{\sum_{i=1}^{n} X_i - n\mathbb{E}X_1 > t\sqrt{n}\right\} = 1 - \Psi(t/\sqrt{\mathrm{Var}(X_1)})$$

$$\leq e^{-t^2/(2\mathrm{Var}(X_1))}$$

so we expect an exponential decrease in $t^2/\mathrm{Var}(X_1)$.
Trick: use Markov's inequality in a more clever way: if $\lambda > 0$,

$$\mathbb{P}\{Z - \mathbb{E}Z > t\} = \mathbb{P}\left\{e^{\lambda(Z-\mathbb{E}Z)} > e^{\lambda t}\right\} \leq \frac{\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}}{e^{\lambda t}}$$

Now derive bounds for the moment generating function $\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}$ and optimize $\lambda$.

# chernoff bounds

If $Z = \sum_{i=1}^{n} X_i$ is a sum of independent random variables,

$$\mathbb{E}e^{\lambda Z} = \mathbb{E}\prod_{i=1}^{n} e^{\lambda X_i} = \prod_{i=1}^{n} \mathbb{E}e^{\lambda X_i}$$

by independence. Now it suffices to find bounds for $\mathbb{E}e^{\lambda X_i}$.

# chernoff bounds

If $Z = \sum_{i=1}^{n} X_i$ is a sum of independent random variables,

$$\mathbb{E}e^{\lambda Z} = \mathbb{E}\prod_{i=1}^{n} e^{\lambda X_i} = \prod_{i=1}^{n} \mathbb{E}e^{\lambda X_i}$$

by independence. Now it suffices to find bounds for $\mathbb{E}e^{\lambda X_i}$.

Serguei Bernstein (1880-1968)

Herman Chernoff (1923–)

# hoeffding's inequality

If $X_1, \ldots, X_n \in [0,1]$, then

$$\mathbb{E}e^{\lambda(X_i - \mathbb{E}X_i)} \leq e^{\lambda^2/8} .$$

# hoeffding's inequality

If $X_1, \ldots, X_n \in [0,1]$, then

$$\mathbb{E}e^{\lambda(X_i - \mathbb{E}X_i)} \leq e^{\lambda^2/8} .$$

We obtain

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

Wassily Hoeffding (1914–1991)

# bernstein's inequality

Hoeffding's inequality is distribution free. It does not take variance information into account.

Bernstein's inequality is an often useful variant:

Let $X_1, \ldots, X_n$ be independent such that $X_i \leq 1$. Let $v = \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right]$. Then

$$\mathbb{P}\left\{\sum_{i=1}^{n} (X_i - \mathbb{E}X_i) \geq t\right\} \leq \exp\left(-\frac{t^2}{2(v + t/3)}\right) .$$

# a maximal inequality

Suppose $Y_1, \ldots, Y_N$ are sub-Gaussian in the sense that

$$\mathbb{E} e^{\lambda Y_i} \leq e^{\lambda^2 \sigma^2 / 2} .$$

Then

$$\mathbb{E} \max_{i=1,\ldots,N} Y_i \leq \sigma \sqrt{2 \log N} .$$

# a maximal inequality

Suppose $Y_1, \ldots, Y_N$ are sub-Gaussian in the sense that

$$\mathbb{E} e^{\lambda Y_i} \leq e^{\lambda^2 \sigma^2 / 2} \ .$$

Then

$$\mathbb{E} \max_{i=1,\ldots,N} Y_i \leq \sigma \sqrt{2 \log N} \ .$$

Proof:

$$e^{\lambda \mathbb{E} \max_{i=1,\ldots,N} Y_i} \leq \mathbb{E} e^{\lambda \max_{i=1,\ldots,N} Y_i} \leq \sum_{i=1}^{N} \mathbb{E} e^{\lambda Y_i} \leq N e^{\lambda^2 \sigma^2 / 2}$$

Take logarithms, and optimize in $\lambda$.

# uniform deviations–finite classes

Let $A_1, \ldots, A_N \subset \mathcal{X}$ and let $X_1, \ldots, X_n$ be i.i.d. random points in $\mathcal{X}$. Let

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

By Hoeffding's inequality, for each $A$,

$$\begin{aligned} \mathbb{E} e^{\lambda(P(A) - P_n(A))} &= \mathbb{E} e^{(\lambda/n) \sum_{i=1}^{n} (P(A) - \mathbb{1}_{X_i \in A})} \\ &= \prod_{i=1}^{n} \mathbb{E} e^{(\lambda/n)(P(A) - \mathbb{1}_{X_i \in A})} \le e^{\lambda^2/(8n)} . \end{aligned}$$

By the maximal inequality,

$$\mathbb{E} \max_{j=1,\ldots,N} (P(A_j) - P_n(A_j)) \le \sqrt{\frac{\log N}{2n}} .$$

# johnson-lindenstrauss

Suppose $\mathbf{A} = \{\mathbf{a_1}, \ldots, \mathbf{a_n}\} \subset \mathbb{R}^{\mathbf{D}}$ is a finite set, $\mathbf{D}$ is large.

We would like to embed $\mathbf{A}$ in $\mathbb{R}^{\mathbf{d}}$ where $\mathbf{d} \ll \mathbf{D}$.

# johnson-lindenstrauss

Suppose $\mathbf{A = \{a_1, \ldots, a_n\}} \subset \mathbb{R}^{\mathbf{D}}$ is a finite set, $\mathbf{D}$ is large.

We would like to embed $\mathbf{A}$ in $\mathbb{R}^{\mathbf{d}}$ where $\mathbf{d} \ll \mathbf{D}$.

Is this possible? In what sense?

# johnson-lindenstrauss

Suppose $\mathbf{A = \{a_1, \ldots, a_n\} \subset \mathbb{R}^D}$ is a finite set, $\mathbf{D}$ is large.

We would like to embed $\mathbf{A}$ in $\mathbb{R}^{\mathbf{d}}$ where $\mathbf{d \ll D}$.

Is this possible? In what sense?

Given $\varepsilon > \mathbf{0}$, a function $\mathbf{f : \mathbb{R}^D \to \mathbb{R}^d}$ is an $\varepsilon$-isometry if for all $\mathbf{a, a' \in A}$,

$$\mathbf{(1-\varepsilon)\left\|a-a'\right\|^2 \leq \left\|f(a)-f(a')\right\|^2 \leq (1+\varepsilon)\left\|a-a'\right\|^2} \ .$$

# johnson-lindenstrauss

Suppose $\mathbf{A = \{a_1, \ldots, a_n\} \subset \mathbb{R}^D}$ is a finite set, $\mathbf{D}$ is large.

We would like to embed $\mathbf{A}$ in $\mathbb{R}^{\mathbf{d}}$ where $\mathbf{d \ll D}$.

Is this possible? In what sense?

Given $\varepsilon > \mathbf{0}$, a function $\mathbf{f : \mathbb{R}^D \to \mathbb{R}^d}$ is an $\varepsilon$-isometry if for all $\mathbf{a, a' \in A}$,

$$\mathbf{(1-\varepsilon)\left\|a-a'\right\|^2 \leq \left\|f(a)-f(a')\right\|^2 \leq (1+\varepsilon)\left\|a-a'\right\|^2} \ .$$

Johnson-Lindenstrauss lemma: If $\mathbf{d \geq (c/\varepsilon^2)\log n}$, then there exists an $\varepsilon$-isometry $\mathbf{f : \mathbb{R}^D \to \mathbb{R}^d}$.

# johnson-lindenstrauss

Suppose $\mathbf{A = \{a_1, \ldots, a_n\} \subset \mathbb{R}^D}$ is a finite set, $\mathbf{D}$ is large.

We would like to embed $\mathbf{A}$ in $\mathbb{R}^{\mathbf{d}}$ where $\mathbf{d \ll D}$.

Is this possible? In what sense?

Given $\varepsilon > \mathbf{0}$, a function $\mathbf{f : \mathbb{R}^D \to \mathbb{R}^d}$ is an $\varepsilon$-isometry if for all $\mathbf{a, a' \in A}$,

$$\mathbf{(1-\varepsilon)\left\|a-a'\right\|^2 \leq \left\|f(a)-f(a')\right\|^2 \leq (1+\varepsilon)\left\|a-a'\right\|^2} \ .$$

Johnson-Lindenstrauss lemma: If $\mathbf{d \geq (c/\varepsilon^2)\log n}$, then there exists an $\varepsilon$-isometry $\mathbf{f : \mathbb{R}^D \to \mathbb{R}^d}$.

Independent of $\mathbf{D}$!

# random projections

We take **f** to be linear. How? At random!

# random projections

We take $\mathbf{f}$ to be linear. How? At random!

Let $\mathbf{f} = (\mathbf{W_{i,j}})_{\mathbf{d} \times \mathbf{D}}$ with

$$\mathbf{W_{i,j}} = \frac{\mathbf{1}}{\sqrt{\mathbf{d}}} \mathbf{X_{i,j}}$$

where the $\mathbf{X_{i,j}}$ are independent standard normal.

# random projections

We take $\mathbf{f}$ to be linear. How? At random!

Let $\mathbf{f} = (\mathbf{W_{i,j}})_{\mathbf{d \times D}}$ with

$$\mathbf{W_{i,j}} = \frac{\mathbf{1}}{\sqrt{\mathbf{d}}} \mathbf{X_{i,j}}$$

where the $\mathbf{X_{i,j}}$ are independent standard normal.

For any $\mathbf{a} = (\alpha_1, \ldots, \alpha_{\mathbf{D}}) \in \mathbb{R}^{\mathbf{D}}$,

$$\mathbb{E}\|\mathbf{f(a)}\|^2 = \frac{\mathbf{1}}{\mathbf{d}} \sum_{\mathbf{i=1}}^{\mathbf{d}} \sum_{\mathbf{j=1}}^{\mathbf{D}} \alpha_{\mathbf{j}}^2 \mathbb{E}\mathbf{X_{i,j}^2} = \|\mathbf{a}\|^2 \ .$$

The expected squared distances are preserved!

# random projections

We take $\mathbf{f}$ to be linear. How? At random!

Let $\mathbf{f} = (\mathbf{W_{i,j}})_{\mathbf{d}\times\mathbf{D}}$ with

$$\mathbf{W_{i,j}} = \frac{\mathbf{1}}{\sqrt{\mathbf{d}}}\mathbf{X_{i,j}}$$

where the $\mathbf{X_{i,j}}$ are independent standard normal.

For any $\mathbf{a} = (\alpha_1, \ldots, \alpha_D) \in \mathbb{R}^D$,

$$\mathbb{E}\|\mathbf{f(a)}\|^2 = \frac{\mathbf{1}}{\mathbf{d}} \sum_{\mathbf{i=1}}^{\mathbf{d}} \sum_{\mathbf{j=1}}^{\mathbf{D}} \alpha_j^2 \mathbb{E}\mathbf{X_{i,j}^2} = \|\mathbf{a}\|^2 \ .$$

The expected squared distances are preserved!
$\|\mathbf{f(a)}\|^2/\|\mathbf{a}\|^2$ is a weighted sum of squared normals.

# random projections

Let $\mathbf{b} = \mathbf{a_i} - \mathbf{a_j}$ for some $\mathbf{a_i}, \mathbf{a_j} \in \mathbf{A}$. Then

$$\mathbb{P}\left\{\exists \mathbf{b} : \left|\frac{\|\mathbf{f(b)}\|^2}{\|\mathbf{b}\|^2} - \mathbf{1}\right| > \sqrt{\frac{\mathbf{8\log(n/\sqrt{\delta})}}{\mathbf{d}}} + \frac{\mathbf{8\log(n/\sqrt{\delta})}}{\mathbf{d}}\right\}$$

$$\leq \binom{\mathbf{n}}{\mathbf{2}}\mathbb{P}\left\{\left|\frac{\|\mathbf{f(b)}\|^2}{\|\mathbf{b}\|^2} - \mathbf{1}\right| > \sqrt{\frac{\mathbf{8\log(n/\sqrt{\delta})}}{\mathbf{d}}} + \frac{\mathbf{8\log(n/\sqrt{\delta})}}{\mathbf{d}}\right\}$$

$$\leq \delta \quad \text{(by a Bernstein-type inequality)}\ .$$

If $\mathbf{d} \geq \mathbf{(c/\varepsilon^2)\log(n/\sqrt{\delta})}$, then

$$\sqrt{\frac{\mathbf{8\log(n/\sqrt{\delta})}}{\mathbf{d}}} + \frac{\mathbf{8\log(n/\sqrt{\delta})}}{\mathbf{d}} \leq \varepsilon$$

and $\mathbf{f}$ is an $\varepsilon$-isometry with probability $\geq \mathbf{1} - \delta$.

# martingale representation

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

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

Denote $\mathbb{E}_i[\cdot] = \mathbb{E}[\cdot|X_1, \ldots, X_i]$. Thus, $\mathbb{E}_0 Z = \mathbb{E}Z$ and $\mathbb{E}_n Z = Z$.

# martingale representation

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

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

Denote $\mathbb{E}_i[\cdot] = \mathbb{E}[\cdot | X_1, \ldots, X_i]$. Thus, $\mathbb{E}_0 Z = \mathbb{E}Z$ and $\mathbb{E}_n Z = Z$.

Writing

$$\Delta_i = \mathbb{E}_i Z - \mathbb{E}_{i-1} Z \ ,$$

we have

$$Z - \mathbb{E}Z = \sum_{i=1}^{n} \Delta_i$$

This is the Doob martingale representation of $Z$.

# martingale representation

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

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

Denote $\mathbb{E}_i[\cdot] = \mathbb{E}[\cdot | X_1, \ldots, X_i]$. Thus, $\mathbb{E}_0 Z = \mathbb{E}Z$ and $\mathbb{E}_n Z = Z$.

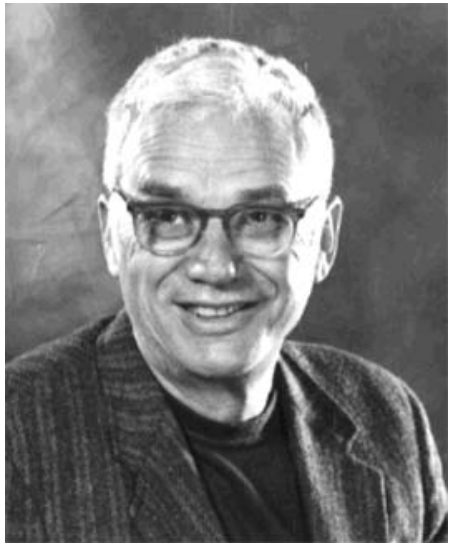
Writing

$$\Delta_i = \mathbb{E}_i Z - \mathbb{E}_{i-1} Z \ ,$$

we have

$$Z - \mathbb{E}Z = \sum_{i=1}^{n} \Delta_i$$

This is the Doob martingale representation of $Z$.



Joseph Leo Doob (1910–2004)

# martingale representation: the variance

$$\mathbf{Var}\,(Z) = \mathbb{E}\left[\left(\sum_{i=1}^{n} \Delta_i\right)^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\Delta_i^2\right] + 2\sum_{j>i} \mathbb{E}\Delta_i\Delta_j \ .$$

Now if $j > i$, $\mathbb{E}_i\Delta_j = 0$, so

$$\mathbb{E}_i\Delta_j\Delta_i = \Delta_i\mathbb{E}_i\Delta_j = 0 \ ,$$

We obtain

$$\mathbf{Var}\,(Z) = \mathbb{E}\left[\left(\sum_{i=1}^{n} \Delta_i\right)^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\Delta_i^2\right] \ .$$

# martingale representation: the variance

$$\mathbf{Var}\,(Z) = \mathbb{E}\left[\left(\sum_{i=1}^{n} \Delta_i\right)^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\Delta_i^2\right] + 2\sum_{j>i} \mathbb{E}\Delta_i\Delta_j \ .$$

Now if $j > i$, $\mathbb{E}_i\Delta_j = 0$, so

$$\mathbb{E}_i\Delta_j\Delta_i = \Delta_i\mathbb{E}_i\Delta_j = 0 \ ,$$

We obtain

$$\mathbf{Var}\,(Z) = \mathbb{E}\left[\left(\sum_{i=1}^{n} \Delta_i\right)^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\Delta_i^2\right] \ .$$

From this, using independence, it is easy derive the Efron-Stein inequality.

# efron-stein inequality (1981)

Let $X_1, \ldots, X_n$ be independent random variables taking values in $\mathcal{X}$. Let $f : \mathcal{X}^n \to \mathbb{R}$ and $Z = f(X_1, \ldots, X_n)$.
Then

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

where $\mathbb{E}^{(i)} Z$ is expectation with respect to the $i$-th variable $X_i$ only.

# efron-stein inequality (1981)

Let $X_1, \ldots, X_n$ be independent random variables taking values in $\mathcal{X}$. Let $f : \mathcal{X}^n \to \mathbb{R}$ and $Z = f(X_1, \ldots, X_n)$.
Then

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

where $\mathbb{E}^{(i)} Z$ is expectation with respect to the $i$-th variable $X_i$ only.

We obtain more useful forms by using that

$$\mathrm{Var}(X) = \frac{1}{2} \mathbb{E}(X - X')^2 \quad \text{and} \quad \mathrm{Var}(X) \leq \mathbb{E}(X - a)^2$$

for any constant $a$.

# efron-stein inequality (1981)

If $X_1', \ldots, X_n'$ are independent copies of $X_1, \ldots, X_n$, and

$$Z_i' = f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n),$$

then

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

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

# efron-stein inequality (1981)

If $X_1', \ldots, X_n'$ are independent copies of $X_1, \ldots, X_n$, and

$$Z_i' = f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n),$$

then

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

$Z$ is concentrated if it doesn't depend too much on any of its variables.
If $Z = \sum_{i=1}^{n} X_i$ then we have an equality. Sums are the "least concentrated" of all functions!

# efron-stein inequality (1981)

If for some arbitrary functions $f_i$

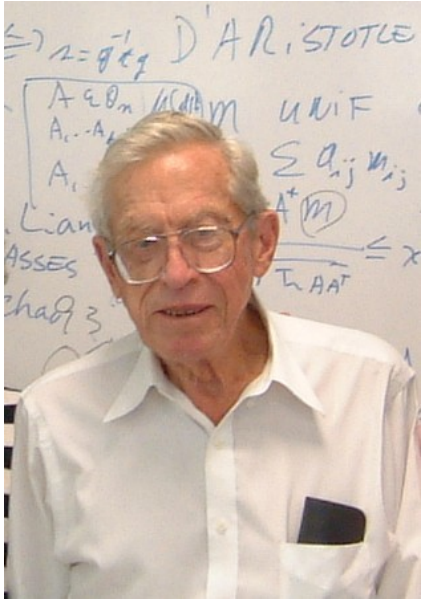
$$Z_i = f_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \ ,$$

then

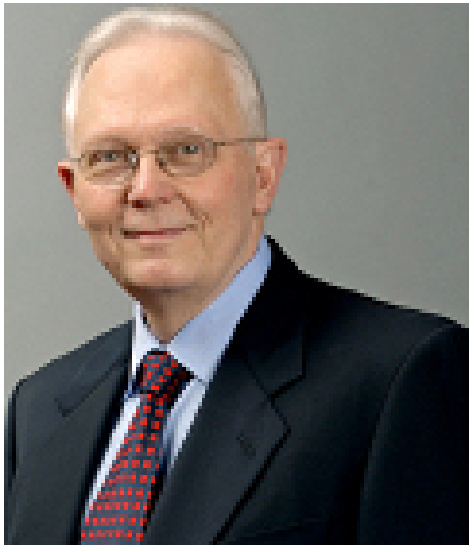
$$\mathbf{Var}(Z) \leq \mathbb{E}\left[\sum_{i=1}^{n}(Z - Z_i)^2\right]$$

# efron, stein, and steele

Bradley Efron



Charles Stein



Mike Steele

# example: uniform deviations

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

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

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

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

# example: uniform deviations

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

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

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

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

regardless of the distribution and the richness of $\mathcal{A}$.

# example: kernel density estimation

Let $\mathbf{X_1, \ldots, X_n}$ be i.i.d. real samples drawn according to some density $\phi$. The kernel density estimate is

$$\phi_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $h > 0$, and $K$ is a nonnegative "kernel" $\int K = 1$. The $L_1$ error is

$$Z = f(X_1, \ldots, X_n) = \int |\phi(x) - \phi_n(x)| dx \; .$$

# example: kernel density estimation

Let $X_1, \ldots, X_n$ be i.i.d. real samples drawn according to some density $\phi$. The kernel density estimate is

$$\phi_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $h > 0$, and $K$ is a nonnegative "kernel" $\int K = 1$. The $L_1$ error is

$$Z = f(X_1, \ldots, X_n) = \int |\phi(x) - \phi_n(x)| dx \; .$$

It is easy to see that

$$\begin{aligned} &|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \\ &\leq \frac{1}{nh} \int \left| K\left(\frac{x - x_i}{h}\right) - K\left(\frac{x - x_i'}{h}\right) \right| dx \leq \frac{2}{n} \; , \end{aligned}$$

so we get $\mathbf{Var}(Z) \leq \frac{2}{n}$ .

# bounding the expectation

Let $\mathbf{P'_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X'_i \in A}}$ and let $\mathbb{E}'$ denote expectation only with respect to $\mathbf{X'_1, \ldots, X'_n}$.

$$\mathbb{E} \sup_{A \in \mathcal{A}} |\mathbf{P_n(A) - P(A)}| = \mathbb{E} \sup_{A \in \mathcal{A}} |\mathbb{E}'[\mathbf{P_n(A) - P'_n(A)}]|$$

$$\leq \mathbb{E} \sup_{A \in \mathcal{A}} |\mathbf{P_n(A) - P'_n(A)}| = \frac{1}{n} \mathbb{E} \sup_{A \in \mathcal{A}} \left| \sum_{i=1}^{n} (\mathbb{1}_{X_i \in A} - \mathbb{1}_{X'_i \in A}) \right|$$

# bounding the expectation

Let $\mathbf{P'_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X'_i \in A}}$ and let $\mathbb{E}'$ denote expectation only with respect to $\mathbf{X'_1, \ldots, X'_n}$.

$$\mathbb{E} \sup_{A \in \mathcal{A}} |P_n(A) - P(A)| = \mathbb{E} \sup_{A \in \mathcal{A}} |\mathbb{E}'[P_n(A) - P'_n(A)]|$$

$$\leq \mathbb{E} \sup_{A \in \mathcal{A}} |P_n(A) - P'_n(A)| = \frac{1}{n} \mathbb{E} \sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n (\mathbb{1}_{X_i \in A} - \mathbb{1}_{X'_i \in A}) \right|$$

Second symmetrization: if $\varepsilon_1, \ldots, \varepsilon_n$ are independent Rademacher variables, then

$$= \frac{1}{n} \mathbb{E} \sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \varepsilon_i (\mathbb{1}_{X_i \in A} - \mathbb{1}_{X'_i \in A}) \right| \leq \frac{2}{n} \mathbb{E} \sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \varepsilon_i \mathbb{1}_{X_i \in A} \right|$$

# conditional rademacher average

If

$$\mathbf{R_n} = \mathbb{E}_{\varepsilon} \sup_{\mathbf{A} \in \mathcal{A}} \left| \sum_{\mathbf{i}=1}^{\mathbf{n}} \varepsilon_{\mathbf{i}} \mathbb{1}_{\mathbf{X_i} \in \mathbf{A}} \right|$$

then

$$\mathbb{E} \sup_{\mathbf{A} \in \mathcal{A}} |\mathbf{P_n(A)} - \mathbf{P(A)}| \leq \frac{\mathbf{2}}{\mathbf{n}} \mathbb{E} \mathbf{R_n} \ .$$

# conditional rademacher average

If

$$\mathbf{R_n} = \mathbb{E}_{\varepsilon} \sup_{\mathbf{A} \in \mathcal{A}} \left| \sum_{\mathbf{i}=1}^{\mathbf{n}} \varepsilon_{\mathbf{i}} \mathbb{1}_{\mathbf{X_i} \in \mathbf{A}} \right|$$

then

$$\mathbb{E} \sup_{\mathbf{A} \in \mathcal{A}} |\mathbf{P_n(A)} - \mathbf{P(A)}| \leq \frac{\mathbf{2}}{\mathbf{n}} \mathbb{E} \mathbf{R_n} \ .$$

$\mathbf{R_n}$ is a data-dependent quantity!

# concentration of conditional rademacher average

Define

$$\mathbf{R_n^{(i)}} = \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{A}\in\mathcal{A}} \left| \sum_{\mathbf{j}\neq\mathbf{i}} \boldsymbol{\varepsilon}_\mathbf{j} \mathbb{1}_{\mathbf{X_j}\in\mathbf{A}} \right|$$

One can show easily that

$$\mathbf{0} \leq \mathbf{R_n} - \mathbf{R_n^{(i)}} \leq \mathbf{1} \quad \text{and} \quad \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} (\mathbf{R_n} - \mathbf{R_n^{(i)}}) \leq \mathbf{R_n} \ .$$

By the Efron-Stein inequality,

$$\mathbf{Var(R_n)} \leq \mathbb{E} \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} (\mathbf{R_n} - \mathbf{R_n^{(i)}})^2 \leq \mathbb{E}\mathbf{R_n} \ .$$

# concentration of conditional rademacher average

Define

$$R_n^{(i)} = \mathbb{E}_\varepsilon \sup_{A \in \mathcal{A}} \left| \sum_{j \neq i} \varepsilon_j \mathbb{1}_{X_j \in A} \right|$$

One can show easily that

$$0 \leq R_n - R_n^{(i)} \leq 1 \quad \text{and} \quad \sum_{i=1}^{n} (R_n - R_n^{(i)}) \leq R_n \ .$$

By the Efron-Stein inequality,

$$\mathbf{Var}(R_n) \leq \mathbb{E} \sum_{i=1}^{n} (R_n - R_n^{(i)})^2 \leq \mathbb{E} R_n \ .$$

Standard deviation is at most $\sqrt{\mathbb{E} R_n}$!

# concentration of conditional rademacher average

Define

$$\mathbf{R_n^{(i)}} = \mathbb{E}_\varepsilon \sup_{\mathbf{A}\in\mathcal{A}} \left| \sum_{\mathbf{j}\neq\mathbf{i}} \varepsilon_\mathbf{j} \mathbb{1}_{\mathbf{X_j}\in\mathbf{A}} \right|$$

One can show easily that

$$\mathbf{0 \le R_n - R_n^{(i)} \le 1} \quad \text{and} \quad \mathbf{\sum_{i=1}^{n} (R_n - R_n^{(i)}) \le R_n} \ .$$

By the Efron-Stein inequality,

$$\mathbf{Var(R_n)} \le \mathbb{E} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{(R_n - R_n^{(i)})^2} \le \mathbb{E}\mathbf{R_n} \ .$$

Standard deviation is at most $\sqrt{\mathbb{E}\mathbf{R_n}}$!

Such functions are called self-bounding.

# bounding the conditional rademacher average

If $\mathbf{S}(\mathbf{X}_1^n, \mathcal{A})$ is the number of different sets of form

$$\{\mathbf{X}_1, \ldots, \mathbf{X}_n\} \cap \mathbf{A} : \mathbf{A} \in \mathcal{A}$$

then $\mathbf{R_n}$ is the maximum of $\mathbf{S}(\mathbf{X}_1^n, \mathcal{A})$ sub-Gaussian random variables. By the maximal inequality,

$$\frac{1}{2}\mathbf{R_n} \leq \sqrt{\frac{\log \mathbf{S}(\mathbf{X}_1^n, \mathcal{A})}{2n}} \; .$$

# bounding the conditional rademacher average

If $\mathbf{S}(\mathbf{X}_1^n, \mathcal{A})$ is the number of different sets of form

$$\{\mathbf{X}_1, \ldots, \mathbf{X}_n\} \cap \mathbf{A} : \mathbf{A} \in \mathcal{A}$$

then $\mathbf{R_n}$ is the maximum of $\mathbf{S}(\mathbf{X}_1^n, \mathcal{A})$ sub-Gaussian random variables. By the maximal inequality,

$$\frac{1}{2}\mathbf{R_n} \leq \sqrt{\frac{\log \mathbf{S}(\mathbf{X}_1^n, \mathcal{A})}{2n}} \; .$$

In particular,

$$\mathbb{E} \sup_{\mathbf{A} \in \mathcal{A}} |\mathbf{P_n}(\mathbf{A}) - \mathbf{P}(\mathbf{A})| \leq 2\mathbb{E}\sqrt{\frac{\log \mathbf{S}(\mathbf{X}_1^n, \mathcal{A})}{2n}} \; .$$

# random VC dimension

Let $\mathbf{V} = \mathbf{V}(x_1^n, \mathcal{A})$ be the size of the largest subset of $\{x_1, \ldots, x_n\}$ shattered by $\mathcal{A}$.
By Sauer's lemma,

$$\log S(X_1^n, \mathcal{A}) \leq V(X_1^n, \mathcal{A}) \log(n+1) \ .$$

# random VC dimension

Let $V = V(x_1^n, \mathcal{A})$ be the size of the largest subset of $\{x_1, \ldots, x_n\}$ shattered by $\mathcal{A}$.
By Sauer's lemma,

$$\log S(X_1^n, \mathcal{A}) \leq V(X_1^n, \mathcal{A}) \log(n+1) \ .$$

$V$ is also self-bounding:

$$\sum_{i=1}^{n} (V - V^{(i)})^2 \leq V$$

so by Efron-Stein,

$$\mathbf{Var}(V) \leq \mathbb{E}V$$

# vapnik and chervonenkis

Vladimir Vapnik

Alexey Chervonenkis

# beyond the variance

$X_1, \ldots, X_n$ are independent random variables taking values in some set $\mathcal{X}$. Let $f : \mathcal{X}^n \to \mathbb{R}$ and $Z = f(X_1, \ldots, X_n)$. Recall the Doob martingale representation:

$$Z - \mathbb{E}Z = \sum_{i=1}^{n} \Delta_i \quad \text{where} \quad \Delta_i = \mathbb{E}_i Z - \mathbb{E}_{i-1} Z \ ,$$

with $\mathbb{E}_i[\cdot] = \mathbb{E}[\cdot | X_1, \ldots, X_i]$.

To get exponential inequalities, we bound the moment generating function $\mathbb{E}e^{\lambda(Z - \mathbb{E}Z)}$.

# azuma's inequality

Suppose that the martingale differences are bounded: $|\Delta_i| \leq c_i$. Then

$$\begin{aligned}
\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} &= \mathbb{E}e^{\lambda\left(\sum_{i=1}^n \Delta_i\right)} = \mathbb{E}\mathbb{E}_n e^{\lambda\left(\sum_{i=1}^{n-1} \Delta_i\right)+\lambda\Delta_n} \\
&= \mathbb{E}e^{\lambda\left(\sum_{i=1}^{n-1} \Delta_i\right)}\mathbb{E}_n e^{\lambda\Delta_n} \\
&\leq \mathbb{E}e^{\lambda\left(\sum_{i=1}^{n-1} \Delta_i\right)} e^{\lambda^2 c_n^2/2} \text{ (by Hoeffding)} \\
&\cdots \\
&\leq e^{\lambda^2\left(\sum_{i=1}^n c_i^2\right)/2} .
\end{aligned}$$

This is the Azuma-Hoeffding inequality for sums of bounded martingale differences.

# bounded differences inequality

If $Z = f(X_1, \ldots, X_n)$ and $f$ is such that

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$$

then the martingale differences are bounded.

# bounded differences inequality

If $Z = f(X_1, \ldots, X_n)$ and $f$ is such that

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$$

then the martingale differences are bounded.

Bounded differences inequality: if $X_1, \ldots, X_n$ are independent, then

$$\mathbb{P}\{|Z - \mathbb{E}Z| > t\} \leq 2e^{-2t^2/\sum_{i=1}^n c_i^2} .$$

# bounded differences inequality

If $Z = f(X_1, \ldots, X_n)$ and $f$ is such that

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$$

then the martingale differences are bounded.

Bounded differences inequality: if $X_1, \ldots, X_n$ are independent, then

$$\mathbb{P}\{|Z - \mathbb{E}Z| > t\} \leq 2e^{-2t^2/\sum_{i=1}^n c_i^2} .$$

McDiarmid's inequality.

Colin McDiarmid

# hoeffding in a hilbert space

Let $\mathbf{X_1, \ldots, X_n}$ be independent zero-mean random variables in a separable Hilbert space such that $\mathbf{\|X_i\| \leq c/2}$ and denote $\mathbf{v = nc^2/4}$. Then, for all $\mathbf{t \geq \sqrt{v}}$,

$$\mathbb{P}\left\{\left\|\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}\right\| > \mathbf{t}\right\} \leq \mathbf{e^{-(t-\sqrt{v})^2/(2v)}} \ .$$

# hoeffding in a hilbert space

Let $\mathbf{X_1, \ldots, X_n}$ be independent zero-mean random variables in a separable Hilbert space such that $\|\mathbf{X_i}\| \leq \mathbf{c/2}$ and denote $\mathbf{v = nc^2/4}$. Then, for all $\mathbf{t \geq \sqrt{v}}$,

$$\mathbb{P}\left\{\left\|\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}\right\| > \mathbf{t}\right\} \leq \mathbf{e}^{-(\mathbf{t}-\sqrt{\mathbf{v}})^2/(\mathbf{2v})} \ .$$

Proof: By the triangle inequality, $\left\|\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}\right\|$ has the bounded differences property with constants $\mathbf{c}$, so

$$\mathbb{P}\left\{\left\|\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}\right\| > \mathbf{t}\right\} = \mathbb{P}\left\{\left\|\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}\right\| - \mathbb{E}\left\|\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}\right\| > \mathbf{t} - \mathbb{E}\left\|\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}\right\|\right\}$$
$$\leq \mathbf{exp}\left(-\frac{\left(\mathbf{t} - \mathbb{E}\left\|\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}\right\|\right)^2}{\mathbf{2v}}\right) \ .$$

Also,

$$\mathbb{E}\left\|\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}\right\| \leq \sqrt{\mathbb{E}\left\|\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}\right\|^2} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbb{E}\left\|\mathbf{X_i}\right\|^2} \leq \sqrt{\mathbf{v}} \ .$$

# bounded differences inequality

- Easy to use.
- Distribution free.
- Often close to optimal (e.g., $L_1$ error of kernel density estimate).
- Does not exploit "variance information."
- Often too rigid.
- Other methods are necessary.

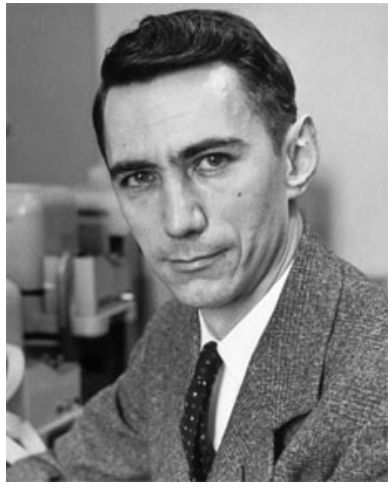
# shannon entropy

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

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

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

$\mathbf{H(X)} \leq \log \mathbf{N}$ and $\mathbf{H(X|Y)} \leq \mathbf{H(X)}$



Claude Shannon
(1916–2001)

# han's inequality

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

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

Te Sun Han

Proof:

$$\begin{aligned} \mathbf{H(X)} &= \mathbf{H(X^{(i)}) + H(X_i|X^{(i)})} \\ &\leq \mathbf{H(X^{(i)}) + H(X_i|X_1, \ldots, X_{i-1})} \end{aligned}$$

Since $\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{H(X_i|X_1, \ldots, X_{i-1}) = H(X)}$, summing the inequality, we get

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

# subadditivity of entropy

The entropy of a random variable $Z \geq 0$ is

$$\mathbf{Ent}(Z) = \mathbb{E}\Phi(Z) - \Phi(\mathbb{E}Z)$$

where $\Phi(x) = x \log x$. By Jensen's inequality, $\mathbf{Ent}(Z) \geq 0$.

# subadditivity of entropy

The entropy of a random variable $Z \geq 0$ is

$$\mathbf{Ent}(Z) = \mathbb{E}\Phi(Z) - \Phi(\mathbb{E}Z)$$

where $\Phi(x) = x \log x$. By Jensen's inequality, $\mathbf{Ent}(Z) \geq 0$.

Han's inequality implies the following sub-additivity property. Let $X_1, \ldots, X_n$ be independent and let $Z = f(X_1, \ldots, X_n)$, where $f \geq 0$.
Denote

$$\mathbf{Ent}^{(i)}(Z) = \mathbb{E}^{(i)}\Phi(Z) - \Phi(\mathbb{E}^{(i)}Z)$$

Then

$$\mathbf{Ent}(Z) \leq \mathbb{E}\sum_{i=1}^{n} \mathbf{Ent}^{(i)}(Z) \ .$$

# a logarithmic sobolev inequality on the hypercube

Let $X = (X_1, \ldots, X_n)$ be uniformly distributed over $\{-1, 1\}^n$. If $f : \{-1, 1\}^n \to \mathbb{R}$ and $Z = f(X)$,

$$\mathrm{Ent}(Z^2) \leq \frac{1}{2}\mathbb{E}\sum_{i=1}^{n}(Z - Z_i')^2$$

The proof uses subadditivity of the entropy and calculus for the case $n = 1$.

Implies Efron-Stein and the edge-isoperimetric inequality.

# herbst's argument: exponential concentration

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

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

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

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

Differential inequalities are obtained for $F(\lambda)$.

# herbst's argument

As an example, suppose $f$ is such that $\sum_{i=1}^n (Z - Z_i')_+^2 \leq v$. Then by the log-Sobolev inequality,

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

If $G(\lambda) = \log F(\lambda)$, this becomes

$$\left(\frac{G(\lambda)}{\lambda}\right)' \leq \frac{v}{4} .$$

This can be integrated: $G(\lambda) \leq \lambda \mathbb{E}Z + \lambda v/4$, so

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

This implies

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

# herbst's argument

As an example, suppose $f$ is such that $\sum_{i=1}^n (Z - Z_i')_+^2 \leq v$. Then by the log-Sobolev inequality,

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

If $G(\lambda) = \log F(\lambda)$, this becomes

$$\left(\frac{G(\lambda)}{\lambda}\right)' \leq \frac{v}{4}\,.$$

This can be integrated: $G(\lambda) \leq \lambda \mathbb{E}Z + \lambda v/4$, so

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

This implies

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

Stronger than the bounded differences inequality!

# gaussian log-sobolev and concentration inequalities

Let $\mathbf{X = (X_1, \ldots, X_n)}$ be a vector of i.i.d. standard normal If $\mathbf{f : \mathbb{R}^n \to \mathbb{R}}$ and $\mathbf{Z = f(X)}$,

$$\mathbf{Ent(Z^2) \leq 2\mathbb{E}\left[\|\nabla f(X)\|^2\right]}$$

This can be proved using the central limit theorem and the Bernoulli log-Sobolev inequality.

# gaussian log-sobolev and concentration inequalities

Let $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_n})$ be a vector of i.i.d. standard normal If $\mathbf{f} : \mathbb{R}^{\mathbf{n}} \to \mathbb{R}$ and $\mathbf{Z} = \mathbf{f(X)}$,

$$\mathbf{Ent(Z^2)} \leq \mathbf{2}\mathbb{E}\left[\|\nabla \mathbf{f(X)}\|^{\mathbf{2}}\right]$$

This can be proved using the central limit theorem and the Bernoulli log-Sobolev inequality.
It implies the Gaussian concentration inequality:
Suppose $\mathbf{f}$ is Lipschitz: for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathbf{n}}$,

$$|\mathbf{f(x)} - \mathbf{f(y)}| \leq \mathbf{L}\|\mathbf{x} - \mathbf{y}\| \ .$$

Then, for all $\mathbf{t} > \mathbf{0}$,

$$\mathbb{P}\left\{\mathbf{f(X)} - \mathbb{E}\mathbf{f(X)} \geq \mathbf{t}\right\} \leq \mathbf{e}^{-\mathbf{t}^2/(\mathbf{2L}^2)} \ .$$

# an application: supremum of a gaussian process

Let $(X_t)_{t\in\mathcal{T}}$ be an almost surely continuous centered Gaussian process. Let $Z = \sup_{t\in\mathcal{T}} X_t$. If

$$\sigma^2 = \sup_{t\in\mathcal{T}} \left( \mathbb{E}\left[ X_t^2 \right] \right) ,$$

then

$$\mathbb{P}\left\{ |Z - \mathbb{E}Z| \geq u \right\} \leq 2e^{-u^2/(2\sigma^2)}$$

# an application: supremum of a gaussian process

Let $(X_t)_{t\in\mathcal{T}}$ be an almost surely continuous centered Gaussian process. Let $Z = \sup_{t\in\mathcal{T}} X_t$. If

$$\sigma^2 = \sup_{t\in\mathcal{T}} \left( \mathbb{E}\left[ X_t^2 \right] \right) \ ,$$

then

$$\mathbb{P}\left\{ |Z - \mathbb{E}Z| \geq u \right\} \leq 2e^{-u^2/(2\sigma^2)}$$

Proof: We may assume $\mathcal{T} = \{1, ..., n\}$. Let $\Gamma$ be the covariance matrix of $X = (X_1, \ldots, X_n)$. Let $A = \Gamma^{1/2}$. If $Y$ is a standard normal vector, then

$$f(Y) = \max_{i=1,\ldots,n} (AY)_i \overset{\text{distr.}}{=} \max_{i=1,\ldots,n} X_i$$

By Cauchy-Schwarz,

$$|(Au)_i - (Av)_i| = \left| \sum_j A_{i,j} (u_j - v_j) \right| \leq \left( \sum_j A_{i,j}^2 \right)^{1/2} \|u - v\|$$
$$\leq \sigma \|u - v\|$$

# beyond bernoulli and gaussian: the entropy method

For general distributions, logarithmic Sobolev inequalities are not available.

Solution: modified logarithmic Sobolev inequalities.
Suppose $X_1, \ldots, X_n$ are independent. Let $Z = f(X_1, \ldots, X_n)$ and $Z_i = f_i(X^{(i)}) = f_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$.

Let $\phi(x) = e^x - x - 1$. Then for all $\lambda \in \mathbb{R}$,

$$\lambda \mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda Z}\right]$$
$$\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z} \phi\left(-\lambda(Z - Z_i)\right)\right].$$

Michel Ledoux

# the entropy method

Define $Z_i = \inf_{x_i'} f(X_1, \ldots, x_i', \ldots, X_n)$ and suppose

$$\sum_{i=1}^{n} (Z - Z_i)^2 \leq v \ .$$

Then for all $t > 0$,

$$\mathbb{P}\left\{Z - \mathbb{E}Z > t\right\} \leq e^{-t^2/(2v)} \ .$$

# the entropy method

Define $Z_i = \inf_{x_i'} f(X_1, \ldots, x_i', \ldots, X_n)$ and suppose

$$\sum_{i=1}^{n} (Z - Z_i)^2 \leq v \ .$$

Then for all $t > 0$,

$$\mathbb{P}\left\{Z - \mathbb{E}Z > t\right\} \leq e^{-t^2/(2v)} \ .$$

This implies the bounded differences inequality and much more.

# example: the largest eigenvalue of a symmetric matrix

Let $\mathbf{A} = (\mathbf{X_{i,j}})_{\mathbf{n}\times\mathbf{n}}$ be symmetric, the $\mathbf{X_{i,j}}$ independent ($\mathbf{i \le j}$) with $|\mathbf{X_{i,j}}| \le \mathbf{1}$. Let

$$\mathbf{Z} = \lambda_1 = \sup_{\mathbf{u}:\|\mathbf{u}\|=1} \mathbf{u}^\mathsf{T}\mathbf{A}\mathbf{u} \ .$$

and suppose $\mathbf{v}$ is such that $\mathbf{Z} = \mathbf{v}^\mathsf{T}\mathbf{A}\mathbf{v}$.
$\mathbf{A'_{i,j}}$ is obtained by replacing $\mathbf{X_{i,j}}$ by $\mathbf{x'_{i,j}}$. Then

$$\begin{aligned}(\mathbf{Z} - \mathbf{Z_{i,j}})_+ &\le \left(\mathbf{v}^\mathsf{T}\mathbf{A}\mathbf{v} - \mathbf{v}^\mathsf{T}\mathbf{A'_{i,j}}\mathbf{v}\right) \mathbb{1}_{Z>Z_{i,j}} \\ &= \left(\mathbf{v}^\mathsf{T}(\mathbf{A} - \mathbf{A'_{i,j}})\mathbf{v}\right) \mathbb{1}_{Z>Z_{i,j}} \le 2\left(\mathbf{v_i v_j}(\mathbf{X_{i,j}} - \mathbf{X'_{i,j}})\right)_+ \\ &\le 4|\mathbf{v_i v_j}| \ .\end{aligned}$$

Therefore,

$$\sum_{1\le i\le j\le n} (\mathbf{Z} - \mathbf{Z'_{i,j}})_+^2 \le \sum_{1\le i\le j\le n} 16|\mathbf{v_i v_j}|^2 \le 16\left(\sum_{i=1}^{n} \mathbf{v_i^2}\right)^2 = 16 \ .$$

# example: convex lipschitz functions

Let $f : [0,1]^n \to \mathbb{R}$ be a convex function. Let $Z_i = \inf_{x_i'} f(X_1, \ldots, x_i', \ldots, X_n)$ and let $X_i'$ be the value of $x_i'$ for which the minimum is achieved. Then, writing $\overline{X}^{(i)} = (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$,

$$\begin{aligned}
\sum_{i=1}^{n}(Z - Z_i)^2 &= \sum_{i=1}^{n}(f(X) - f(\overline{X}^{(i)})^2 \\
&\leq \sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}(X)\right)^2 (X_i - X_i')^2 \\
&\quad \text{(by convexity)} \\
&\leq \sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}(X)\right)^2 \\
&= \|\nabla f(X)\|^2 \leq L^2 \,.
\end{aligned}$$

# self-bounding functions

Suppose $Z$ satisfies

$$0 \leq Z - Z_i \leq 1 \quad \text{and} \quad \sum_{i=1}^{n} (Z - Z_i) \leq Z \ .$$

Recall that $\mathbf{Var}(Z) \leq \mathbb{E}Z$. We have much more:

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

and

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

# self-bounding functions

Suppose $Z$ satisfies

$$0 \leq Z - Z_i \leq 1 \quad \text{and} \quad \sum_{i=1}^{n} (Z - Z_i) \leq Z \ .$$

Recall that $\mathbf{Var}(Z) \leq \mathbb{E}Z$. We have much more:

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

and

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

Rademacher averages and the random VC dimension are self bounding.

# self-bounding functions

Suppose $Z$ satisfies

$$0 \leq Z - Z_i \leq 1 \quad \text{and} \quad \sum_{i=1}^{n} (Z - Z_i) \leq Z \ .$$

Recall that $\mathbf{Var}(Z) \leq \mathbb{E}Z$. We have much more:

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

and

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

Rademacher averages and the random VC dimension are self bounding.

Configuration functions.

# weakly self-bounding functions

$f : \mathcal{X}^n \to [0, \infty)$ is weakly $(a, b)$-self-bounding if there exist $f_i : \mathcal{X}^{n-1} \to [0, \infty)$ such that for all $x \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \left( f(x) - f_i(x^{(i)}) \right)^2 \leq a f(x) + b \, .$$

# weakly self-bounding functions

$f : \mathcal{X}^n \to [0, \infty)$ is weakly $(a, b)$-self-bounding if there exist $f_i : \mathcal{X}^{n-1} \to [0, \infty)$ such that for all $x \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \left( f(x) - f_i(x^{(i)}) \right)^2 \leq a f(x) + b \, .$$

Then

$$\mathbb{P}\left\{ Z \geq \mathbb{E}Z + t \right\} \leq \exp\left( - \frac{t^2}{2\left(a\mathbb{E}Z + b + at/2\right)} \right) \, .$$

# weakly self-bounding functions

$f : \mathcal{X}^n \to [0, \infty)$ is weakly $(a, b)$-self-bounding if there exist $f_i : \mathcal{X}^{n-1} \to [0, \infty)$ such that for all $x \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \left( f(x) - f_i(x^{(i)}) \right)^2 \leq a f(x) + b \, .$$

Then

$$\mathbb{P}\left\{ Z \geq \mathbb{E}Z + t \right\} \leq \exp\left( - \frac{t^2}{2\left(a\mathbb{E}Z + b + at/2\right)} \right) \, .$$

If, in addition, $f(x) - f_i(x^{(i)}) \leq 1$, then for $0 < t \leq \mathbb{E}Z$,

$$\mathbb{P}\left\{ Z \leq \mathbb{E}Z - t \right\} \leq \exp\left( - \frac{t^2}{2\left(a\mathbb{E}Z + b + c\_t\right)} \right) \, .$$

where $c = (3a - 1)/6$.

# the isoperimetric view

Let $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_n})$ have independent components, taking values in $\mathcal{X}^{\mathbf{n}}$. Let $\mathbf{A} \subset \mathcal{X}^{\mathbf{n}}$.

The Hamming distance of $\mathbf{X}$ to $\mathbf{A}$ is

$$\mathbf{d(X, A)} = \min_{\mathbf{y} \in \mathbf{A}} \mathbf{d(X, y)} = \min_{\mathbf{y} \in \mathbf{A}} \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbb{1}_{\mathbf{X_i} \neq \mathbf{y_i}} \ .$$

Michel Talagrand

# the isoperimetric view

Let $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_n})$ have independent components, taking values in $\mathcal{X}^{\mathbf{n}}$. Let $\mathbf{A} \subset \mathcal{X}^{\mathbf{n}}$.

The Hamming distance of $\mathbf{X}$ to $\mathbf{A}$ is

$$\mathbf{d(X, A)} = \min_{\mathbf{y} \in \mathbf{A}} \mathbf{d(X, y)} = \min_{\mathbf{y} \in \mathbf{A}} \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbb{1}_{\mathbf{X_i} \neq \mathbf{y_i}} \ .$$

Michel Talagrand

$$\mathbb{P}\left\{ \mathbf{d(X, A)} \geq \mathbf{t} + \sqrt{\frac{\mathbf{n}}{\mathbf{2}} \log \frac{\mathbf{1}}{\mathbb{P}[\mathbf{A}]}} \right\} \leq \mathbf{e}^{-\mathbf{2t^2/n}} \ .$$

# the isoperimetric view

Let $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_n})$ have independent components, taking values in $\mathcal{X}^{\mathbf{n}}$. Let $\mathbf{A} \subset \mathcal{X}^{\mathbf{n}}$.

The Hamming distance of $\mathbf{X}$ to $\mathbf{A}$ is

$$\mathbf{d(X, A)} = \min_{\mathbf{y} \in \mathbf{A}} \mathbf{d(X, y)} = \min_{\mathbf{y} \in \mathbf{A}} \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbb{1}_{\mathbf{X_i} \neq \mathbf{y_i}} \ .$$

Michel Talagrand

$$\mathbb{P}\left\{ \mathbf{d(X, A)} \geq \mathbf{t} + \sqrt{\frac{\mathbf{n}}{\mathbf{2}} \log \frac{\mathbf{1}}{\mathbb{P}[\mathbf{A}]}} \right\} \leq \mathbf{e}^{-\mathbf{2t^2/n}} \ .$$

Concentration of measure!

# the isoperimetric view

Proof: By the bounded differences inequality,

$$\mathbb{P}\{\mathbb{E}d(X,A) - d(X,A) \geq t\} \leq e^{-2t^2/n}.$$

Taking $t = \mathbb{E}d(X,A)$, we get

$$\mathbb{E}d(X,A) \leq \sqrt{\frac{n}{2}\log\frac{1}{\mathbb{P}\{A\}}}.$$

By the bounded differences inequality again,

$$\mathbb{P}\left\{d(X,A) \geq t + \sqrt{\frac{n}{2}\log\frac{1}{\mathbb{P}\{A\}}}\right\} \leq e^{-2t^2/n}$$

# talagrand's convex distance

The weighted Hamming distance is

$$d_\alpha(x, A) = \inf_{y \in A} d_\alpha(x, y) = \inf_{y \in A} \sum_{i : x_i \neq y_i} |\alpha_i|$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)$. The same argument as before gives

$$\mathbb{P}\left\{ d_\alpha(X, A) \geq t + \sqrt{\frac{\|\alpha\|^2}{2} \log \frac{1}{\mathbb{P}\{A\}}} \right\} \leq e^{-2t^2/\|\alpha\|^2} ,$$

This implies

$$\sup_{\alpha : \|\alpha\| = 1} \min\left(\mathbb{P}\{A\}, \mathbb{P}\left\{d_\alpha(X, A) \geq t\right\}\right) \leq e^{-t^2/2} .$$

# convex distance inequality

convex distance:

$$\mathbf{d_T(x, A) = \sup_{\alpha \in [0,\infty)^n : \|\alpha\| = 1} d_\alpha(x, A)} \ .$$

$$\mathbf{\mathbb{P}\{A\}\mathbb{P}\left\{d_T(X, A) \geq t\right\} \leq e^{-t^2/4}} \ .$$

# convex distance inequality

convex distance:

$$\mathbf{d_T(x, A)} = \sup_{\alpha \in [0,\infty)^n : \|\alpha\| = 1} \mathbf{d_\alpha(x, A)} \ .$$

$$\mathbf{\mathbb{P}\{A\}\mathbb{P}\left\{d_T(X, A) \geq t\right\} \leq e^{-t^2/4}} \ .$$

Follows from the fact that $\mathbf{d_T(X, A)^2}$ is $\mathbf{(4, 0)}$ weakly self bounding (by a saddle point representation of $\mathbf{d_T}$).

Talagrand's original proof was different.

# convex lipschitz functions

For $\mathbf{A} \subset [0,1]^n$ and $\mathbf{x} \in [0,1]^n$, define

$$\mathbf{D(x, A)} = \inf_{\mathbf{y} \in \mathbf{A}} \|\mathbf{x} - \mathbf{y}\| \ .$$

If $\mathbf{A}$ is convex, then

$$\mathbf{D(x, A)} \leq \mathbf{d_T(x, A)} \ .$$

# convex lipschitz functions

For $\mathbf{A} \subset [0,1]^n$ and $\mathbf{x} \in [0,1]^n$, define

$$D(x, A) = \inf_{y \in A} \|x - y\| \ .$$

If $\mathbf{A}$ is convex, then

$$D(x, A) \leq d_T(x, A) \ .$$

Proof:

$$\begin{aligned} D(x, A) &= \inf_{\nu \in \mathcal{M}(A)} \|x - \mathbb{E}_\nu Y\| \quad \text{(since } A \text{ is convex)} \\ &\leq \inf_{\nu \in \mathcal{M}(A)} \sqrt{\sum_{j=1}^n \left(\mathbb{E}_\nu \mathbb{1}_{x_j \neq Y_j}\right)^2} \quad \text{(since } x_j, Y_j \in [0,1]\text{)} \\ &= \inf_{\nu \in \mathcal{M}(A)} \sup_{\alpha : \|\alpha\| \leq 1} \sum_{j=1}^n \alpha_j \mathbb{E}_\nu \mathbb{1}_{x_j \neq Y_j} \quad \text{(by Cauchy-Schwarz)} \\ &= d_T(x, A) \quad \text{(by minimax theorem)} \ . \end{aligned}$$

# convex lipschitz functions

Let $\mathbf{X = (X_1, \ldots, X_n)}$ have independent components taking values in $\mathbf{[0,1]}$. Let $\mathbf{f : [0,1]^n \to \mathbb{R}}$ be quasi-convex such that $\mathbf{|f(x) - f(y)| \leq \|x - y\|}$. Then

$$\mathbf{\mathbb{P}\{f(X) > \mathbb{M}f(X) + t\} \leq 2e^{-t^2/4}}$$

and

$$\mathbf{\mathbb{P}\{f(X) < \mathbb{M}f(X) - t\} \leq 2e^{-t^2/4}} \ .$$

# convex lipschitz functions

Let $\mathbf{X = (X_1, \ldots, X_n)}$ have independent components taking values in $\mathbf{[0,1]}$. Let $\mathbf{f : [0,1]^n \to \mathbb{R}}$ be quasi-convex such that $\mathbf{|f(x) - f(y)| \leq \|x - y\|}$. Then

$$\mathbf{\mathbb{P}\{f(X) > \mathbb{M}f(X) + t\} \leq 2e^{-t^2/4}}$$

and

$$\mathbf{\mathbb{P}\{f(X) < \mathbb{M}f(X) - t\} \leq 2e^{-t^2/4}} \ .$$

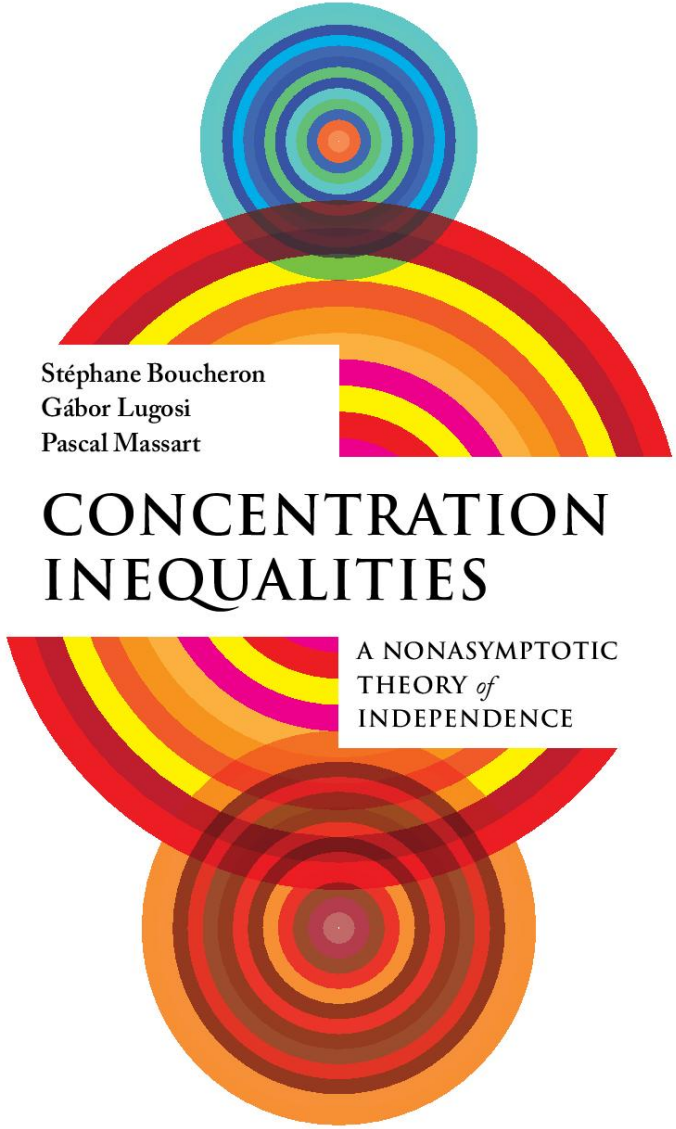
Proof: Let $\mathbf{A_s = \{x : f(x) \leq s\} \subset [0,1]^n}$. $\mathbf{A_s}$ is convex. Since $\mathbf{f}$ is Lipschitz,

$$\mathbf{f(x) \leq s + D(x, A_s) \leq s + d_T(x, A_s)} \ ,$$

By the convex distance inequality,

$$\mathbf{\mathbb{P}\{f(X) \geq s + t\}\mathbb{P}\{f(X) \leq s\} \leq e^{-t^2/4}} \ .$$

Take $\mathbf{s = \mathbb{M}f(X)}$ for the upper tail and $\mathbf{s = \mathbb{M}f(X) - t}$ for the lower tail.



Stéphane Boucheron
Gábor Lugosi
Pascal Massart

# CONCENTRATION INEQUALITIES

## A NONASYMPTOTIC THEORY *of* INDEPENDENCE

OXFORD