Concentration of measure in probability and high-dimensional statistical learning

Guillaume Aubrun, Aurélien Garivier, Rémi Gribonval

remi.gribonval@inria.fr

http://perso.ens-lyon.fr/remi.gribonval

Practical information

- **Lecturers**
  - ✓ **Guillaume Aubrun** **Aurélien Garivier** **Rémi Gribonval**



- **Joint course: maths & computer science**
  - ✓ Monday 13:30-15:30
  - ✓ Friday 10:15-12:15
  - ✓ this week: CS only
- **Language** : ~~french or~~ english ?
- **Attendance** : physical (and/or virtual as needed)

Remote attendance via https://ent-services.ens-lyon.fr/entVisio

Pour un bon usage de ce service, merci de vérifier que vous disposez de :

- Un ordinateur avec un micro-casque
- Ou Un téléphone avec son kit oreillette
- Une simple webcam pour la vidéo suffit (facultatif)
- Pour le son, nous recommandons vivement l'utilisation d'un micro-casque audio (l'utilisation d'un micro indépendant associé à des haut-
  externes peuvent provoquer un phénomène d'écho désagréable)
- Une connexion ADSL minimum

ENS DE LYON

Meeting		concentration
Nom d'utilisateur :		<your name> (for presence sheet)
Mot de passe :		m2concentration

Valider

Practical information

- **Official pad:**
  - ✓ latest general information on all courses
  - ✓ presence sheets (in class & online)

  https://pad.inria.fr/p/r.f0843991855b3c5006ef30aeb674d272
- **Web page that I maintain**

  https://people.irisa.fr/Remi.Gribonval/talks-and-tutorials/m2-ens-lyon-concentration/
  - ✓ course specific information
  - ✓ links, bibliographical references …
- **Evaluation**
  - ✓ Principle
    - ✦ Homework, in-class exercises & final exam: 50%
    - ✦ Final exam: 50%
  - ✓ More details in due time

Questions ?

Course context and objectives

High dimensional statistical learning

- **Goal**
  - use **training data** to infer parameters θ to achieve a certain **task**
  - **avoid overfitting**: ensure **generalization to unseen data** of similar type
- **Training collection = large point cloud $\mathcal{X}$**
  - signals, images, …
  - feature vectors, labels, …

Digit recognition (MNIST)

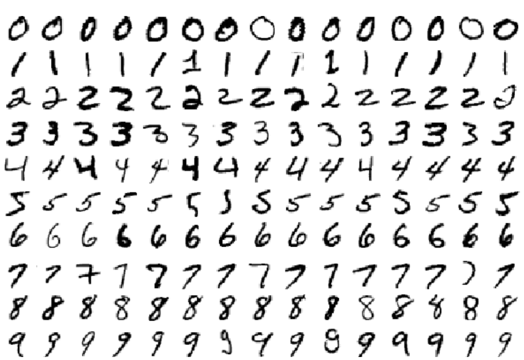
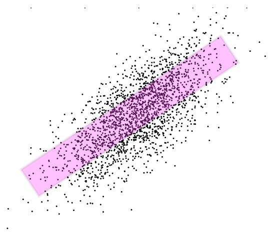


Image classification

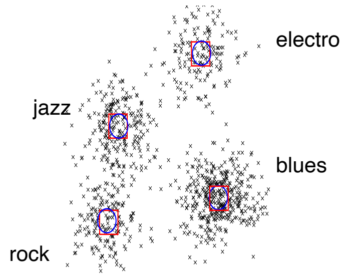
Sound classification

High dimensional statistical learning

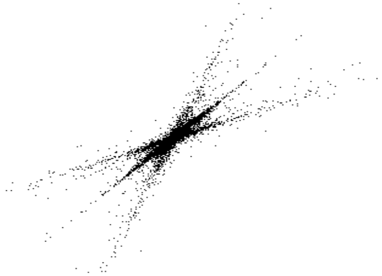
- **Goal**
  - use **training data** to infer parameters θ to achieve a certain **task**
  - **avoid overfitting**: ensure **generalization to unseen data** of similar type
- **Training collection = large point cloud** $\mathcal{X}$
  - signals, images, …
  - feature vectors, labels, …
- **Examples of tasks & parameters**



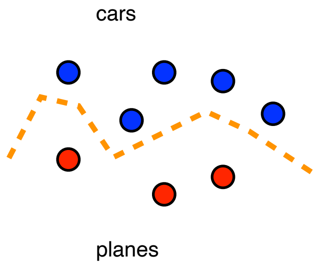
- PCA
  - θ = principal subspace



- Clustering
  - θ = centroids



- Dictionary learning
  - θ = dictionary atoms



- Classification
  - θ = classifier parameters (e.g. support vectors)

High dimensional **statistical learning**

- **Machine learning:**
  - ✓ focus on design of computationally efficient **algorithms**
- **Statistical learning:**
  - ✓ focus on proving **statistical guarantees**

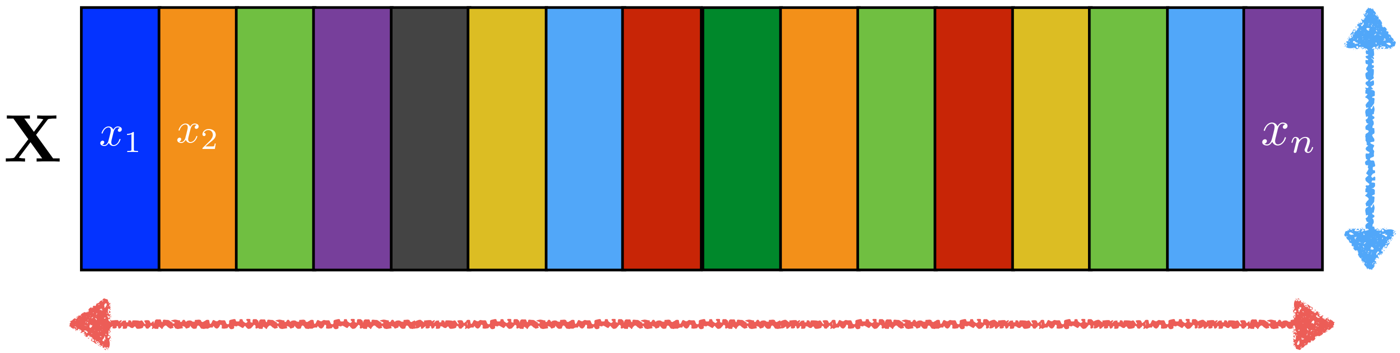
➡ **the PAC (Probably Approximately Correct) framework**

How many training samples do I need to learn accurately ?

➡ **notions of complexity / dimension of a learning task**

High dimensional statistical learning

- **Training collection** = collection of feature vectors



✓ High feature dimension *d*
✓ Large collection size *n* = *"volume"*

High dimensional statistical learning

- **Challenges of high dimension**
  - ✓ statistical *significance* of results
  - ✓ computational *scalability* of algorithms
  - ➡ **sparsity promoting algorithms** $d \gg n$
  - ➡ **dimension reduction when d "too large"**
  - ➡ **model selection**

Organization - CS viewpoint

- **Important tools for (high-dim) statistical learning**
  - ➡ **the PAC (Probably Approximately Correct) framework**
  - ➡ **notions of complexity / dimension of a learning task**
  - ➡ **sparsity promoting algorithms**
  - ➡ **dimension reduction with random projections**

  ✓ **Swiss knife:**
  - ✦ *measure concentration (probability theory)*

Organization - Maths viewpoint

- **Important concepts in probability**
  - ➡ **deviation inequalities** for averages of independent variables
  - ➡ **concentration** of high-dimensional random functions
  - ➡ **isoperimetry** in the sphere and Gaussian spaces

  ✓ **Applications:**
  - ✦ *analysis of random graphs*
  - ✦ *random projections for dimension reduction*
  - ✦ *structural risk minimization in machine learning*

Introduction to measure concentration

Why measure concentration ?

- **Experiment: draw a dice**

  - ✓ observation: a random number

$$X \in \{1, \ldots, 6\}$$

- **Repeat the experiment n times**
  - ✓ independent & identically distributed (**i.i.d.**) random numbers

$$X_i \in \{1, \ldots, 6\} \qquad 1 \leq i \leq n$$

  - ✓ compute the **empirical average**

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

  - ➡ **value ?**

Distribution of a random variable

$$X \sim \mathbb{P}$$

- **Discrete random variables**
  - ✓ ex: uniform distribution on $\{1, \ldots, 6\}$

$$P(X = 1) = \ldots = P(X = 6) = 1/6$$

- **Scalar random variables**
  - ✓ ex: Gaussian distribution

$$P(a \leq X \leq b) = \int_a^b \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}}_{p(x)} dx$$

- **Also: vector random variables …**

Expectation of a random variable

$$\mathbb{E}(X) = \mathbb{E}_{X \sim \mathbb{P}}(X)$$

- **Discrete random variables**

$$\mathbb{E}(X) = \sum_{x \in \Omega} x P(X = x)$$

= ????

  - ✓ ex: uniform distribution on $\Omega = \{1, \ldots, 6\}$

- **Scalar random variables**

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x p(x) dx$$

= ????

  - ✓ ex: Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Back to our problem …

- **Experiment: draw a dice**
  - ✓ observation: a random number

$$X \in \{1, \ldots, 6\}$$

- **Repeat the experiment n times**
  - ✓ independent & identically distributed (**i.i.d.**) random numbers

$$X_i \in \{1, \ldots, 6\} \qquad 1 \leq i \leq n$$

  - ✓ compute the **empirical average**

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

➡ **value** = ????

Overview

- **Law of large numbers**
- Central Limit Theorem
- Markov / Chebyshev / Chernoff / Hoeffding
- Summary

Law of large numbers

- **Expectation of the empirical average**

$$\mathbb{E}(\bar{X}_n)$$ = ????

- **Property**:
  - ✓ *when n gets large, the empirical average tends to the expectation*
  - ✓ mathematical expression

$$\lim_{n\to\infty} |\bar{X}_n - \mathbb{E}(X)| = 0$$?

Law of large numbers

- **Expectation of the empirical average**

$$\mathbb{E}(\bar{X}_n)$$ = ????

- **Property**:
  - ✓ *when n gets large, the empirical average tends to the expectation*
  - ✓ mathematical expression

$$\cancel{\lim_{n\to\infty} |\bar{X}_n - \mathbb{E}(X)| = 0} \; ?$$

Law of large numbers

- **Expectation of the empirical average**

$$\mathbb{E}(\bar{X}_n) \quad = ????$$

- **Property**:
  - ✓ *when n gets large, the empirical average tends to the expectation*
  - ✓ mathematical expression

$$\cancel{\lim_{n\to\infty} |\bar{X}_n - \mathbb{E}(X)| = 0} \; ?$$

$$\lim_{n\to\infty} P(|\bar{X}_n - \mathbb{E}(X)| > \epsilon) = 0$$

Law of large numbers

- **Expectation of the empirical average**

$$\mathbb{E}(\bar{X}_n) \quad = ????$$

- **Property**:
  - ✓ *when n gets large, the empirical average tends to the expectation*
  - ✓ mathematical expression

$$\cancel{\lim_{n\to\infty} |\bar{X}_n - \mathbb{E}(X)| = 0} \ ?$$

$$\forall \epsilon > 0 \qquad \lim_{n\to\infty} P(|\bar{X}_n - \mathbb{E}(X)| > \epsilon) = 0$$

Limits of law of large numbers

- **Law of large numbers**
  - ✓ randomness captured by one quantity

$$\rho(n,\epsilon) := P(|\bar{X}_n - \mathbb{E}(X)| > \epsilon)$$

  - ✓ asymptotic behavior = qualitative

$$\forall \epsilon > 0, \lim_{n\to\infty} \rho(n,\epsilon) = 0$$

- **Quantitative results ?** Target probability level $\rho(n,\epsilon) \leq \rho$
  - ✓ How many training samples $n(\rho,\epsilon)$?
  - ✓ What precision $\epsilon(n,\rho)$?
- **Order of magnitude of** $\rho(n,\epsilon)$ **?**

Overview

- Law of large numbers
- **Central Limit Theorem**
- Markov / Chebyshev / Chernoff / Hoeffding
- Summary

Variance of a random variable

- **Definition** $\mathtt{Var}(X) := \mathbb{E}[(X - \mathbb{E}(X))^2]$
- **Property** $\mathtt{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$

  Proof ?

- **Examples**
  - ✓ uniform distribution on $\Omega = \{1, \ldots, 6\}$

    $\mathtt{Var}(X)$ = ????

  - ✓ Gaussian distribution $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$

    $\mathtt{Var}(X)$ = ????

Variance of the empirical average

- **Consider i.i.d. samples with finite variance**

$$X_i \sim \mathbb{P} \qquad \mathtt{Var}(X_i) = \sigma^2 < \infty, 1 \leq i \leq n$$

  - ✓ *Expectation*

$$\mathbb{E}(\bar{X}_n)$$ = ????

  - ✓ *Variance*

$$\mathtt{Var}(\bar{X}_n)$$ = ????

  - ✓ *Rescaled variance*

$$\mathtt{Var}[\sqrt{n}(\bar{X}_n - \mathbb{E}(X))]$$ = ????

Variance of the empirical average

- **Consider i.i.d. samples with finite variance**

$$X_i \sim \mathbb{P} \qquad \mathtt{Var}(X_i) = \sigma^2 < \infty, 1 \leq i \leq n$$

  - ✓ *Expectation*

$$\mathbb{E}(\bar{X}_n)$$ = ????

  - ✓ *Variance*

$$\mathtt{Var}(\bar{X}_n)$$ = ????

  - ✓ *Rescaled variance*

$$\mathtt{Var}[\sqrt{n}(\bar{X}_n - \mathbb{E}(X))]$$ = ????

- example where variance is infinite ?

Example: Histograms with binomial variables

$n = 1 \longrightarrow$

$\longleftarrow n = 512$

Source [github](https://github.com/resendedaniel/math/tree/master/17-central-limit-theorem)
Daniel Resende (Creative Commons Attribution-Share Alike 4.0 International license)

Central Limit Theorem (CLT)

- **Consider i.i.d. samples with finite variance** σ^2

- **Theorem**
  - ✓ *the distribution of the empirical average converges to a Gaussian distribution*
  - ✓ mathematical expression

$$\forall a, b, \lim_{n \to \infty} P(\sqrt{n}(\bar{X}_n - \mathbb{E}(X)) \in [a, b]) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}} dt$$

Central Limit Theorem (CLT)

- **Consider i.i.d. samples with finite variance** σ^2
- **Theorem**
  - ✓ *the distribution of the empirical average converges to a Gaussian distribution*
  - ✓ mathematical expression

$$\forall a, b, \lim_{n \to \infty} P(\sqrt{n}(\bar{X}_n - \mathbb{E}(X)) \in [a, b]) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}} dt$$

$$\lim_{n \to \infty} P(|\bar{X}_n - \mathbb{E}(X)| > \tfrac{t}{\sqrt{n}})$$ = ????

Complementary error function

- **Definition** $$\mathtt{erfc}(z) := \frac{2}{\sqrt{\pi}} \cdot \int_z^{+\infty} e^{-u^2} du$$
- **Property** $$\mathtt{erfc}(z) \leq e^{-z^2}, \quad \forall z > 0$$
  - ✓ see [1]

Complementary error function

- **Definition** $$\mathtt{erfc}(z) := \frac{2}{\sqrt{\pi}} \cdot \int_z^{+\infty} e^{-u^2} du$$
- **Property** $$\mathtt{erfc}(z) \leq e^{-z^2}, \quad \forall z > 0$$
  - ✓ see [1]
- **Consequence of CLT**
  - ✓ for n "large enough"

$$P(|\bar{X}_n - \mathbb{E}(X)| > \tfrac{\epsilon}{\sqrt{n}}) \approx 2 \int_\epsilon^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}} dt$$

Complementary error function

- **Definition** $$\mathtt{erfc}(z) := \frac{2}{\sqrt{\pi}} \cdot \int_z^{+\infty} e^{-u^2} du$$
- **Property** $$\mathtt{erfc}(z) \leq e^{-z^2}, \quad \forall z > 0$$
  - ✓ see [1]
- **Consequence of CLT**
  - ✓ for n “large enough”

$$P(|\bar{X}_n - \mathbb{E}(X)| > \tfrac{\epsilon}{\sqrt{n}}) \approx 2 \int_\epsilon^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}} dt \quad = \frac{2}{\sqrt{\pi}} \cdot \int_{\frac{\epsilon}{\sigma\sqrt{2}}}^{+\infty} e^{-u^2} du$$

Complementary error function

- **Definition**

$$\texttt{erfc}(z) := \frac{2}{\sqrt{\pi}} \cdot \int_z^{+\infty} e^{-u^2} du$$

- **Property**
  - ✓ see [1]

$$\texttt{erfc}(z) \leq e^{-z^2}, \quad \forall z > 0$$

- **Consequence of CLT**
  - ✓ for n “large enough”

$$P(|\bar{X}_n - \mathbb{E}(X)| > \tfrac{\epsilon}{\sqrt{n}}) \approx 2\int_{\epsilon}^{+\infty} \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}} dt \quad = \frac{2}{\sqrt{\pi}} \cdot \int_{\frac{\epsilon}{\sigma\sqrt{2}}}^{+\infty} e^{-u^2} du \quad = \texttt{erfc}(\tfrac{\epsilon}{\sigma\sqrt{2}}) \leq e^{-\frac{\epsilon^2}{2\sigma^2}}$$

Complementary error function

- **Definition** $$\texttt{erfc}(z) := \frac{2}{\sqrt{\pi}} \cdot \int_z^{+\infty} e^{-u^2} du$$
- **Property** $$\texttt{erfc}(z) \leq e^{-z^2}, \quad \forall z > 0$$
  - ✓ see [1]
- **Consequence of CLT**
  - ✓ for n "large enough"

$$P(|\bar{X}_n - \mathbb{E}(X)| > \tfrac{\epsilon}{\sqrt{n}}) \approx 2 \int_\epsilon^{+\infty} \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}} dt \quad = \tfrac{2}{\sqrt{\pi}} \cdot \int_{\frac{\epsilon}{\sigma\sqrt{2}}}^{+\infty} e^{-u^2} du \quad = \texttt{erfc}(\tfrac{\epsilon}{\sigma\sqrt{2}}) \leq e^{-\frac{\epsilon^2}{2\sigma^2}}$$

- **Asymptotically: exponential decay with n**

$$P(|\bar{X}_n - \mathbb{E}(X)| > t) \lesssim e^{-\frac{nt^2}{2\sigma^2}}$$

Complementary error function

- **Definition** $$\mathtt{erfc}(z) := \frac{2}{\sqrt{\pi}} \cdot \int_z^{+\infty} e^{-u^2} du$$
- **Property** $$\mathtt{erfc}(z) \leq e^{-z^2}, \quad \forall z > 0$$
  - ✓ see [1]
- **Consequence of CLT**
  - ✓ for n “large enough”

➡ **How many training samples ?**

$$P(|\bar{X}_n - \mathbb{E}(X)| > \tfrac{\epsilon}{\sqrt{n}}) \approx 2\int_\epsilon^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}} dt \quad = \frac{2}{\sqrt{\pi}} \cdot \int_{\frac{\epsilon}{\sigma\sqrt{2}}}^{+\infty} e^{-u^2} du \quad = \mathtt{erfc}(\tfrac{\epsilon}{\sigma\sqrt{2}}) \leq e^{-\frac{\epsilon^2}{2\sigma^2}}$$

- **Asymptotically: exponential decay with n**

$$P(|\bar{X}_n - \mathbb{E}(X)| > t) \lesssim e^{-\frac{nt^2}{2\sigma^2}}$$

Complementary error function

- **Definition**

$$\texttt{erfc}(z) := \frac{2}{\sqrt{\pi}} \cdot \int_z^{+\infty} e^{-u^2} du$$

- **Property**
  - see [1]

$$\texttt{erfc}(z) \leq e^{-z^2}, \quad \forall z > 0$$

- **Consequence of CLT**
  - for n "large enough"

➡ **How many training samples ?**

$$P(|\bar{X}_n - \mathbb{E}(X)| > \tfrac{\epsilon}{\sqrt{n}}) \approx 2\int_\epsilon^{+\infty} \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}} dt \quad = \tfrac{2}{\sqrt{\pi}} \cdot \int_{\frac{\epsilon}{\sigma\sqrt{2}}}^{+\infty} e^{-u^2} du \quad = \texttt{erfc}(\tfrac{\epsilon}{\sigma\sqrt{2}}) \leq e^{-\frac{\epsilon^2}{2\sigma^2}}$$

- **Asymptotically: exponential decay with n**

$$P(|\bar{X}_n - \mathbb{E}(X)| > t) \lesssim e^{-\frac{nt^2}{2\sigma^2}}$$

➡ **How accurately ?**

Need for *finite-sample* results

- **Probability of a given deviation**

$$P(\bar{X}_n \geq \mathbb{E}(X) + t) \leq ?$$

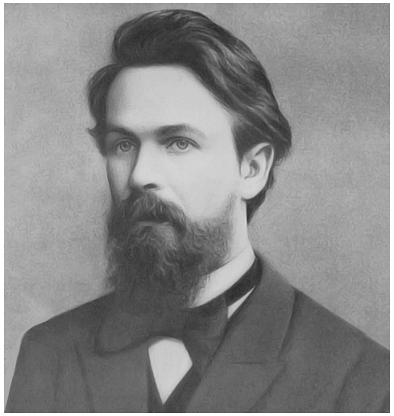
- **Will be achieved with a series of tools to control**

$$P(Z \geq \mathbb{E}(Z) + t)$$

  - ✓ with various assumptions on the random variable Z
  - ✓ applied to certain random variables $Z = f(\bar{X}_n)$

Overview

- Law of large numbers
- Central Limit Theorem
- **Markov / Chebyshev / Chernoff / Hoeffding**
- Summary



Markov's inequality

(due to Chebyshev, Markov's teacher)

Andreï A. Markov
1856-1922
Russian

- **Property**
  - ✓ For a *non-negative* random variable *Z* we have for any *t>0*

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

- **Remark**: decay $O(1/t)$, not exponential

Crédit photo: domaine public

Proof of Markov's inequality

- Blackboard

Chebyshev’s inequality (Exercise)

Pafnouti Tchebychev
1821-1894
Russian

- **Property**
  - ✓ Consider a random variable *Z* with finite variance

$$\sigma^2 = \mathtt{Var}(Z) < \infty$$

  - ✓ Then for any *t>0*

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

- **Proof:** = ????
- **Remark**: decay $O(1/t^2)$, still not exponential

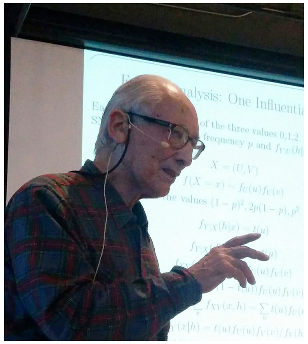
Crédit photo: domaine public

Proof of Chebyshev's inequality

- Exercise

Beyond Chebyshev's inequality ?

- Can you propose extensions of Chebyshev's inequality that yield faster 'concentration' to the mean (under stronger assumptions) ?



in 2015

Chernoff bound

(due to Herman Rubin)

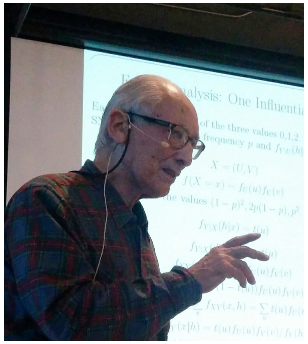
Herman Chernoff
Born 1923

- **Property**
  - ✓ For any $t, \lambda > 0$ we have

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

- **Remark**: exponential decay !
  - ✓ need to assume that (for small enough λ) we have

$$\mathbb{E}(e^{\lambda Z}) < \infty$$



in 2015

Chernoff bound

(due to Herman Rubin)

Herman Chernoff
Born 1923
American

- **Property**
  - ✓ For any $t, \lambda > 0$ we have

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

- **Remark**: exponential decay !
  - ✓ need to assume that (for small enough λ) we have

$$\mathbb{E}(e^{\lambda Z}) < \infty$$

Historical note

Lin, X., Genest, C., Banks, D., Molenberghs, G., Scott, D., & Wang, J.-L. (Eds.). (2014). Past, Present and Future of Statistical Science (pp. 1–1). Chapman and Hall/CRC. http://doi.org/10.1201/b16720-2

In working on an artificial example, I discovered that I was using the Central Limit Theorem for large deviations where it did not apply. This led me to derive the asymptotic upper and lower bounds that were needed for the tail probabilities. ***Rubin claimed he could get these bounds with much less work and I challenged him. He produced a rather simple argument****, using the Markov inequality, for the upper bound. Since that seemed to be a minor lemma in the ensuing paper I published (Chernoff, 1952), I neglected to give him credit. I now consider it a serious error in judgment, especially because his result is stronger, for the upper bound, than the asymptotic result I had derived.*

Shannon had published a paper using the Central Limit Theorem as an approximation for large deviations and had been criticized for that. My paper permitted him to modify his results and led to a great deal of publicity in the computer science literature for ***the so-called Chernoff bound which was really Rubin's result****.*

Proof of Chernoff's inequality

- Exercise

Example: Bounded Random Variable

- **Property 1**:
  - ✓ Consider a bounded random variable $a \leq Z \leq b$
  - ✓ Denote $\mu := \mathbb{E}(Z)$
  - ✓ For any $\lambda > 0$ we have

$$\mathbb{E}(e^{\lambda(Z-\mu)}) \leq e^{\frac{\lambda^2(b-a)^2}{8}}$$

Proof: next time

Hoeffding's inequality

Wassily Hoeffding
1914-1991
Born in Finland
American

- **Theorem**
  - ✓ Consider ***independent bounded*** random variables with common expectation

$$a \leq X_i \leq b \qquad \mathbb{E}(X_i) = \mu$$

  - ✓ **For any n and any** *t*>0 the empirical average $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ satisfies

$$P(|\bar{X}_n - \mu| > t) \leq 2e^{-\frac{2nt^2}{(b-a)^2}}$$

Crédit photo: licence inconnue, https://www.nap.edu/read/11429/chapter/12 The National Academies Press

Proof: Hoeffding's inequality (1)

- **Step 1**

$$\begin{aligned}\mathbb{E}(e^{\lambda(\bar{X}_n-\mu)}) &= \mathbb{E}(e^{\frac{\lambda}{n}\sum_{i=1}^n (X_i-\mu)}) \\ &= \Pi_{i=1}^n \mathbb{E}(e^{\frac{\lambda}{n}(X_i-\mu)}) \\ &\leq \Pi_{i=1}^n \exp\left(\frac{\lambda^2(b-a)^2}{8n^2}\right) \\ &= e^{\frac{\lambda^2(b-a)^2}{8n}}\end{aligned}$$

- **Step 2: Chernoff's bound**

$$P(\bar{X}_n-\mu>t) \leq \frac{\mathbb{E}(e^{\lambda(\bar{X}_n-\mu)})}{e^{\lambda t}} \leq e^{\frac{\lambda^2(b-a)^2}{8n}-\lambda t}$$

Proof: Hoeffding's inequality (2)

- **Step 3: optimize $\lambda > 0$ for bound** $e^{\frac{\lambda^2(b-a)^2}{8n} - \lambda t}$

$$\min_{\lambda>0} \left\{ \frac{\lambda^2(b-a)^2}{8n} - \lambda t \right\} = -\frac{2nt^2}{(b-a)^2}$$

  - ✓ achieved for $\lambda = \dfrac{4nt}{(b-a)^2}$

- **Step 4: repeat with** $X_i' = -X_i$

Overview

- Law of large numbers
- Central Limit Theorem
- Markov / Chebyshev / Chernoff / Hoeffding
- **Summary**

Summary: Chernoff's method

- **Theorem** (sometimes known as Chernoff's bound)
  - For any random variable *Z*, with $\mu := \mathbb{E}(Z)$

$$\log P(Z - \mu > t) \leq -\sup_{\lambda > 0} \left\{ \lambda t - \log \mathbb{E}(e^{\lambda(Z-\mu)}) \right\}$$

- **Definitions**:
  - *Moment-generating function* $M_Z(\lambda) := \mathbb{E}(e^{\lambda Z})$
  - ***Cumulant-generating function*** $K_Z(\lambda) := \log \mathbb{E}(e^{\lambda Z})$

Summary: measure concentration

- **Nature of the results:**
  - ✓ probability bounds valid for any finite *n*
  - ✓ exponential decay with number *n* of samples
- **Technical ingredients:**
  - ✓ Chernoff's method
  - ✓ Bounds on cumulant generating function
- **Hoeffding's inequality**
  - ✓ valid for independent & bounded random variables
- **Other tools:**
  - ✓ beyond boundedness: sub-Gaussian/sub-exponential r.v.
  - ✓ beyond empirical average: McDiarmid's inequality
  - ✓ Lipschitz functions of Gaussian random variables

That’s all folks !

References

[1] M. Chiani, D. Dardari, and M. K. Simon, "New exponential bounds and approximations for the computation of error probability in fading channels," IEEE Transactions on Wireless Communications, vol. 24, no. 5, pp. 840–845, 2003.