CLT, Hoeffdings ≠

Prof: Yaniv Plan
("Yuh-neev Plan")

Class website:
www.yanivplan.com/math-608d

Theory: Understanding
"probabilistic objects" in
high dimensions, e.g., random variables,
random matrices, functions with random
input, shapes as seen through "prob. lens."

Applications: Compressed sensing,
statistical estimation, dimension
reduction, etc.
Concentration of measure

1. Deviation inequalities for sums of indep r.v.'s

1.1 CLT: asymptotic and non-asymptotic

Normal r.v.: \( g \sim \mathcal{N}(0,1) \)

Density: \( f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \)

Ubiquitous in science. Why?

Thm: (CLT, Lindeberg-Levy)

Let \( X_1, X_2, \ldots \) be iid r.v.'s

with \( \mathbb{E} X_i = 0, \text{ Var}(X_i) = 1 \). Then

\[
\lim_{n \to \infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \to \mathcal{N}(0,1)
\]

uniformly in distribution.

Remarks:

0: Role of \( \frac{1}{\sqrt{n}} \): \( \text{ Var}\left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \right) = \text{Var}(g) = 1 \)
"Uniformly in distribution"

\[
\lim_{n \to \infty} \sup_{t \in \mathbb{R}} |P(\sum_{i=1}^{n} \frac{x_i}{\sqrt{n}} < t) - \Phi(t)| \to 0
\]

\[\text{c.d.f. of normal c.d.f.}\]

\[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i\]

Why is \(\mathcal{N}(0,1)\) so useful?

- Fast tail decay

\[f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\]

0.1%

Roughly: \(P(g > t) \approx \frac{1}{t} e^{-\frac{t^2}{2}} = \frac{1}{t} f(t)\)

See: [Durrett, Probability: Theory & Examples, Thm 1.4]

Precisely: \(((\frac{1}{t} - \frac{1}{t^2}) f(t) \leq P(g > t) \leq \frac{1}{t} f(t)\)

\[
P(g > t) \leq \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \quad t \geq 1
\]

\[\text{super-exponential tail decay}\]
Question Does \( \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \) have the same fast tail decay?

CLT suggests "yes", but this does not follow from CLT. Depends on convergence rate of CLT.

**Thm (CLT: Berry-Esseen)**

\[
\sup_{t \in \mathbb{R}} \left| P\left( \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i < t \right) - P(g > t) \right| \leq \frac{C \rho}{\sqrt{n}}
\]

where \( \rho := \mathbb{E}|X_i|^3 \), \( C = \text{abs. const} \)

**HW**

Show by example that the rate \( \frac{C \rho}{\sqrt{n}} \)

cannot be improved, i.e., find a distribution for \( X_i \) and an abs. const \( C \), so that

\[
\sup_{t \in \mathbb{R}} \left| P\left( \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i < t \right) - P(g > t) \right| \geq \frac{C \rho}{\sqrt{n}} \quad \text{for all } n \geq 1.
\]

**Remark:** Non-asymptotic version of CLT.

This gives a non-asymptotic tail bound:

\[
P\left( \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i > t \right) \leq P(g > t) + \frac{C \rho}{\sqrt{n}} \leq e^{-t^2} + \frac{C \rho}{\sqrt{n}}
\]

\( \text{hides abs. const.} \)

- Unfortunately, \( \frac{C \rho}{\sqrt{n}} \) does not depend on \( t \), and can ruin exponential decay.
Further one cannot hope for exponential decay with only the condition \(|x|^3 < \infty\). (Why?)

\section{Inequality: Deviation inequalities: Bound tail of \(\frac{1}{n} \sum_{i=1}^{n} X_i\) directly, without comparing to \(\mathcal{N}(0,1)\).}

\subsection{Hoeffding's inequality for Rademacher r.v.'s}

\textbf{Def.} \(X \sim \text{Rademacher}\) iff

\[
P(X = 1) = P(X = -1) = 0.5
\]

Consider \(\frac{1}{n} \sum_{i=1}^{n} X_i\), \(X_i \sim \text{Rademacher}\)

\textbf{Normalization: Require variance} \(\text{var} X_i = 1\).

\[
\text{Var} \left( \sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) = \sum_{i=1}^{n} a_i^2
\]

\textbf{Req.} \(\sum_{i=1}^{n} a_i^2 = 1\)

\textbf{First attempt: Bound tail with moment bound.}

\[
P \left( \sum_{i=1}^{n} a_i X_i > t \right) = P \left( \left( \sum_{i=1}^{n} a_i X_i \right)^2 > t^2 \right)
\]

\[
\leq \frac{\mathbb{E} \left( \sum_{i=1}^{n} a_i X_i \right)^2}{t^2} = \frac{1}{t^2}
\]

\textbf{Chebychev}
\( \frac{1}{e^t} \) is not exponential!

2nd attempt:

\[
P(\exists a_i, x_i > t) = P(\exp(\sum a_i, x_i) > e^{\lambda t})
\]

\[
\leq e^{-\lambda t} \exp(\sum a_i x_i)
\]

(by Chebyshev)

\[
= e^{-\lambda t} \prod_{i=1}^n e^{a_i x_i}
\]

(by indep.)

\[
\text{Note: } \cosh(x) = 1 + \frac{x^2}{2} + \frac{x^4}{4!} + \ldots
\]

\[
\exp\left(\frac{x^2}{2}\right) = 1 + \frac{x^2}{2} + \frac{x^4}{4!} + \ldots
\]

Exercise: \( \cosh(x) \leq \exp\left(\frac{x^2}{2}\right) \)

\[
\Rightarrow (\star) \leq e^{-\lambda t} \prod e^{\frac{(a_i x_i)^2}{\sqrt{2}}} = \exp(-\lambda t + \frac{1}{2} \sum a_i^2)
\]

\[
= \exp(-\lambda t + \frac{\lambda^2}{2})
\]

Optimize \( \lambda : \lambda = t \)

\[
\Rightarrow (\star) \leq \exp\left(\frac{-t^2}{2}\right)
\]

We have proven:

\[
\text{Thm. Hoeffding } \frac{1}{e^t} \text{ not exponential!}
\]

Let \( X_i \) be indep. Rademacher.
\[ \frac{1}{n} \sum_{i=1}^{n} a_i^2 = 1. \] Then

\[ P \left( \frac{1}{n} \sum_{i=1}^{n} x_i^2 > t \right) \leq \exp \left( -\frac{t^2}{2} \right), \quad t > 0. \]

Remarks:

1. \( \sim \) matches normal tail: \( P(X > t) \leq e^{-\frac{t^2}{2}} \)

2. Non-asymptotic

3. Method is very flexible.
   (Due to S. Bernstein.)
   Generalizes to other distributions.
   (Even to random matrices.)

**Literature:** [G. Lugosi, concentration of measure inequalities]

Next lecture: Largest class of r.v.'s st. result of above flavor occurs.