U Statistic and Concentration Inequality

Zhe Gao

School of Management
University of Science and Technology of China

20, March, 2025

Outline

The Importance of Euclidean Space Methods

Why Study Euclidean Space Methods?

- Foundation of classical statistical theory.
- Critical for understanding basic tools like concentration inequalities, CLT, empirical processes.
- Essential for traditional hypothesis testing and model building.
- Provides tools and insights that are applicable even in more complex spaces.

Goals:

- To grasp fundamental statistical concepts and theories.
- To understand the limits and applicability of these methods in modern contexts.

Extending to Non-Euclidean Spaces

Why Extend to Non-Euclidean Spaces?

- Emerging data structures (like graphs, manifolds) are not well-handled by traditional methods.
- Critical for advancements in fields such as machine learning, data science, and network analysis.
- Offers new perspectives and tools to tackle complex datasets.

Goals:

- To adapt and extend Euclidean methodologies to more complex, abstract spaces.
- To explore novel statistical techniques that cater to the geometrical properties of non-Euclidean data.

Euclidean Space: Key Topics

Topics Covered:

- Theories of U-statistics and Central Limit Theorems.
- Concentration inequalities.
- Time series and weakly dependent sequences.
- Fundamentals of graphical models.
- Theories of empirical processes and regression error analysis.
- Introduction to neural networks and existing theoretical frameworks.

Non-Euclidean Space: Key Topics

Topics Covered:

- Basic theories and classical examples of non-Euclidean spaces.
- Two-sample tests and independence tests in non-Euclidean settings.
- Change-point problems and theories of weakly dependent sequences in non-Euclidean spaces.
- Graphical models in non-Euclidean spaces.
- Principal Component Analysis (PCA) in non-Euclidean spaces.
- Regression models for non-Euclidean data.
- Generative models in non-Euclidean spaces.

Seminar Outline

Time: Every Thursdays, 2:00 - 4:00 PM.

序号	主要内容	重点	汇报人
1	U 统计量、极限定理、集中不等式	U 统计量的基础理论，常见的集中不等式证明技巧	高哲
2	弱依赖性序列与图模型	弱依赖性序列的极限理论，图模型的基础理论	彭辉阳
3	经验过程理论	经验过程的基础理论，基础统计模型的误差理论分析	高哲
4	神经网络基础理论	神经网络的逼近定理，估计误差分析	张小可
5	非欧空间的基础理论	非欧空间中常见量的定义，经典的例子	章寒露
6	非欧空间的两样本检验与独立性检验问题	两样本检验的方法，非欧空间的推广	陈鹏
7	非欧空间的变点问题以及弱依赖性序列理论	变点问题的理论，弱依赖性序列极限定理	夏伽其
8	非欧空间的图模型	概率图模型在非欧空间的推广	陈书延
9	非欧空间的主成分分析	因子分析，主成分分析在非欧空间的推广	胡佳琪
10	非欧空间的回归模型	非欧空间经典回归方法，Fréchet 回归、随机森林等	兰敬国
11	非欧空间的生成模型	diffusion model, flow model 在非欧空间的推广	金璋

Data and Initial Exploration

Problem Setup:

- We have a dataset containing a response variable Y and a predictor variable X.
- Objective: model the relationship between X and Y.

Measure and Testing Independence:

- Establish divergence measures (or dependency measures), metric of the measure.
- To determine whether there is a statistically significant relationship between X and Y.
- Apply theories of U-statistics and concentration inequalities to support the inference process.

Modeling and Analysis

Exploring the Intrinsic Structure of X:

- Explore if X has a low-dimensional representation or heterogeneity, for example through factor analysis or principal component analysis.
- Analyze if X exhibits characteristics of time series or weak dependency structures.

Regression Models and Predictive Performance:

- Apply advanced regression methods such as neural networks and random forests to model the impact of X on Y.
- Evaluate the predictive performance of the models, using theories of empirical processes to analyze regression errors and model stability.

Outline

Definition of U-Statistics

Definition 2.1

For a real-valued symmetric measurable function, $h(x_1, \ldots, x_m)$ and for a sample, $X_1, \ldots, X_n$, of size $n \geq m$ from a distribution P, a U-statistic with kernel h is defined as

$$U_n = \frac{1}{\binom{n}{m}} \sum_{(n,m)} h\left(X_{i_1}, \ldots, X_{i_m}\right).$$

Optimality of U-Statistics

- If $\theta(P) = E_P[h(X_1, \ldots, X_m)]$ exists for all $P \in \mathcal{P}$, then the U-statistic is an unbiased estimate of $\theta(P)$.
- Under some mild conditions, for example if $\mathcal{P}$ contains all distributions, P, for which $\theta(P)$ is finite. Then the order statistics form a complete sufficient statistic from $P \in \mathcal{P}$. And U_n is a function of the order statistics, and so is the best unbiased estimate of its expectation (Hodges-Lehmann theorem). This means no unbiased estimate of $\theta(P)$, based on $X_1, \ldots, X_n$, can have a variance smaller than the variance of U_n. (Theorem 3 in page 3 in Lee 1990)

Example

Example 1: Sample Variance

Let $\mathcal{P}$ be the set of all distributions with second moment finite:

$$\mathcal{P} = \left\{ P : \int |x|^2 dP(x) < \infty \right\}.$$

Then we can define the variance functional on $\mathcal{P}$ by

$$\operatorname{Var} P = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2} (x_1 - x_2)^2 \, dP(x_1) \, dP(x_2)$$

which is estimated by the sample variance $s_n^2 = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \frac{1}{2} \left(X_i - X_j \right)^2$.

Example

Example 2: The Wilcoxon Signed Rank Test

Let $\mathcal{P}$ be the family of continuous distributions on the real line. The Wilcoxon signed rank test is based on the statistic

$$W_n^+ = \sum_{i=1}^{n} R_i^+ I\left(Z_i > 0\right) = nU_n^{(1)} + \binom{n}{2} U_n^{(2)}.$$

where $Z_i = X_i - Y_i$, R_i^+ is the rank of $|Z_i|$ among $|Z_1|, |Z_2|, \ldots, |Z_n|$. W_n^+ is a linear combination of two U-statistics, the first U-statistic is $U_n^{(1)} = n^{-1} \sum_1^n I\left(Z_i > 0\right)$ with the kernel $h(z) = I(z > 0)$. The second U-statistic is $U_n^{(2)} = \binom{n}{2}^{-1} \sum_{i<j} I\left(Z_i + Z_j > 0\right)$ with kernel $h\left(z_1, z_2\right) = I\left(z_1 + z_2 > 0\right)$. For large n the second term dominates the first, so asymptotically W_n^+ behaves like $n^2 U_n^{(2)}/2$.

The variance of a U-statistic

Define for $c = 1, 2, \ldots, m$ the conditional expectations

$$h_c(x_1, \ldots, x_c) = E\{h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_k)\}$$

and their variances

$$\sigma_c^2 = \mathrm{Var}\{h_c(X_1, \ldots, X_c)\}.$$

Theorem 2.2

For $P \in \mathcal{P}$, the variance of a U-statistic is

$$\mathrm{Var}(U_n) = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2.$$

The H-decomposition

Define kernels $h^{(1)}, h^{(2)}, \ldots, h^{(k)}$ of degrees $1, 2, \ldots, k$ which are defined recursively by the equations

$$h^{(1)}(x_1) = h_1(x_1) - \theta$$

and

$$h^{(c)}(x_1, \ldots, x_c) = h_c(x_1, \ldots, x_c) - \sum_{j=1}^{c-1} \sum_{(c,j)} h^{(j)}\left(x_{i_1}, \ldots, x_{i_j}\right) - \theta$$

for $c = 2, 3, \ldots, m$.

The H-decomposition

Theorem 2.3

For $j = 1, 2, \ldots, k$, let $H_n^{(j)}$ be the U-statistic based on the kernel $h^{(j)}$. Then

$$U_n = \theta + \sum_{j=1}^{k} \binom{k}{j} H_n^{(j)}.$$

The asymptotic distribution of a U-statistic

Theorem 2.4

If $\sigma_1 > 0$, *then*

$$\sqrt{n}(U_n - \theta) \to N(m^2\sigma_1^2).$$

The asymptotic distribution of a U-statistic

Consider the U-statistic, $U_n^{(2)}$ with kernel, $h(x_1, x_2) = I(x_1 + x_2 > 0)$ of degree $m = 2$, associated with the Wilcoxon signed rank test. The parameter estimated is $\theta = \mathrm{E}h(X_1, X_2) = \mathrm{P}(X_1 + X_2 > 0)$. Note that

$$\sigma_1^2 = \mathrm{Cov}(h(X_1, X_2), h(X_1, X_3)) = \mathrm{P}(X_1 + X_2 > 0, X_1 + X_3 > 0) - \theta^2.$$

Under the null hypothesis that the distribution P is symmetric about 0 , we have $\theta = 1/2$ and $\mathrm{P}(X_1 + X_2 > 0, X_1 + X_3 > 0) = 1/3$. Therefore, under the null hypothesis, $\sigma_1^2 = (1/3) - (1/2)^2 = 1/12$, then,

$$\sqrt{n}\left(U_n^{(2)} - 1/2\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1/3).$$

This test of the null hypothesis based on $U_n^{(2)}$ is consistent only for alternatives P for which $\theta(P) \neq 1/2$.

Degeneracy

Definition 2.5

We say that a U-statistic has a degeneracy of order k if $\sigma_1^2 = \cdots = \sigma_k^2 = 0$ and $\sigma_{k+1}^2 > 0$.

Consider the kernel, $h(x_1, x_2) = x_1 x_2$, $h_1(x_1) = E(x_1 X_2) = x_1 E(X_2) = x_1 \mu$, and $\sigma_1^2 = Var(h_1(X_1)) = \mu^2 \sigma^2$, where $\sigma^2 = Var(X_1)$. If $\mu = E(X_1) = 0$ under the null hypothesis. Then the U-statistic has a degeneracy of order 1.

First order degeneracy

For given i.i.d. random variables, X_1 and X_2, any symmetric, square integrable function,

$$A(x_1, x_2) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(x_1) \varphi_k(x_2)$$

where the λ_k are real numbers, and the φ_k are an orthonormal sequence,

$$\mathrm{E}\varphi_j(X_1)\varphi_k(X_1) = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases}$$

The λ_k are the eigenvalues, and the $\varphi_k(x)$ are corresponding eigenfunctions of the transformation, $g(x) \to \mathrm{E}A(x, X_1) g(X_1)$. Then

$$\sum_{k=1}^{n} \lambda_k \varphi_k(X_1) \varphi_k(X_2) \xrightarrow{L_2} A(X_1, X_2).$$

First order degeneracy

Theorem 2.6

Let U_n be the U-statistic associated with a symmetric kernel of degree 2, degeneracy of order 1, and expectation θ. Then

$$n(U_n - \theta) \to \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1)$$

where $Z_1, Z_2, \ldots$ are independent $N(0,1)$ and $\lambda_1, \lambda_2, \ldots$ are the eigenvalues of $A(x_1, x_2) = h(x_1, x_2) - \theta$.

The Berry-Esseen theorem for U-statistics

Theorem 2.7

Let U_n be a non-degenerate U-statistic of degree 2 , based on a sequence of i.i.d. random variables $\{X_n\}$. Suppose that the kernel ψ has an H-decomposition

$$\psi(x_1,x_2)=\theta+h^{(1)}(x_1)+h^{(1)}(x_2)+h^{(2)}(x_1,x_2)$$

where $E\left|h^{(1)}(X_1)\right|^3<\infty$ and $E\left|h^{(2)}(X_1,X_2)\right|^{5/3}<\infty$. Let ρ denote the quantity $E\left|h^{(1)}(X_1)\right|^3/\sigma_1^3$ and $\lambda_p=E\left|h^{(2)}(X_1,X_2)\right|^p/\sigma_1^p$. Then there exist constants C_1,C_2 and C_3 depending neither on n,ψ nor the distribution of the X' s such that

$$\sup_x\left|\Pr\left(\sqrt{n}\left(U_n-\theta\right)/2\sigma_1\leq x\right)-\Phi(x)\right|\leq\left\{C_1\rho+C_2\lambda_{5/3}+C_3\left(\rho\lambda_{3/2}\right)^{2/3}\right\}n^{-\frac{1}{2}}$$

for all $n\geq 2$.

The law of the iterated logarithm for U-statistics

Theorem 2.8

Let $\{U_n\}$ be a sequence of U-statistic based on a non-degenerate kernel h of order m satisfying $E|h(X_1,\ldots,X_m)|^2 < \infty$ and hence having asymptotic variance $m^2\sigma_1^2/n$. Then the LIL holds for U_n:

$$\limsup_n \frac{n(U_n-\theta)}{\sqrt{2m^2\sigma_1^2 n\log n\log n}} = 1, \quad a.s.$$

V-statistics

Definition 2.9

For a real-valued symmetric measurable function, $h(x_1, \ldots, x_m)$ and for a sample, $X_1, \ldots, X_n$, of size $n \geq m$ from a distribution P, a V-statistic with kernel h is defined as

$$V_n = \frac{1}{n^m} \sum_{i_1, \cdots i_m = 1}^{n} h\left(X_{i_1}, \ldots, X_{i_m}\right).$$

V-statistics

The V-statistics can be written as:

$$V_n = \frac{1}{n^m} \sum_{j=1}^{m} j! S_k^{(j)} \binom{n}{j} U_n^{(j)},$$

where $U_n^{(j)}$ is a U-statistics with kernel

$$\phi_{(j)}(X_1, \ldots, X_j) = \frac{1}{j! S_k^{(j)}} \sum h\left(X_{i_1}, \ldots, X_{i_m}\right),$$

and $S_k^{(j)}$ are Stirling numbers of the second kind.

Example

For $m = 3$,

$$n^3 V_n = 6\binom{n}{3} U_n^{(3)} + 6\binom{n}{2} U_n^{(2)} + \binom{n}{1} U_n^{(1)}.$$

The kernels are

$$\begin{aligned}
\phi_{(3)}(x_1, x_2, x_3) &= \frac{1}{6} \sum_{(3)} \psi\left(x_{i_1}, x_{i_2}, x_{i_3}\right) = \psi(x_1, x_2, x_3) \\
\phi_{(2)}(x_1, x_2) &= \frac{1}{6} (\psi(x_1, x_1, x_2) + \psi(x_1, x_2, x_1) + \psi(x_2, x_1, x_1) \\
&\quad + \psi(x_1, x_2, x_2) + \psi(x_2, x_1, x_2) + \psi(x_2, x_2, x_1)) \\
&= \frac{1}{2} (\psi(x_1, x_1, x_2) + \psi(x_1, x_2, x_2)) \\
\phi_{(1)}(x_1) &= \psi(x_1, x_1, x_1).
\end{aligned}$$

Incomplete U-statistics

The incomplete U-statistics is

$$U_n^{(0)} = \frac{1}{N} \sum_{s \in \mathcal{D}} h(s),$$

where the sum is taken over the N subsets. Then

$$Var(U_n^{(0)}) \geq Var(U_n).$$

Incomplete U-statistics

Theorem 2.10

Let $U_n^{(0)}$ be a U-statistic constructed by selecting N sets at random with replacement from $\mathcal{S}_{n,k}$, and U_n the corresponding complete statistic, assumed to be degenerate of order d. Let $\lim_{n\to\infty} n^{d+1}N^{-1} = \alpha$, and assume all necessary variances exist.

(i) If $\alpha = 0$ then $n^{(d+1)/2}\left(U_n^{(0)} - \theta\right)$ has the same limit distribution as $n^{(d+1)/2}(U_n - \theta)$;

(ii) If $0 < \alpha < \infty$ then the limit distribution of $N^{\frac{1}{2}}\left(U_n^{(0)} - \theta\right)$ is that of the r.v. $\alpha^{\frac{1}{2}}X + \sigma_k Y$, where X has the same distribution as the limiting distribution of $n^{(d+1)/2}(U_n - \theta)$, Y is $N(0,1)$, and X and Y an independent;

(iii) If $\alpha = \infty$, then the limit distribution of $N^{\frac{1}{2}}\left(U_n^{(0)} - \theta\right)$ is $N\left(0, \sigma_k^2\right)$.

Generalised U-statistics

For $j = 1, \ldots, k$, let $X_{j1}, \ldots, X_{jn_j}$ be i.i.d. sample for F_j, X_{ij} and $X_{i'j'}$ are independent for $j \neq j'$. The kernel function is

$$h(x_{11}, \ldots, x_{1m_1}, x_{21}, \ldots, x_{2m_2}, \ldots, x_{k1}, \ldots, x_{km_k}).$$

The Generalised U-statistics is defined as

$$\frac{1}{\binom{n_1}{m_1} \cdots \binom{n_k}{m_k}} \sum_{(n_1, m_1)} \cdots \sum_{(n_k, m_k)} h(S_1, \ldots, S_k).$$

The variance of generalised U-statistics

Theorem 2.11

Let U_{n_1,n_2} be a generalised U-statistic with $k = 2$ based on a kernel h having degrees m_1 and m_2. Then

$$\operatorname{Var} U_{n_1,n_2} = \sum_{c=0}^{m_1} \sum_{d=0}^{m_2} \frac{\binom{m_1}{c}\binom{m_2}{d}\binom{n_1-m_1}{m_1-c}\binom{n_2-m_2}{m_2-d}}{\binom{n_1}{m_1}\binom{n_2}{m_2}} \sigma_{c,d}^2.$$

Example

Example: The two-sample Wilcoxon (Mann-Whitney) statistic.
Let $X_1, \ldots, X_{n1}$ and $Y_1, \ldots, Y_{n_2}$ be two independent samples with $n_1 \geq n_2$ from absolutely continuous distributions F and G, and let R_j denote the rank of Y_j in the combined sample. Then the Wilcoxon rank sum statistic is

$$W = \sum_{j=1}^{n_2} R_j.$$

Example

Example: The two-sample Wilcoxon (Mann-Whitney) statistic.
If we define

$$\phi(x,y) = \begin{cases} 1 & \text{if } x < y \\ 0 & \text{otherwise} \end{cases}$$

and

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi\left(X_i, Y_j\right)$$

then in the absence of ties it can be shown that

$$W = U + n_2\left(n_2+1\right)/2.$$

The statistic U is the Mann-Whitney U-statistic, and the related statistic $U_{n_1,n_2} = \left(n_1 n_2\right)^{-1} U$ is clearly a generalised U-statistic. The mean of U_{n_1,n_2} is just $E\phi\left(X_1, Y_1\right) = \Pr\left(X_1 < Y_1\right)$.

Example

Example: The two-sample Wilcoxon (Mann-Whitney) statistic.
The variance is

$$\operatorname{Var} U_{n_1,n_2} = (n_1 n_2)^{-1} \left\{ (n_1 - 1)\, \sigma_{0,1}^2 + (n_2 - 1)\, \sigma_{1,0}^2 + \sigma_{1,1}^2 \right\}$$

where

$$\sigma_{0,1}^2 = \Pr(X_1 < Y_1, X_2 < Y_1) - \Pr^2(X_1 < Y_1) = \frac{1}{12},$$
$$\sigma_{1,0}^2 = \Pr(X_1 < Y_1, X_1 < Y_2) - \Pr^2(X_1 < Y_1) = \frac{1}{12},$$
$$\sigma_{1,1}^2 = \Pr(X_1 < Y_1) \{1 - \Pr(X_1 < Y_1)\} = \frac{1}{4}.$$

Thus, $\operatorname{Var} U_{n_1,n_2} = \frac{n_1+n_2+1}{12 n_1 n_2}$, $EW = \frac{n_2(n_1+n_2+1)}{2}$, $\operatorname{Var} W = \frac{n_1 n_2 (n_1+n_2+1)}{12}$.

Summary

- H-decomposition.
- Asymptotic distribution.
- LIL.
- V-statistics.
- Incomplete U-statistics.

Definition of Martingales

Definition 2.12

A filtered space is $\left(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n=0}^{\infty}, \mathbb{P}\right)$, where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, and $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F}$ are σ-algebras, jointly called a filtration. We also define $\mathcal{F}_\infty := \sigma\left(\bigcup_n \mathcal{F}_n\right) \subseteq \mathcal{F}$.

Definition 2.13

A process (sequence of random variables, that is) X_n is adapted to the filtration $(\mathcal{F}_n)_{n\geq 0}$, if for every n, the variable X_n is $\mathcal{F}_n$-measurable.

Definition of Martingales

Definition 2.14

A process $(X_n)_{n\geq 0}$ in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a martingale with respect to a filtration $(\mathcal{F}_n)_{n\geq 0}$, if
- it is adapted to $(\mathcal{F}_n)_{n\geq 0}$;
- $\mathbb{E}|X_n| < \infty, \forall n \geq 0$;
- $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) = X_n$ a.s., $\forall n \geq 0$.

If $(X_n)_{n\geq 0}$ is a martingale, then $Y_n = X_n - X_{n-1}$ is called martingales difference sequence (MDS).

Example of Martingales

Example: Simple random walk

Consider the successive tosses of a fair coin and let $\xi_n = 1$ if the nth toss is heads and $\xi_n = -1$ if the nth toss is tails. Let $X_n = \xi_1 + \cdots + \xi_n$ and $\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)$ for $n \geq 1$, $X_0 = 0$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Then $X_n, n \geq 0$ is a martingale with respect to $\mathcal{F}_n$.

Martingales central limit theorem

Theorem 2.15

Let $\{X_i\}$ be a zero-mean, square intergrable martingales difference sequence with filtration $\{\mathcal{F}_i\}$. If

(i) $\sum_i \mathbb{E}\left[X_i^2 \mid \mathcal{F}_{i-1}\right] \xrightarrow{p} \eta$,

(ii) $\forall \varepsilon > 0, \sum_i \mathbb{E}\left[X_i^2 I(X_i > \varepsilon)\right] \xrightarrow{p} 0.$

Then $\sum_i X_i \xrightarrow{d} Z$, $Z \sim N(0, \eta)$.

Convergence rate of MCLT

Theorem 2.16

Let $\{X_i\}$ be a square integrable martingale difference sequence with filtration $\mathcal{F}_i$. If the Lindeberg condition is hold, then for $\delta > 0$, there is some constant $C_\delta > 0$ depending only on δ such that

$$\sup_{x \in R} |P(S_n \leq x) - \Phi(x)| \leq C_\delta \left(L_{n,2\delta} + N_{n,2\delta}\right)^{\frac{1}{3+2\delta}}$$

where $S_n = \sum_{i=1}^n X_i$, $L_{n,2\delta} = \sum_{i=1}^n \mathbb{E}|X_i|^{2+2\delta}$,
$Q_{n,2\delta} = \mathbb{E}\left|\sum_{i=1}^n \mathbb{E}\left[X_i^2 \mid \mathcal{F}_{i-1}\right] - 1\right|^{1+\delta}$.

LIL of Martingales

Theorem 2.17

Let $\{X_i\}$ be a stationary, square integrable martingale difference sequence with filtration $\mathcal{F}_i$, then

$$\frac{1}{\sqrt{2n\log(\log(n))}} \limsup \sum_{i=1}^{n} X_i = 1$$

Darling-Erdős theorems for martingales

Theorem 2.18

Let $\{X_i\}$ be a square integrable martingale difference sequence with filtration $\mathcal{F}_i$. Assume that

$$s_n^2 = \sum_{j=1}^{n} E(X_j^2 | \mathcal{F}_{j-1}) \to \infty$$

and there exists a positive sequence $\epsilon_n \in \mathcal{F}_{n-1}$ with $\epsilon_n \to 0$ such that

$$|X_n| \leq \epsilon_n s_n / (\log(\log(s_n)))^{3/2},$$

then let $S(t) = S_n, s_n^2 \leq t < s_{n+1}^2$, as $T \to \infty$,

$$a(T) \sup_{1 \leq t \leq T} S(t)/\sqrt{t} - b(T) \to E$$

where $a(T) = \sqrt{2\log(\log(T))}$,
$b(T) = 2\log(\log(T)) + \log(\log(\log(T)))/2 - \log(4\pi)/2.$

Outline

What are Concentration Inequalities?

Definition:

- Concentration inequalities provide bounds on how a random variable deviates from some central value such as its mean or median.
- These inequalities are powerful tools in probability theory to measure the extent to which a random variable differs from a typical value.

Key Points:

- They quantify the tail behavior of distributions.
- Useful in scenarios where one needs to understand the probabilities of extreme deviations.

Why are Concentration Inequalities Important?

Applications:

- **Machine Learning:** Assess the reliability of learning algorithms and generalization bounds.
- **Statistics:** Confidence intervals and hypothesis tests are based on understanding the fluctuations of sample means and sums.
- **Operations Research:** Risk assessment and stochastic optimization.

Advantages:

- Provide guarantees about the behavior of random processes.
- Enable rigorous analysis of algorithms and systems under randomness.

Historical Perspective and Key Theorems

Historical Context:

- The development of concentration inequalities can be traced back to the work on the law of large numbers and the central limit theorem.
- Modern advances began with the introduction of the Chernoff bound and have since expanded significantly.

Key Theorems:

- **Markov's Inequality:** Provides an upper bound on the probability that a non-negative random variable is greater than a certain value.
- **Chebyshev's Inequality:** A special case of Markov's inequality when dealing with variance.
- **Chernoff Bound:** Gives exponentially decreasing bounds on tail distributions for sums of independent random variables.

Detailed Introduction to Basic Concentration Inequalities

Markov's Inequality:

- Statement: For any non-negative random variable X and $a > 0$,

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

- Significance: Provides a very general but loose bound, applicable to any non-negative random variable.

Chebyshev's Inequality:

- Statement: For any random variable X with finite expected value μ and finite non-zero variance σ^2, for $k > 0$

$$\Pr(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}.$$

- Significance: Useful for variables with known mean and variance, gives a tighter bound compared to Markov's inequality.

The General Chernoff Bound

- For any upper tail bound $t > 0$, the Chernoff Bound states:

$$\Pr(X \geq a) \leq \inf_{t>0} \frac{\mathbb{E}[e^{tX}]}{e^{ta}},$$

  Here $\mathbb{E}[e^{tX}]$ is the moment-generating function.
- Let $X = \sum_{i=1}^{n} X_i$, where X_i are independent random variables,

$$\Pr(X \geq a) \leq \inf_{t>0} \frac{\mathbb{E}[e^{tX}]}{e^{ta}} = \inf_{t>0} e^{-ta} \prod_{i=1}^{n} M_{X_i}(t)$$

What is the Hoeffding Inequality?

- The Hoeffding Inequality is a fundamental result in probability theory that provides an upper bound on the probability that the sum of bounded independent random variables deviates from its expected value by a certain amount.
- It is particularly useful in scenarios involving sums of independent and bounded random variables.

Theorem 3.1 (Hoeffding's inequality)

Let $X_1, \ldots, X_n$ be independent RVs satisfying bound condition $a_i \leq X_i \leq b_i$. For all $t \geq 0$, we have

$$P\left(\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{2\sum_{i=1}^{n}(b_i - a_i)^2}\right\}$$

Hoeffding's inequality

Empirical distribution function:
Let $\{X_i\}_{i=1}^n$ $i.i.d \sim F(x)$, and $F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x)$ be the empirical distribution. By Hoeffding's inequality,

$$P(|F_n(x) - F(x)| \ge t) \le 2e^{-2nt^2}$$

Lemma 3.2 (Hoeffding's lemma)

Let X be a RV with $EX = 0$, and $a \le X \le b$. For all $\lambda \ge 0$, we have

$$Ee^{\lambda X} \le \exp\left\{\frac{\lambda^2}{8}(b-a)^2\right\}$$

Derivation of the Hoeffding Inequality

Key Concepts in Derivation:

- The inequality is derived using the concept of the moment generating function (MGF) of the bounded random variables.
- By applying Markov's inequality to the exponential of the sum of these random variables, and optimizing over a parameter, the bound is obtained.

Steps in Derivation:

- Consider the exponential $e^{t(S_n - \mathbb{E}[S_n])}$ and apply the expectation.
- Use the independence of X_i to separate the expectations and apply individual bounds based on the MGFs of bounded variables.
- Minimize over t to find the tightest bound.

What is Bernstein's Inequality

- Bernstein's Inequality is a powerful tool in probability theory used to bound the sums of independent random variables.
- It is particularly valuable when dealing with variables that have bounded differences from their means.

Theorem 3.3

Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $|X_i - \mathbb{E}[X_i]| \leq M$ for some $M > 0$, and let $\sigma^2 = \sum_{i=1}^{n} Var(X_i)$. Then, for any $t > 0$,

$$\Pr\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp\left(-\frac{t^2}{2(\sigma^2 + \frac{1}{3}Mt)}\right)$$

Derivation of Bernstein's Inequality

Key Concepts in Derivation:

- Bernstein's Inequality is derived using the moment generating function (MGF) approach, similar to the Chernoff bound.
- The derivation involves bounding the MGFs of the centered variables and then applying Markov's inequality.

Steps in Derivation:

- Use the fact that e^{tx} for $t > 0$ can be bounded above by an exponential series involving the variances and bounded differences.
- Apply the union bound and optimize over t to obtain the final inequality form.

Introduction to Sub-Gaussian Distributions

What are Sub-Gaussian Distributions?

- Sub-Gaussian distributions are a class of probability distributions that exhibit tail behavior similar to or lighter than that of a Gaussian distribution.
- These distributions are characterized by having tails that decay at least as fast as the tails of a normal distribution.

Key Properties:

- **Bounded Moments:** The k-th moment of a sub-Gaussian random variable is bounded by the k-th moment of a Gaussian random variable with some variance σ^2.
- **Tail Behavior:** The probability of deviations from the mean decays exponentially fast, which is expressed as:

$$\Pr(|X-\mu| > t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

Sub-Gaussian

Definition 3.4 (Sub-Gaussian)

A RV X with mean μ is sub-Gaussian with a parameter σ (denoted $X \sim \text{subG}\left(\sigma^2\right)$) if its MGF satisfies

$$Ee^{\lambda(X-\mu)} \leq e^{\frac{\sigma^2\lambda^2}{2}}, \quad \forall \lambda \in \mathbb{R}$$

- Classical examples: Gaussian, Bernoulli, Bounded.
- The σ^2 is indeed the upper bounds of $\text{Var}(X)$.

$$\begin{aligned} Ee^{\lambda(X-\mu)} - 1 &= \lambda E(X-\mu) + \frac{\lambda^2}{2} E(X-\mu)^2 + o\left(\lambda^2\right) \\ &\leq e^{\frac{\sigma^2\lambda^2}{2}} - 1 = \frac{\lambda^2\sigma^2}{2} + o\left(\lambda^2\right) \end{aligned}$$

Remark on sub-Gaussian

- By the Chernoff's inequality, the sub-Gaussian RV satisfies

$$P(|X-\mu| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}, \quad \forall t \in \mathbb{R}$$

- By Hoeffding's lemma, all bounded RVs are essentially sub-Gaussian with $\sigma^2 = \frac{(b-a)^2}{4}$.

Equal definition

Let X be a RV with $EX = 0$. Then, the following are equivalent for finite positive constants $\{K_i\}_{i=1}^5$.

- The tails of X : $P\{|X| \geq t\} \leq 2\exp\left(-t^2/K_1^2\right), \quad \forall t \geq 0.$
- The moments of X : $\left(E|X|^k\right)^{1/k} \leq K_2\sqrt{k}, \quad \forall integer k \geq 1.$
- The local MGF of X^2 : $Ee^{\lambda^2 X^2} \leq e^{K_3^2\lambda^2}, \quad \forall |\lambda| \leq \frac{1}{K_3}.$
- The exponential moment of X^2 : $Ee^{X^2/K_4^2} \leq 2.$
- The MGF of X : $Ee^{\lambda X} \leq e^{K_5^2\lambda^2}, \quad \forall \lambda \in \mathbb{R}.$

Sub-Gaussian norm

Definition 3.5 (sub-Gaussian norm)

For a sub-Gaussian RV X, its sub-Gaussian norm is defined by

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : E[e^{\frac{x^2}{t^2}}] \leq 2\right\}$$

Remark

- Triangle inequality: define $\psi_2(x) = e^{x^2} - 1$, we have $E\psi_2\left(\frac{X}{\|X\|_{\psi_2}}\right) \leq 1$ and $E\psi_2\left(\frac{Y}{\|Y\|_{\psi_2}}\right) \leq 1$, the convexity yields

$$E\psi_2\left(\frac{X+Y}{\|X\|_{\psi_2} + \|Y\|_{\psi_2}}\right) \leq 1$$

- X is sub-Gaussian $\Longleftrightarrow \|X\|_{\psi_2} < \infty$.

Remarks on sub-Gaussian

Suppose that X_1 and X_2 are zeromean and sub-Gaussian with parameters σ_1 and σ_2, respectively.

- If X_1 and X_2 are independent, show that the random variable $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$. (Prove by Hoeffding's inequality)
- The random variable $X_1 + X_2$ is sub-Gaussian with parameter at most $\sqrt{2}\sqrt{\sigma_1^2 + \sigma_2^2}\ (\sigma_1 + \sigma_2)$.

Hoeffding inequality for sub-Gaussian

Theorem 3.6

Let $\{X_i\}_{i=1}^n$ be independent σ_i^2-sub-Gaussian RV s. Then

$$Ee^{\lambda \sum_{i=1}^n (X_i - \mu_i)} \leq e^{\frac{\lambda^2 \sum_{i=1}^n \sigma_i^2}{2}}, \quad \forall \lambda \in \mathbb{R}$$

The sum $\sum_{i=1}^n X_i$ is a $\sum_{i=1}^n \sigma_i^2$-sub-Gaussian RV. Then for all $t \geq 0$, we have

$$P\left(\left|\sum_{i=1}^n (X_i - \mu_i) \geq t\right|\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right)$$

Introduction to Sub-Exponential Distributions

What are Sub-Exponential Distributions?

- Sub-exponential distributions are probability distributions with tails that decay slower than those of Gaussian distributions but faster than those of heavy-tailed distributions.
- They are useful in describing the distribution of sums of independent random variables that themselves have heavier tails than Gaussian variables.

Key Properties:

- **Tail Decay:** For sub-exponential variables, the tails decay according to:

$$\Pr(|X-\mu| > t) \leq 2\exp\left(-\frac{t}{\beta}\right), \text{ for } t \geq 0$$

- **Moment Conditions:** The moment generating function (MGF) exists and behaves in a certain controlled manner beyond a certain threshold.

Sub-exponential

Definition 3.7 (Sub-exponentia)

A random variable X with mean $\mu = EX$ is sub-exponential with parameters (v, α) (denoted $X \sim \text{subE}(v, \alpha)$) if its MGF atisfies

$$Ee^{\lambda(X-\mu)} \leq e^{\frac{v^2\lambda^2}{2}} \quad \forall |\lambda| < \frac{1}{\alpha}$$

Suppose that X is sub-exponential with parameters (v, α). Then

$$P(X - \mu \geq t) \leq \begin{cases} e^{-\frac{t^2}{2v^2}}, & 0 \leq t \leq \frac{v^2}{\alpha} \\ e^{-\frac{t}{2\alpha}}, & t > \frac{v^2}{\alpha}. \end{cases}$$

Example

Exponential distribution: $X \sim \mathrm{Exp}(\mu)$.

$$Ee^{\lambda\left(X-\frac{1}{\mu}\right)} = \begin{cases} \frac{1}{1-\lambda/\mu} e^{-\frac{\lambda}{\mu}}, & \lambda < \mu \\ \infty, & \lambda \geq \mu \end{cases}$$

By the inequality $\frac{1}{1-x} e^{-x} \leq e^{\frac{3}{2}x^2}$, which is valid for all $x \in \left[-\frac{1}{2}, \frac{1}{2}\right]$, we have

$$Ee^{\lambda\left(X-\frac{1}{\mu}\right)} \leq e^{\frac{3\lambda^2}{2\mu^2}}, \quad \forall |\lambda| \leq \frac{1}{2}\mu$$

Example

Chi-square distribution: $X \sim \chi_1^2$.

$$Ee^{\lambda(X-1)} = \begin{cases} \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}, & \lambda < \frac{1}{2} \\ \infty, & \lambda \geq \frac{1}{2} \end{cases}$$

By the inequality $\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2}, \quad \forall |\lambda| \leq \frac{1}{4}$, we have

$$Ee^{\lambda(X-1)} \leq e^{2\lambda^2}, \quad \forall |\lambda| \leq \frac{1}{4}$$

Equal definition

Let X be a RV with $EX = 0$. Then, the following are equivalent for finite positive constants $\{K_i\}_{i=1}^5$.

- The tails of X : $P\{|X| \geq t\} \leq 2e^{-t/K_2}, \quad \forall t \geq 0.$
- The moments of X : $\left(E|X|^k\right)^{1/k} \leq K_3 k, \quad \forall \text{ integer } k \geq 1.$
- The local MGF of $|X|$: $Ee^{\lambda|X|} \leq e^{(K_5\lambda)}, \quad \forall 0 \leq \lambda \leq \frac{1}{K_5}.$
- The exponential moment of $|X|$: $Ee^{|X|/K_4} \leq 2.$
- The MGF of X : $Ee^{(\lambda X)} \leq e^{K_1^2\lambda^2}, \quad \forall |\lambda| \leq \frac{1}{K_1}.$

sub-exponential norm

Definition 3.8 (sub-exponential norm)

For a sub-exponential RV X, its sub-exponential norm is defined by

$$\|X\|_{\psi_1} = \inf\left\{t > 0 : Ee^{\frac{|X|}{t}} \leq 2\right\}$$

- X is sub-exponential $\Longleftrightarrow \|X\|_{\psi_1} < \infty$.
- A random variable X is sub-gaussian if and only if X^2 is sub-exponential, moreover, $\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2$.
- $\|X\|_{\psi_2}^2 \leq \left\|X^2\right\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2$.

Properties

Proposition 1 (Sub-exponential is sub-gaussian squared)

A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover,

$$\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2$$

$\left\|X^2\right\|_{\psi_1}$ is the infimum of the numbers $K > 0$ satisfying $\mathbb{E}\exp\left(X^2/K\right) \leq 2$, while $\|X\|_{\psi_2}$ is the infimum of the numbers $L > 0$ satisfying $\mathbb{E}\exp\left(X^2/L^2\right) \leq 2$.

Properties

Proposition 2 (Product of sub-gaussians is sub-exponential)

Let X and Y be sub-gaussian random variables. Then XY is sub-exponential. Moreover,

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}$$

$$\begin{aligned}
\mathbb{E}\exp(|XY|) &\leq \mathbb{E}\exp\left(\frac{X^2}{2}+\frac{Y^2}{2}\right) \\
&= \mathbb{E}\left[\exp\left(\frac{X^2}{2}\right)\exp\left(\frac{Y^2}{2}\right)\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[\exp\left(X^2\right)+\exp\left(Y^2\right)\right] \qquad \text{(by Young's inequality)} \\
&= \frac{1}{2}(2+2) = 2
\end{aligned}$$

Bernstein's inequality

Theorem 3.9 (Bernstein's inequality)

Let $X_1, \ldots, X_N$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$, we have

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{\sum_{i=1}^{N}\|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right)\right]$$

where $c > 0$ is an absolute constant.

Bernstein's inequality

Theorem 3.10

Let $X_1, \ldots, X_N$ be independent, mean zero, sub-exponential random variables, and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, we have

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right]$$

where $K = \max_i \|X_i\|_{\psi_1}$.

Bernstein's inequality

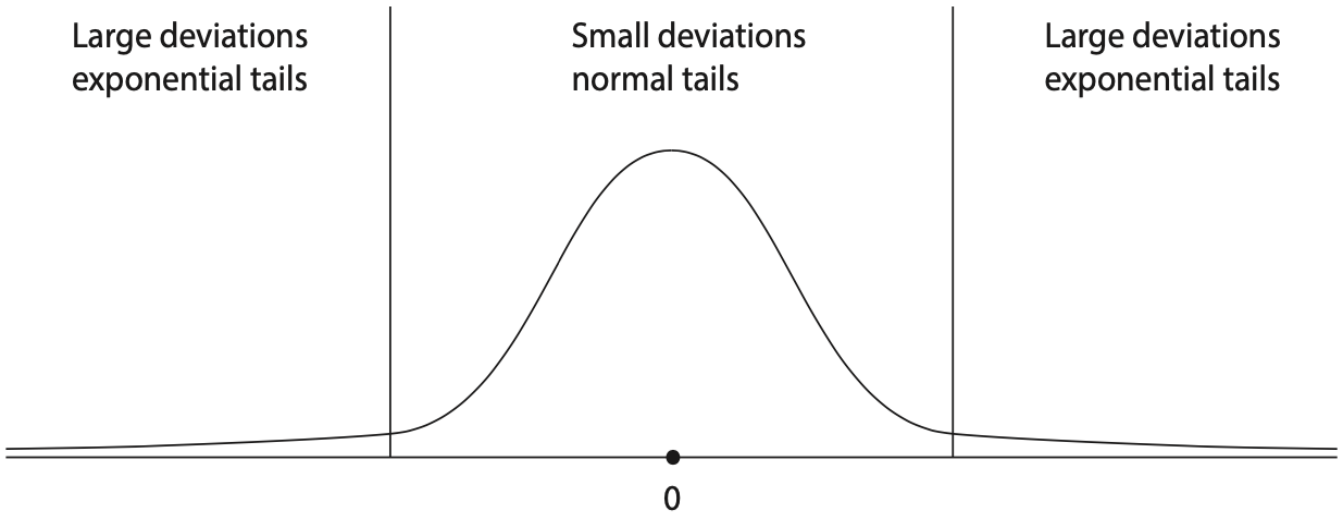

Figure 2.3 Bernstein's inequality for a sum of sub-exponential random variables gives a mixture of two tails: sub-gaussian for small deviations and sub-exponential for large deviations.