课程信息

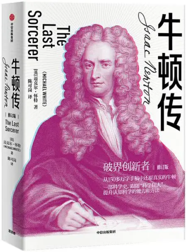
- **《随机过程》**：64学时，4学分，60230014（课号）
- **授课教师**：陈斌，信息大楼1608, cb17@tsinghua.org.cn
- **助教**：高英华、黄钰钧
- **成绩比例**：期中20%, 期末50%, 平时（作业+Project）30%
- **教材:《随机过程及其应用》陆大金**
- **其他参考书:**
  1. 李贤平，《概率论基础》，高等教育出版社
  2. 林元烈，《应用随机过程》，高等教育出版社
  3. R.Gallager, Stochastic Processes: Theory for Applications, Cambridge University Press
  4. S. Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press

Project要求

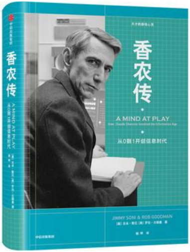
- 组队人数：3人以内，默认姓氏排序，除非特别说明贡献
- Topic（2选1）：
  - 参考Maryland大学课程中与课程相关的主题：http://www.cs.umd.edu/class/fall2020/cmsc828W/
  - 自选随机过程相关的主题；
- Reference数量不少于10篇；
- Tutorial Presentation: (40%)：
  - 内容：Motivation+ Theory+ Emperical Results（复现）+Conclusion+Thinking；
  - Q&A环节表现：任课老师和同学提问；
- Technical Report (姓名+学号）：(60%): English Writing in ICML Style (Latex模板在网络学堂) including Necessary Parts of An Academic Conference Paper.
- Deadline: 10.31，网络学堂提交，过期不能提交！

“应用”数学的基本素养与价值

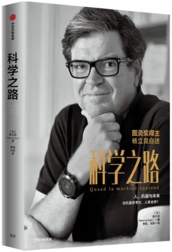
- **Mathematical Abstraction/ Theoretical Formulation**：学会用数学的语言表达；
- **Analog/Transfer Learning:** 迁移/类比的能力；
- **Empirical Observation/ Induction**； 发现现象/归纳能力；
- **Deductive Inference/Logical Implication**： 演绎/逻辑推理；
- **“牛逼”的“三个代表”**：



(a) Newton,1643-1727



(b) Shannon,1916-2001



(c) Lecun,1960-

第一章概率基础

陈斌

Tsinghua Shenzhen International Graduate School (SIGS)

Outline

基本定义

一个试验（或观察），若其结果预先无法确定，称之为**随机试验**。随机试验的可能结果成为**样本点**，记为 ω，样本点的全体构成**样本空间**，记为 Ω. 我们即将在 Ω 的某些子集上定义概率，但事先要对子集进行如下限制。

定义： 样本空间Ω的**某些子集**构成**事件域$\mathcal{F}$**，若$\mathcal{F}$满足

1. $\phi \in \mathcal{F}$;
2. $A \in \mathcal{F} \Rightarrow \bar{A} = \Omega \setminus A \in \mathcal{F}$;
3. $A_n \in \mathcal{F}, n \in \mathbb{N} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

则称**$\mathcal{F}$为σ域**，$(\Omega, \mathcal{F})$为**可测空间**。

定义：设$(\Omega,\mathcal{F})$为可测空间，若定义在$\mathcal{F}$上的**集函数P**满足：

1. $\forall A\in\mathcal{F}$, $P(A)\geqslant 0$; （非负性）
2. $P(\Omega)=1$; （规一性）
3. 设 $A_1,A_2,\ldots,\in\mathcal{F}$ 两两不相交，i.e., $A_iA_j\triangleq A_i\cap A_j=\phi$, $\forall i\neq j$, 则$P(\bigcup\limits_{n=1}^{\infty}A_i)=\sum\limits_{i=1}^{\infty}P(A_i)$. （可列可加性）

则称**P为概率**, $(\Omega,\mathcal{F},P)$为**概率空间**。

性质：

1. **不可能事件概率为0:** $P(\phi)=0$
2. **有限可加性:** $A_iA_j=\phi,\forall i\neq j\Rightarrow P(\bigcup\limits_{n=1}^{n}A_i)=\sum\limits_{i=1}^{n}P(A_i)$,
   $P(\bar{A})=1-P(A)$,
3. 若$B\subseteq A$，则$P(A-B)=P(A)-P(B)$, $P(B)\leq P(A)$ **(单调性)**;
4. $P(A\cup B)=P(A)+P(B)-P(AB)$;
   $P(A\cup B)\leq P(A)+P(B)$ **(Union Bound)**;
5. **全概率公式** 设 $A_1,A_2,\ldots,$ 为Ω的划分，即 A_i两两不相交
   且 $\bigcup_{i=1}^{\infty}A_i=\Omega$, 则 $P(B)=\sum_{i=1}^{n}P(A_i)P(B|A_i)$.

定义：设$(\Omega,\mathcal{F},P)$为概率空间，$\xi(\omega)$为定义在 Ω 上的**单值实函数**：$\xi:\Omega\to\mathbb{R}$。若$\forall x\in\mathbb{R}$，$\{\omega:\xi(\omega)\le x\}\in\mathcal{F}$，则称 $\xi(\omega)$ 为**随机变量**。

$$F(x)\triangleq P\{\xi(\omega)\le x\}$$称为随机变量 ξ 的**分布函数**。

性质（证明略）：

1. $0\le F(x)\le 1$，$F(x)$单调不减;
2. $F(-\infty)=\lim_{x\to-\infty}F(x)=0$，$F(+\infty)=1$;
3. $F(x)$右连续，且至多有可数个间断点.

离散型随机变量：状态的数目可数
连续型随机变量：存在概率密度函数$f(x)$使得$F'(x)=f(x)$

$$\text{状态}\triangleq\text{随机变量的取值}\begin{cases}\text{离散}\\ \text{连续}\end{cases}$$

常用分布（离散型）

检验概率分布：$p_i \geq 0, \sum p_i = 1.$

- 贝努力分布：
  $$P\{\xi = 1\} = p, \quad P\{\xi = 0\} = 1 - p, \quad 0 \leq p \leq 1.$$
- 二项分布：n次中恰好有i次成功
  $$P\{\xi = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \ldots, n,$$
  $$n \geq 1, \quad 0 \leq p \leq 1.$$
- 泊松分布:
  $$\lambda > 0, \quad P\{\xi = i\} = \frac{\lambda^i}{i!} e^{-\lambda}, \quad i = 0, 1, 2, \ldots$$
- 几何分布: 第i次首次成功
  $$0 < p < 1, \quad P\{\xi = i\} = (1-p)^{i-1} p, \quad i = 1, 2, \ldots$$

常用分布（连续性）

检验概率分布：$f(x) \geq 0$, $\int_{\mathbb{R}} f(x)\,dx = 1$.

- 指数分布：

$$\lambda > 0, \quad f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

- 均匀分布：

$$a < b, \quad U(a,b), \quad f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{其它}. \end{cases}$$

- 正态分布：

$$N(\mu, \sigma^2), \quad f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad N(0,1).$$

均值？方差？

数字特征

定义： **数学期望**（均值）

若$\int_{-\infty}^{+\infty} |x|\, dF(x) < +\infty$，$\quad \mu_\xi \triangleq E\xi \triangleq \int_{-\infty}^{+\infty} x\, dF(x).$

注： 针对离散型和连续型，统一用$dF(x)$来表示，若分开写 $E\xi$ 就分别是$\sum_n x_n \cdot p_n$ 和 $\int_{-\infty}^{+\infty} x f(x)\, dx.$

定义： **方差**（二阶矩）

若$\int_{-\infty}^{+\infty} x^2\, dF(x) < +\infty$，

$$\sigma_\xi^2 \triangleq D\xi \triangleq E[\xi - E\xi]^2 = E\xi^2 - E^2\xi.$$

定义： r阶绝对矩 $\quad E|\xi|^r \triangleq \int_{-\infty}^{+\infty} |x|^r\, dF(x).$

性质： 1). 线性； 2). $g(x)$函数，则 $E\,g(\xi)=\int_{-\infty}^{+\infty} g(x)\,dF(x)$.
几个例子：

- 贝努力分布：
  $P\{\xi=1\}=p$, $P\{\xi=0\}=1-p$, $\mu=p$, $\sigma^2=p(1-p)$;
- 二项分布：
  $P\{\xi=i\}=\binom{n}{i}p^i(1-p)^{n-i}$, $\mu=np$, $\sigma^2=np(1-p)$;
- 泊松分布:
  $\lambda>0,\quad P\{\xi=i\}=\frac{\lambda^i}{i!}e^{-\lambda}$, $\mu=\lambda$, $\sigma^2=\lambda$;

注： 二项分布为n个独立贝努力分布之和。

$$I_{\{\text{expression}\}}=\begin{cases}1, & \text{当表达式expression成立时},\\ 0, & \text{否则}.\end{cases}$$

- 指数分布：

  $f(x)=\lambda e^{-\lambda x}\cdot I_{\{x\geq 0\}}$, $F(x)=(1-e^{-\lambda x})I_{\{x\geq 0\}}$.
  $\mu=1/\lambda$, $\sigma^2=1/\lambda^2$.

- 正态分布：

  $N(\mu,\sigma^2)$, 高斯分布由均值和方差唯一确定。

- 其它分布：

  Γ分布、χ^2分布、Rayleigh分布、Rice分布等。

随机向量　复随机变量

定义： 设 $(\xi_1,\ldots,\xi_n)$ 为 n 维随机变量（随机向量）。

分布函数 $F(x_1,\ldots,x_n)\triangleq P\{\xi_1\le x_1,\ldots,\xi_n\le x_n\}$，

若 $\frac{\partial^n F(x_1,\ldots,x_n)}{\partial x_1\ldots\partial x_n}$ 存在，

$$F(x_1,\ldots,x_n)=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n} f(t_1,\ldots,t_n)\,dt_1\ldots dt_n.$$

定义： $(\xi_1,\ldots,\xi_n)$，协方差矩阵 $C\triangleq[C_{ij}]_{n\times n}$，其中

$C_{ij}\triangleq C(\xi_i,\xi_j)\triangleq E(\xi_i-E\xi_i)(\xi_j-E\xi_j)$称为$\xi_i$与$\xi_j$的协方差

C 对称，对角线 $c_{ii}=C(\xi_i,\xi_i)=D\xi_i=\sigma^2_{\xi_i}$.

二维随机变量 (η,ζ):

$$C(\eta,\zeta) \triangleq E(\eta-E\eta)(\zeta-E\zeta)=E\eta\zeta-E\eta E\zeta.$$

若$C(\eta,\zeta)=0$，称η与ζ**不相关**$\Longleftrightarrow$ $E\eta\zeta=E\eta\cdot E\zeta.$

若η,ζ独立，则η与ζ不相关。反之不然

$R(\eta,\zeta)\triangleq E\eta\zeta$ **相关函数**

$r\triangleq \frac{C(\eta,\zeta)}{\sigma_\eta\cdot\sigma_\zeta}$ 相关系数（标准化的协方差）

定义： $\xi\triangleq\eta+j\zeta,\quad j=\sqrt{-1}$称为$(\Omega,\mathcal{F},P)$上的**复随机变量**。

$$E\xi\triangleq E\eta+jE\zeta,\ D\xi\triangleq E|\xi-E\xi|^2=E(\xi-E\xi)\overline{(\xi-E\xi)}.$$

实质上，ξ是η与ζ组成的二维随机变量。$D\xi=D\eta+D\zeta$, 证明?

母函数

定义： ξ, $P\{\xi=k\}=p_k$, $k=0,1,2,\ldots$, **整值随机变量**

称 $G(s)\triangleq E\,s^k=\sum\limits_{k=0}^{\infty}p_k s^k$ 为 ξ 的**母函数**。

性质：

- $G(s)$在 $|s|\le 1$ 时，一致收敛且绝对收敛
- $p_k=G^k(0)/k!$ 反演公式或逆转公式
- $G(s)$与$F(x)$一一对应
- $\eta=a\xi+b\ (a>0,b\ge 0)\implies G_\eta(s)=s^bG(s^a)$.
- **可用来求数字特征:**

$$E\xi=G^{'}(1),\ E\xi(\xi-1)=G^{''}(1),\ D\xi=G^{''}(1)+G^{'}(1)-[G^{'}(1)]^2.$$

- 独立随机变量之和：

  $\xi_1, \ldots, \xi_n$ 相互独立，$G_1(s), \ldots, G_n(s)$, $\eta = \xi_1 + \cdots + \xi_n$.
  则$G_\eta(s) = G_1(s) \cdots G_n(s)$.　乘积

- 随机个 i.i.d.随机变量之和

  $\xi_1, \ldots, \xi_n$ 独立同分布，$G(s)$，整值随机变量 ν, $H(s)$,
  与 ξ_i 独立，$\eta = \xi_1 + \cdots + \xi_\nu$, 则 $G_\eta(s) = H[G(s)]$.　复合

母函数主要用来处理离散型随机变量。

- 求数字特征；
- 求独立随机变量之和；
- 与分布一一对应，且分析性质更好，可用来处理分布。

例：二项分布（独立贝努力分布之和）
　　泊松分布（独立泊松分布之和仍为泊松分布）

特征函数

定义： $\Phi(t) \triangleq E\,e^{jt\xi} = \int_{-\infty}^{+\infty} e^{jtx}\,dF(x) = \int_{-\infty}^{+\infty} e^{jtx} f(x)\,dx.$

直观：$f(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-jtx}\Phi(t)\,dt$

$\Phi(t)$与$f(x)$是一对Fourier变换。 $\begin{array}{l}\Phi(t): \mathbb{R}\to\mathbb{C}\\ f(x): \mathbb{R}\to\mathbb{R}\end{array}$ 1-1对应

性质：

- $\Phi(0)=1$, $|\Phi(t)|\le 1$, $\Phi(-t)=\overline{\Phi(t)}$.
- $\Phi(t)$在$(-\infty,+\infty)$一致连续.
- $\eta = a\xi + b \Rightarrow \Phi_\eta(t) = e^{jbt}\Phi_\xi(at)$.
- $\xi_1,\ldots,\xi_n$ 独立，$\eta = \xi_1+\cdots+\xi_n$，$\Phi_\eta(t) = \Phi_1(t)\cdots\Phi_n(t)$.

  证：$Ee^{jt\eta} = Ee^{jt(\xi_1+\cdots+\xi_n)} \overset{\text{独立}}{=} Ee^{jt\xi_1}\cdots Ee^{jt\xi_n}$.

- 若 ξ 的 n 阶绝对矩存在，则 $\forall k \le n$, $\Phi^k(0) = j^k E\xi^k$.
  证：$\Phi^{(k)}(0) = \int (jx)^k e^{jtx} dF\big|_{t=0} = j^k \int x^k dF$
- 例：$N(\mu, \sigma^2)$ 正态分布 $\quad \Phi(t) = \exp\left[jt\mu - \frac{t^2\sigma^2}{2}\right]$.
- **非负定性：**
  $\forall n \in \mathbb{N}$, $t_1, \ldots, t_n \in \mathbb{R}$, $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$（复数域），
  则 $\sum_{k=1}^{n} \sum_{i=1}^{n} \lambda_k \Phi(t_k - t_i) \overline{\lambda_i} \ge 0$.

证：

$$
\begin{aligned}
\text{左边} &= \sum_k \sum_i \int e^{j(t_k - t_i)x} dF \lambda_k \bar{\lambda}_l = \int \sum_k \lambda_k e^{jt_k x} \sum_i \bar{\lambda}_l e^{-jt_i x} dF \\
&= \int \left| \sum_k \lambda_k e^{jt_k x} \right|^2 dF \ge 0
\end{aligned}
$$

$\Phi(t)$ 非负定 $\Rightarrow$ 其Fourier变换为非负实值函数。
Bochner-Khintchine 定理，Herglotz定理

多维随机变量的特征函数

随机向量 $(\xi_1,\ldots,\xi_n)$，分布 $F(x_1,\ldots,x_n)$，密度 $f(x_1,\ldots,x_n)$，

特征函数

$$
\begin{aligned}
\Phi(t_1,\ldots,t_n) &\triangleq E\, e^{j(t_1\xi_1+\cdots+t_n\xi_n)} \\
&= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} \exp[jt_1x_1+\ldots+jt_nx_n]\, dF(x_1,\ldots,x_n).
\end{aligned}
$$

性质： 与一维情形类似

- $\Phi(t_1,\ldots,t_n)$ 在 $\mathbb{R}^n$ 中一致连续，
  $|\Phi(t_1,\ldots,t_n)|\le 1,\quad \Phi(-t_1,\ldots,-t_n)=\overline{\Phi(t_1,\ldots,t_n)}.$

- $\eta_i = \sigma_i \xi_i + a_i$, 其中 $\sigma_i, a_i \in \mathbb{R}$ 为常数, η 为 n-维随机向量,
  则 $\Phi_\eta(t_1, \ldots, t_n) = \exp(j \sum_{i=1}^n a_i t_i) \Phi_\xi(\sigma_1 t_1, \ldots, \sigma_n t_n)$.
- $(\xi_1, \ldots, \xi_n)$, $\eta = a_1 \xi_1 + \cdots + a_n \xi_n$, η 为1-维随机变量,
  则 $\Phi_\eta(t) = \Phi_\xi(a_1 t_1, \ldots, a_n t_n)$.
- $E\, \xi_1^{k_1} \cdots \xi_n^{k_n} = j^{-\sum_{i=1}^n k_i} \cdot \frac{\partial^{k_1+\cdots+k_n} \Phi(t_1,\ldots,t_n)}{\partial t_1^{k_1} \cdots \partial t_n^{k_n}} |_{t_1=\cdots=t_n=0}$.
- $(\xi_1, \ldots, \xi_n)$, $k < n$, $(\xi_1, \ldots, \xi_k)$, 边际分布的特征函数
  $$\Phi_k(t_1, \ldots, t_k) = \Phi(t_1, \ldots, t_k, 0, \ldots, 0).$$
- $(\xi_1, \ldots, \xi_n) \sim \Phi$, $\quad \xi_i \sim \Phi_i(t_i)$.
  则 ξ_i 两两独立 $\Leftrightarrow$ $\Phi(t_1, \ldots, t_n) = \Phi_1(t_1) \cdots \Phi_n(t_n)$.
- **独立与相关:** $E\xi\eta = E\xi \cdot E\eta$ $\Leftrightarrow$ 不相关.
  $E\, e^{jt_1\xi + jt_2\eta} = E\, e^{jt_1\xi} \cdot E\, e^{jt_2\eta}$ $\Leftrightarrow$ 独立.
  $F(x_1, x_2) = F_\xi(x_1) \cdot F_\eta(x_2)$ $\Leftrightarrow$ $f(x_1, x_2) = f_\xi(x_1) \cdot f_\eta(x_2)$.
- **一般:** 独立$\Rightarrow$ 不相关;
  特殊: 两个高斯随机变量ξ, η, 则ξ, η独立$\Leftrightarrow$$\xi, \eta$不相关;

离散分布的特征函数?

例: 求泊松分布的特征函数, 并计算期望和方差.

证:

$$\text{因为:}\ \Phi(t) = Ee^{jt\xi} = \sum_k e^{jtk}\frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}\sum_k \frac{\left(\lambda e^{jt}\right)^k}{k!} = e^{\lambda\left(e^{jt}-1\right)}$$

$$\text{则}\quad E\xi = \frac{1}{j}\Phi(0) = \lambda$$

$$E\xi^2 = -\Phi''(0) = \lambda^2 + \lambda$$

$$D\xi = E\xi^2 - E^2\xi = \lambda$$

特征函数更具有通用性, 且可以用来求解数字特征!

概率不等式及其应用

Empirical Average

- Let us look at 1D case.
- You have random variables $X_1, X_2, \ldots, X_N$
- Assume independently identically distributed i.i.d.
- This implies

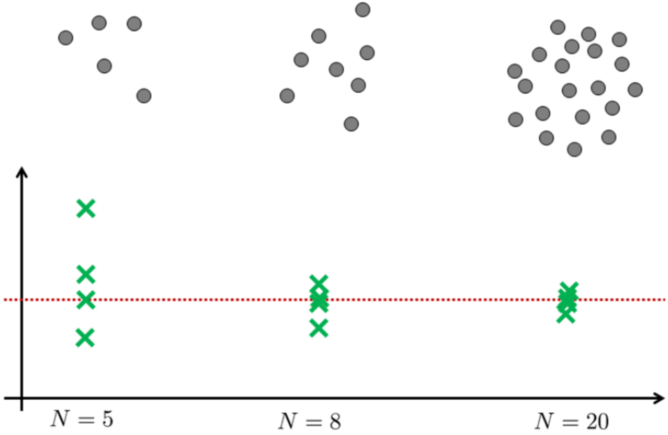
$$\mathbb{E}[X_1] = \mathbb{E}[X_2] = \ldots = \mathbb{E}[X_N] = \mu$$

- You compute the **empirical average**

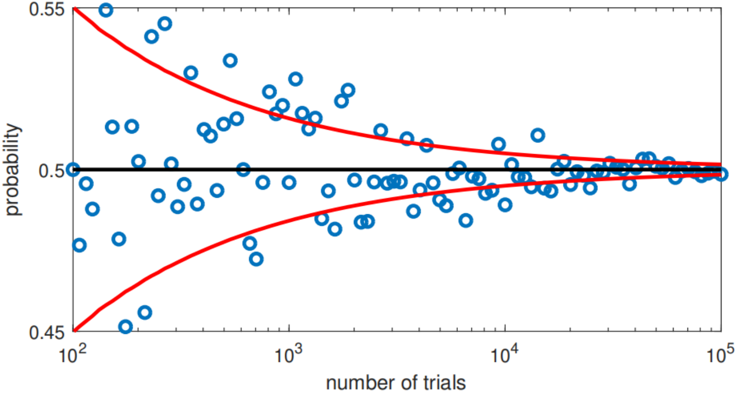
$$\nu = \frac{1}{N} \sum_{n=1}^{N} X_n$$

- How close is ν to μ?

As N grows ...



As N grows ...



Empirical Average: $\nu = \frac{1}{N}\sum_{n=1}^{N} X_n$

- ν is a random variable
- ν has CDF and PDF
- ν has mean:

$$\begin{aligned}\mathbb{E}[\nu] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[X_n\right] \\ &= \frac{1}{N}N\mu = \mu\end{aligned}$$

- Note that "$\mathbb{E}[\nu] = \mu$ " is not the same as " $\nu = \mu$".
- What is the probability ν deviates from μ?

$$\mathbb{P}[|\nu - \mu| > \epsilon] = ?$$

- The **Bad event:** $\mathcal{B} = \{|\nu - \mu| > \epsilon\}$: ν deviates from μ by at least ϵ
- $\mathbb{P}[\mathcal{B}]$ = probability that this bad event happens.
- Want $\mathbb{P}[\mathcal{B}]$ small. So upper bound it by δ

$$\mathbb{P}[|\nu - \mu| > \epsilon] \leq \delta$$

- With probability no greater than δ, **bad event** happens.
- Rearrange the equation:

$$\mathbb{P}[|\nu - \mu| \leq \epsilon] > 1 - \delta$$

- With probability at least $1 - \delta$, the **Bad** event will not happen.

Markov Inequality

Theorem (Markov Inequality)

For any $X > 0$ *and* $\epsilon > 0$

$$\mathbb{P}[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon}$$

Proof.

$$\begin{aligned}
\epsilon\mathbb{P}[X \geq \epsilon] &= \epsilon \int_{\epsilon}^{\infty} p(x)dx \\
&= \int_{\epsilon}^{\infty} \epsilon p(x)dx \\
&\leq \int_{\epsilon}^{\infty} xp(x)dx \\
&\leq \int_{0}^{\infty} xp(x)dx = \mathbb{E}[X]
\end{aligned}$$

$\square$

Chebyshev Inequality

Theorem (Chebyshev Inequality)

Let $X_1, \ldots, X_N$ *be i.i.d. with* $\mathbb{E}[X_n] = \mu$ *and* $\mathrm{Var}[X_n] = \sigma^2$. *Define*

$$\nu = \frac{1}{N} \sum_{n=1}^{N} X_n$$

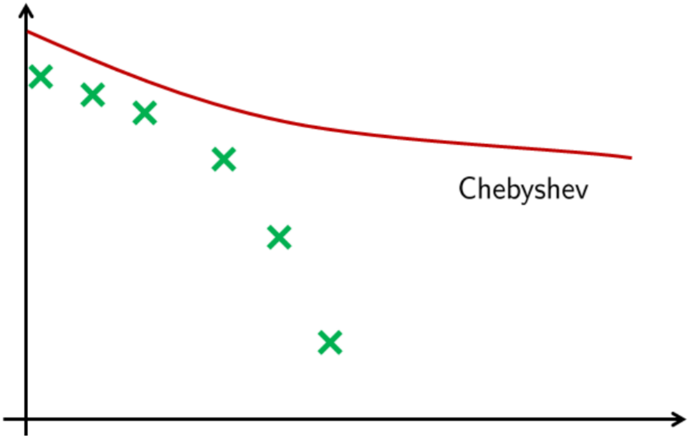
Then,

$$\mathbb{P}[|\nu - \mu| > \epsilon] \leq \frac{\sigma^2}{N\epsilon^2}$$

Proof.

$$\mathbb{P}\left[|\nu - \mu|^2 > \epsilon^2\right] \underbrace{\leq \frac{\mathbb{E}\left[|\nu - \mu|^2\right]}{\epsilon^2}}_{\text{Markov}} \underbrace{= \frac{\mathrm{Var}[\nu]}{\epsilon^2}}_{\mathbb{E}[(\nu-\mu)^2]=\mathrm{var}[\nu]} \underbrace{= \frac{\sigma^2}{N\epsilon^2}}_{\mathrm{var}[\nu]=\frac{\sigma^2}{N}} \qquad \square$$

How Good is Chebyshev Inequality?



Hoeffding Inequality

Let us revisit the Bad event:

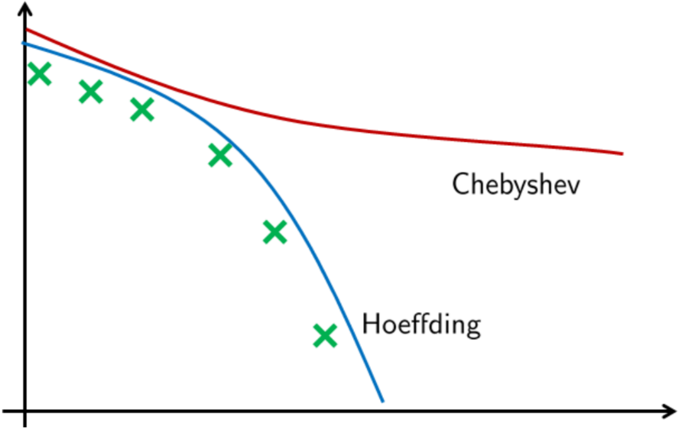
$$\begin{aligned}
\mathbb{P}[|\nu-\mu|\geq\epsilon] &= \mathbb{P}[\nu-\mu\geq\epsilon \quad \text{or} \quad \nu-\mu\leq-\epsilon] \\
&\leq \underbrace{\mathbb{P}[\nu-\mu\geq\epsilon]}_{\leq A}+\underbrace{\mathbb{P}[\nu-\mu\leq-\epsilon]}_{\leq A}, \qquad \text{Union bound} \\
&\leq 2A, \qquad \text{(What is } A \text{ ? To be discussed.)}
\end{aligned}$$

Theorem (Hoeffding Inequality)

Let $X_1,\ldots,X_N$ be random variables with $0\leq X_n\leq 1$, then

$$\mathbb{P}[|\nu-\mu|>\epsilon]\leq 2\underbrace{e^{-2\epsilon^2 N}}_{=A}$$

Chebyshev Inequality v.s. Hoeffding Inequality



Outline of Proof

Let us check **one side:**

$$\begin{aligned}\mathbb{P}[\nu-\mu\geq\epsilon] &= \mathbb{P}\left[\frac{1}{N}\sum_{n=1}^{N}X_n-\mu\geq\epsilon\right]=\mathbb{P}\left[\sum_{n=1}^{N}(X_n-\mu)\geq\epsilon N\right]\\ &= \mathbb{P}\left[e^{s\sum_{n=1}^{N}(X_n-\mu)}\geq e^{s\epsilon N}\right], \quad \forall s>0\\ &\leq \frac{\mathbb{E}\left[e^{s\sum_{n=1}^{N}(X_n-\mu)}\right]}{e^{s\epsilon N}}, \quad \text{Markov Inequality}\\ &= \left(\frac{\mathbb{E}\left[e^{s(X_n-\mu)}\right]}{e^{s\epsilon}}\right)^N, \quad \text{Independence}\end{aligned}$$

So now we have

$$\mathbb{P}[\nu-\mu\geq\epsilon]\leq\left(\frac{\mathbb{E}\left[e^{s(X_n-\mu)}\right]}{e^{s\epsilon}}\right)^N$$

Outline of Proof

Lemma (Hoeffding Lemma)

If $a \le X_n \le b$, *then*

$$\mathbb{E}\left[e^{s(X_n-\mu)}\right] \le e^{\frac{s^2(b-a)^2}{8}}$$

(**Proof Omitted, see [3. Appendix B]**)

This leads to

$$\mathbb{P}[\nu - \mu \ge \epsilon] = \left(\frac{\mathbb{E}\left[e^{s(X_n-\mu)}\right]}{e^{s\epsilon}}\right)^N \le \left(\frac{e^{\frac{s^2}{8}}}{e^{s\epsilon}}\right)^N$$

$$= e^{\frac{s^2 N}{8} - s\epsilon N}, \quad \forall s > 0.$$

Outline of Proof

Finally, we arrive at:

$$\mathbb{P}[\nu - \mu \geq \epsilon] \leq e^{\frac{s^2 N}{8} - s\epsilon N}$$

Since holds for **all** $s > 0$, in particular it holds for the minimizer:

$$\mathbb{P}[\nu - \mu \geq \epsilon] \leq e^{\frac{s_{\min}^2 N}{8} - s_{\min}\epsilon N} = \min_{s>0}\left\{ e^{\frac{\mathbf{s^2 N}}{\mathbf{8}} - \mathbf{s\epsilon N}} \right\}$$

Minimizing the exponent gives:
$\frac{d}{ds}\left\{ \frac{\mathbf{s^2 N}}{\mathbf{8}} - \mathbf{s\epsilon N} \right\} = \frac{sN}{4} - \epsilon N = 0$. So $s = 4\epsilon$, we have

$$\mathbb{P}[\nu - \mu \geq \epsilon] \leq e^{\frac{(4\epsilon)^2 N}{8} - (4\epsilon^2 N)} = e^{-2\epsilon^2 N}$$

Q(课后作业): What about another side $\mathbb{P}[\nu - \mu \leq -\epsilon]$?

$$\textbf{Chebyshev: } \mathbb{P}[|\nu - \mu| \geq \epsilon] \leq \frac{\sigma^2}{N\epsilon^2}.$$

$$\textbf{Hoeffding: } \mathbb{P}[|\nu - \mu| \geq \epsilon] \leq 2e^{-2\epsilon^2 N}$$

Both are in the form of

$$\mathbb{P}[|\nu - \mu| \geq \epsilon] \leq \delta$$

Equivalent to: For probability **at least** $1 - \delta$, we have

$$\mu - \epsilon \leq \nu \leq \mu + \epsilon$$

Error bar / Confidence interval of ν

$$\delta = \frac{\sigma^2}{N\epsilon^2} \Rightarrow \epsilon = \frac{\sigma}{\sqrt{\delta N}}, \quad \delta = 2e^{-2\epsilon^2 N} \Rightarrow \epsilon = \sqrt{\frac{1}{2N} \log \frac{2}{\delta}}$$

Chebyshev: For probability at least $1-\delta$, we have

$$\mu - \frac{\sigma}{\sqrt{\delta N}} \le \nu \le \mu + \frac{\sigma}{\sqrt{\delta N}}$$

Hoeffding: For probability at least $1-\delta$, we have

$$\mu - \sqrt{\frac{1}{2N}\log\frac{2}{\delta}} \le \nu \le \mu + \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}$$

Example:

- Alex: I have data $X_1, \ldots, X_N$. I want to estimate μ. How many data points N do I need?
- Bob: How much δ can you tolerate?
- Alex: Alright. I only have limited number of data points. How good my estimate is? (ϵ)
- Bob: How many data points N do you have?

Numerical Result

Chebyshev: For probability at least $1-\delta$, we have

$$\mu - \frac{\sigma}{\sqrt{\delta N}} \le \nu \le \mu + \frac{\sigma}{\sqrt{\delta N}}$$

Hoeffding: For probability at least $1-\delta$, we have

$$\mu - \sqrt{\frac{1}{2N}\log\frac{2}{\delta}} \le \nu \le \mu + \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}$$

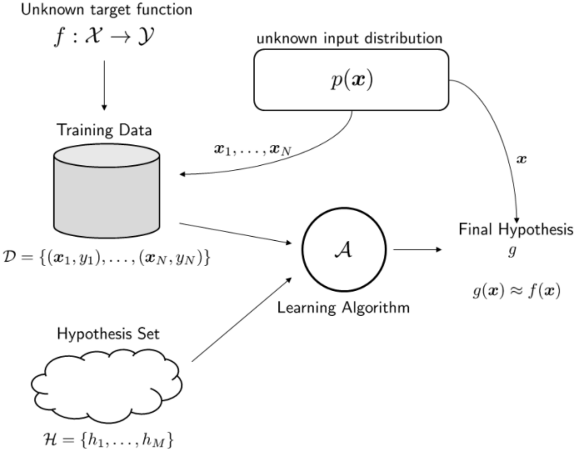
Let $\delta = 0.01, N = 10000, \sigma = 1$.

$$\epsilon = \frac{\sigma}{\sqrt{\delta N}} = 0.1, \quad \epsilon = \sqrt{\frac{1}{2N}\log\frac{2}{\delta}} = 0.016$$

Let $\delta = 0.01, \epsilon = 0.01, \sigma = 1$

$$N \ge \frac{\sigma^2}{\epsilon^2 \delta} = 1,000,000. \quad N \ge \frac{\log\frac{2}{\delta}}{2\epsilon^2} \approx 26,500$$

应用：机器学习的泛化


Unknown target function
$f : \mathcal{X} \to \mathcal{Y}$
unknown input distribution
$p(\boldsymbol{x})$
Training Data
$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$
$\boldsymbol{x}$
$\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$
$\mathcal{A}$
Final Hypothesis
g
$g(\boldsymbol{x}) \approx f(\boldsymbol{x})$
Learning Algorithm
Hypothesis Set
$\mathcal{H} = \{h_1, \ldots, h_M\}$

In-Sample Error

- Let $\boldsymbol{x}_n$ be a training sample
- h : Your hypothesis
- f : The unknown target function: **Oracle**
- If $h(\boldsymbol{x}_n) = f(\boldsymbol{x}_n)$, then say training sample $\boldsymbol{x}_n$ is **correctly classified**.

Definition (In-sample Error / Training Error)

Consider a training set $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, and a target function f. The in-sample error (or the training error) of a hypothesis function $h \in \mathcal{H}$ is the empirical average of $\{h(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n)\}$:

$$E_{\text{in}}(h) \overset{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(h(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n))$$

where $\mathbb{I}(\cdot) = 1$ if the statement inside is true, and $= 0$ otherwise.

Out-Sample Error

- Let $\boldsymbol{x}$ be a testing sample drawn from $p(\boldsymbol{x})$
- If $h(\boldsymbol{x}) = f(\boldsymbol{x})$, then say testing sample $\boldsymbol{x}$ is correctly classified.
- Since $\boldsymbol{x} \sim p(\boldsymbol{x})$, you need to compute the probability of error, called the out-sample error

Definition (Out-sample Error / Testing Error)

Consider an input space $\mathcal{X}$ containing elements $\boldsymbol{x}$ drawn from a distribution $p_{\boldsymbol{X}}(\boldsymbol{x})$, and a target function f. The out-sample error (or the testing error) of a hypothesis function $h \in \mathcal{H}$ is

$$E_{\text{out}}(h) \overset{\text{def}}{=} \mathbb{P}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})]$$

where $\mathbb{P}[\cdot]$ measures the probability of the statement based on the distribution $p_{\boldsymbol{X}}(\boldsymbol{x})$.

In-sample VS Out-sample

In-Sample Error:

$$E_{\text{in}}(h)=\frac{1}{N}\sum_{n=1}^{N}\mathbb{I}(h(\boldsymbol{x}_n)\neq f(\boldsymbol{x}_n))$$

Out-Sample Error:

$$\begin{aligned}E_{\text{out}}(h)=&\mathbb{P}[h(\boldsymbol{x})\neq f(\boldsymbol{x})]\\=&\underbrace{\mathbb{I}(h(\boldsymbol{x}_n)\neq f(\boldsymbol{x}_n))}_{=1}\mathbb{P}\{h(\boldsymbol{x}_n)\neq f(\boldsymbol{x}_n)\}\\&+\underbrace{\mathbb{I}(h(\boldsymbol{x}_n)=f(\boldsymbol{x}_n))}_{=0}(1-\mathbb{P}\{h(\boldsymbol{x}_n)\neq f(\boldsymbol{x}_n)\})\\=&\mathbb{E}\left\{\underbrace{\mathbb{I}(h(\boldsymbol{x}_n)\neq f(\boldsymbol{x}_n))}_{\text{贝努力分布！}}\right\}\end{aligned}$$

A Mathematical Tool

Beside in-sample and out-sample error, we also need a mathematical tool.

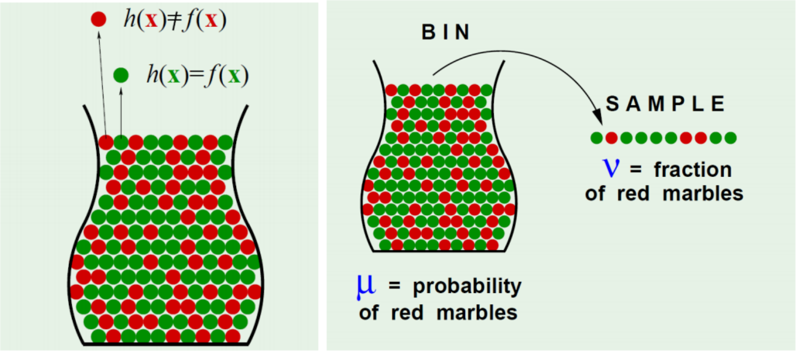
Theorem (Hoeffding Inequality)

Let $X_1, \ldots, X_N$ be random variables with $0 \leq X_n \leq 1$, then

$$\mathbb{P}[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

- We will use Hoeffding inequality to analyze the generalization error
- Hoeffding requires $0 \leq X_n \leq 1$
- $\nu = \frac{1}{N}\sum_{n=1}^{N} X_n$ is the empirical average
- Probability of how close ν compared to μ
- $\epsilon =$ tolerance level
- $N =$ number of samples

Applying Hoeffding Inequality to Our Problem



- $X_n = \mathbb{I}(h(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n))$: one sample training error $=$ either 0 or 1
- $\nu = E_{\text{in}} = \frac{1}{N}\sum_{n=1}^{N} X_n$: training error
- $\mu = E_{\text{out}}$: testing error

- Therefore, the inequality can be stated as

$$\mathbb{P}\left[|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}$$

- N = number of training samples
- ϵ = tolerance level
- Hoeffding is **applicable because $\mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))$ is either 1 or 0**.
- If you want to be more explicit, then

$$\mathbb{P}_{\boldsymbol{x}_n \sim \mathcal{D}}\left[\left|\frac{1}{N}\sum_{n=1}^{N}\mathbb{I}(h(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n)) - E_{\text{out}}(h)\right| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}$$

- The probability is evaluated **with respect to x_n drawn from the dataset $\mathcal{D}$**

Interpreting the Bound

- Let us look at the bound again:

$$\mathbb{P}\left[|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}$$

Message 1:

- You can bound $E_{\text{out}}(h)$ using $E_{\text{in}}(h)$.
- $E_{\text{in}}(h)$: You know. $E_{\text{out}}(h)$: You don't know, but you want to know.
- They are close if N is large.

Message 2 :

- The right hand side is independent of h and $p(\boldsymbol{x})$
- So it is a universal upper bound
- Works for any $\mathcal{A}$, any $\mathcal{H}$, any f, and any $p(\boldsymbol{x})$

Probably Approximately Correct (PAC)

- **Probably:** Quantify error using probability:

$$\mathbb{P}\left[|E_{\text{in}}(h) - E_{\text{out}}(h)| \leq \epsilon\right] \geq 1 - \delta$$

- **Approximately Correct:** In-sample error is an approximation of the out-sample error:

$$\mathbb{P}\left[|E_{\text{in}}(h) - E_{\text{out}}(h)| \leq \epsilon\right] \geq 1 - \delta$$

- If you can find an algorithm $\mathcal{A}$ such that for any ϵ and δ, there exists an N which can make the above inequality holds, then we say that the target function is **PAC-learnable**.

One Hypothesis versus the Final Hypothesis

- In this equation

$$\mathbb{P}\left[|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}$$

  the hypothesis h is **fixed**.
- This h is chosen **before** we look at the dataset.
- If h is chosen **after** we look at the dataset, then Hoeffding is **invalid**.
- We have to choose a h from $\mathcal{H}$ **during the learning process.**
- The h we choose **depends on $\mathcal{D}$**, i.e., This h is the **final hypothesis g**.
- When you need to choose g from $h_1, \ldots, h_M$, you need to **repeat Hoeffding M times**.

The Factor “M”

You can show that

$$
\begin{aligned}
|E_{\text{in}}(g)-E_{\text{out}}(g)|>\epsilon \Longrightarrow & \quad |E_{\text{in}}(h_1)-E_{\text{out}}(h_1)|>\epsilon \\
\text{or} & \quad |E_{\text{in}}(h_2)-E_{\text{out}}(h_2)|>\epsilon \\
\ldots & \\
\text{or} & \quad |E_{\text{in}}(h_M)-E_{\text{out}}(h_M)|>\epsilon
\end{aligned}
$$

- To have g, you need to consider $h_1, \ldots, h_M$
- You don’t know which h_m to pick; So it is a "OR"
- So there is a sequence of "OR"

The Factor “M”

$$
\begin{aligned}
\mathbb{P}\{|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\} \overset{(a)}{\leq} & \quad \mathbb{P}\{|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon \\
\text{or} & \quad |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)| > \epsilon \\
& \quad \ldots \\
\text{or} & \quad |E_{\text{in}}(h_M) - E_{\text{out}}(h_M)| > \epsilon\} \\
\overset{(b)}{\leq} & \quad \sum_{m=1}^{M} \mathbb{P}\{|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon\}
\end{aligned}
$$

- We need two identities

- (a) If-statement. $\mathbb{P}[A] \leq \mathbb{P}[B]$ if $A \subseteq B$
- (b) Union Bound. $\mathbb{P}[A \text{ or } B] \leq \mathbb{P}[A] + \mathbb{P}[B]$

The Factor “M”

- Change this equation

$$\mathbb{P}\left\{\left|E_{\text{in}}\left(h\right)-E_{\text{out}}\left(h\right)\right|>\epsilon\right\}\leq 2e^{-2\epsilon^2 N}$$

- to this equation

$$\mathbb{P}\left\{\left|E_{\text{in}}(g)-E_{\text{out}}(g)\right|>\epsilon\right\}\leq 2Me^{-2\epsilon^2 N}$$

- So what? M is a constant.
- **Bad news:** M can be large, or even ∞, e.g., A linear regression has $M=\infty$.
- **Good news:** It is possible to bound M in machine learning.

Learning Goal

- The ultimate goal of learning is to make

$$E_{\text{out}}(g) \approx 0$$

- To achieve this we need

$$E_{\text{out}}(g) \underbrace{\approx}_{\text{Hoeffding Inequality}} E_{\text{in}}(g) \underbrace{\approx}_{\text{Training Error}} 0$$

- Hoeffding inequality holds when N **is large**;
- Training error is small when you train well;

Rewriting the Hoeffding Inequality

- Recall the Hoeffding Inequality

$$\mathbb{P}\left\{|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\right\} \leq 2Me^{-2\epsilon^2 N}$$

- This is the same as

$$\mathbb{P}\left\{|E_{\text{in}}(g) - E_{\text{out}}(g)| \leq \epsilon\right\} > 1 - \delta$$

- Equivalently, we can say: with probability $1 - \delta$,

$$E_{\text{in}}(g) - \epsilon \leq E_{\text{out}}(g) \leq E_{\text{in}}(g) + \epsilon$$

Generalization Bound

- Move around the terms, then we have

$$2Me^{-2\epsilon^2 N} = \delta \Rightarrow \epsilon = \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}}$$

- Plug this result into the previous bound:

$$E_{\text{in}}(g) - \epsilon \leq E_{\text{out}}(g) \leq E_{\text{in}}(g) + \epsilon$$

- This gives us

$$E_{\text{in}}(g) - \sqrt{\frac{\mathbf{1}}{\mathbf{2N}}\log\frac{\mathbf{2M}}{\delta}} \leq E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{\mathbf{1}}{\mathbf{2N}}\log\frac{\mathbf{2M}}{\delta}}$$

- This is called the **generalization bound**.
- Many unsolved problems in Deep Learning Generalization. (**Interesting Project Topic!**)