不等式

- 不等式
  - 有些量很难计算，不等式可以对这些量给出一个界
  - 不等式也是下一章讨论收敛理论的基础

  - 关于概率的不等式
    - Markov不等式
    - Chebyshev不等式
    - Hoeffding不等式

  - 关于期望的不等式
    - Cauchy-Schwarze不等式
    - Jensen不等式

Markov不等式

- 4.1 定理（ Markov不等式）：令X为非负随机变量且假设$\mathbb{E}(X)$存在，则对任意 $t>0$，有

$$P(X>t)\le\frac{\mathbb{E}(X)}{t} \tag{4.1}$$

- 当 $t=k\mu,\mu=\mathbb{E}(X)$，$\mathbb{P}(X>k\mu)\le\frac{1}{k}$
  - 当 k>1时，表示随机变量的取值离不会期望不会太远（离期望较远的概率很小，小于$1/k$ ）
    - $\mathbb{P}(X>2\mu)\le 0.5$, $\mathbb{P}(X>3\mu)\le 0.33$
  - 当$0<k\le 1$ 时，$1/k\ge 1$，上式总是成立表示（$\mathbb{P}(A)\le 1$）

ningMarkov 不等式证明：X 为非负随机变量，对任何 $t>0$，有 $\mathbb{P}(X>t)\leq\dfrac{\mathbb{E}(X)}{t}$

证明：$\because\ X>0$

$\therefore\ \mathbb{E}(X)=\int_0^{\infty}xf(x)\,dx$ （x 非负，所以积分区间从 0 开始）

$=\int_0^{t}xf(x)\,dx+\int_t^{\infty}xf(x)\,dx$ （分成两段积分）

$\geq\int_t^{\infty}xf(x)\,dx$ （$x\geq 0,t>0,\therefore\int_0^{t}xf(x)\,dx>0$，求和部分去掉正的一部分，不等号成立）

$\geq t\int_t^{\infty}f(x)\,dx$ （将 t 放到积分符号外面，相当于令 由于 不等号成立）

$=t\mathbb{P}(X>t)$

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

Markov不等式

- 将X换成满足条件的$r(X)$，上述结论也成立！

$$\mathbb{P}\left(r\left(X\right) > t\right) \leq \frac{\mathbb{E}\left(r\left(X\right)\right)}{t}$$

- 当 $r\left(X\right) = \frac{\left(X - \mu\right)^2}{\sigma^2}$ ？
  - Chebyshev不等式：Markov不等式的应用

Chebyshev不等式

- 4.2 定理（Chebyshev不等式）：令$\mu = \mathbb{E}(X), \sigma^2 = \mathbb{V}(X)$
- 则 $\mathbb{P}(|X-\mu| \geq t) \leq \dfrac{\sigma^2}{t^2} \qquad \mathbb{P}(|Z| \geq k) \leq \dfrac{1}{k^2}$
- 其中 $Z = (X-\mu)/\sigma$
- $\mathbb{P}(|Z| > 2) \leq 1/4, \mathbb{P}(|Z| > 3) \leq 1/9.$
- X在其期望附近（t邻域）的概率与方差σ^2有关
  - σ^2越大，随机变量远离期望的概率越大（方差用于度量随机变量围绕均值的散布程度）
  - σ^2越小，随机变量在期望附近，远离期望的概率越小
  - 可用来证明样本均值会在其期望附件（样本数越多越接近，因为样本方差随n增大而减小）

Chebyshev 不等式证明：令 $\mathbb{E}(X)=\mu, \mathbb{V}(X)=\sigma^2$，则 $\mathbb{P}\left(\left|X-\mu\right|\geq t\right)\leq\dfrac{\sigma^2}{t^2}$

证明：$\mathbb{P}\left(\left|X-\mu\right|\geq t\right)=\mathbb{P}\left(\left(X-\mu\right)^2\geq t^2\right)$ （两边同时平方）

$$\leq\frac{\mathbb{E}\left(\left(X-\mu\right)^2\right)}{t^2}$$ （Markov 不等式）

$$\leq\frac{\sigma^2}{t^2}$$

当 $t=k\sigma$，则 $\mathbb{P}\left(\left|X-\mu\right|\geq k\sigma\right)=\mathbb{P}\left(\dfrac{\left|X-\mu\right|}{\sigma}\geq k\right)\leq\dfrac{\sigma^2}{\left(k\sigma\right)^2}=\dfrac{1}{k^2}$

Chebyshev不等式

- X在其期望附近（t邻域）的概率与方差σ^2有关
- 另外一个变形：

$$\mathbb{P}\left(\left|X-\mu\right|\geq k\sigma\right)\leq\frac{1}{k^2}$$

  - k=2？ $\mathbb{P}\left(\left|X-\mu\right|\geq 2\sigma\right)\leq 0.25$
  - k=3？ $\mathbb{P}\left(\left|X-\mu\right|\geq 3\sigma\right)\leq 0.11$
    - 高斯分布为0.9997

- 这个界很松，因为Chebyshev不等式没有限定分布的形式，所以应用广泛
  - 对某些具体的分布来说，可以得到更紧致的界，如高斯分布 $Z\sim N(0,1)$

$$\mathbb{P}\left(Z\geq t\right)=\frac{1}{\sqrt{2\pi}}\int_t^\infty xe^{-x^2/2}dx\leq\frac{1}{\sqrt{2\pi}}\int_t^\infty\frac{x}{t}e^{-x^2/2}dx=\frac{1}{\sqrt{2\pi}}\frac{e^{-t^2/2}}{t}$$

Mill’s inequality

$$\mathbb{P}\left(\left|Z\right|\geq t\right)=2\mathbb{P}\left(Z\geq t\right)\leq\sqrt{\frac{2}{\pi}}\frac{e^{-t^2/2}}{t}$$

Chebyshev不等式

- 4.3例：假设我们在一个有n个测试样本的测试集上测试一个预测方法（以神经网络为例）。若预测错误置 $X_i=1$ 预测正确则置 $X_i=0$。则 $\bar{X}_n=n^{-1}\sum_{i=1}^{n}X_i$ 为观测到的错误率。每个 X_i 可视为有未知均值 p 的Bernoulli分布。我们想知道真正的错误率 p。
- 直观地，我们希望 $\bar{X}_n$ 接近 p。但 $\bar{X}_n$ 有多大可能不在 p 的 ε 邻域内？
- $\mathbb{V}\left(\bar{X}_n\right)=\mathbb{V}\left(X_1\right)/n=p\left(1-p\right)/n,$

$$\mathbb{P}\left(\left|\bar{X}_n-p\right|\geq\varepsilon\right)\leq\frac{\mathbb{V}\left(\bar{X}_n\right)}{\varepsilon^2}=\frac{p\left(1-p\right)}{n\varepsilon^2}\leq\frac{1}{4n\varepsilon^2}.$$

- 由于对任意 p 有 $p\left(1-p\right)\leq\frac{1}{4}$，所以当 $\varepsilon=0.2,\ n=100$ 时，边界为0.0625。

例2：依分布收敛

- 考虑随机序列 ，其中
- 直观： 集中在0处， 收敛到0
- 但

（Chebyshev不等式）

Hoeffding不等式

- 作用与Chebyshev不等式类似，但区间更紧致（增加了独立性约束）
- 4.4 定理（ Hoeffding不等式）：设 $Y_1, ... Y_n$ 相互独立，且 $\mathbb{E}(Y_i)=0$, 且 $a_i \le Y_i \le b_i$ 令 $\varepsilon > 0$ 则对任意 $t > 0$

$$Y_i = \frac{1}{n}(X_i - p) \qquad \mathbb{P}\left(\sum_{i=1}^{n} Y_i \ge \varepsilon\right) \le e^{-t\varepsilon} \prod_{i=1}^{n} e^{t^2(b_i - a_i)^2/8}$$

- 4.5 定理（ Hoeffding不等式）：令 $X_1, ... X_n \sim Bernoulli(p)$ 则对任意 $\varepsilon > 0$，有

$$\mathbb{P}\left(\left|\bar{X}_n - p\right| > \varepsilon\right) \le 2e^{-2n\varepsilon^2}$$

- 其中 $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$

Hoeffding 不等式证明：

证明：$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq \varepsilon\right) = \mathbb{P}\left(t\sum_{i=1}^{n} Y_i \geq t\varepsilon\right) = \mathbb{P}\left(e^{t\sum_{i=1}^{n} Y_i} \geq e^{t\varepsilon}\right) \leq e^{-t\varepsilon}\mathbb{E}\left(e^{t\sum_{i=1}^{n} Y_i}\right)$ （Markov 不等式）

$$= e^{-t\varepsilon}\prod_{i=1}^{n}\mathbb{E}\left(e^{tY_i}\right) \qquad （1）$$

$\because\ a_i \leq Y_i \leq b_i$，$\therefore Y_i = \alpha b_i + (1-\alpha)a_i$，$\alpha = (Y_i - a_i)/(b_i - a_i)$

由于　　为凸函数，所以 $e^{tY_i} \leq \frac{(Y_i - a_i)}{(b_i - a_i)}e^{tb_i} + \frac{(b_i - Y_i)}{(b_i - a_i)}e^{ta_i}$

$\mathbb{E}\left(e^{tY_i}\right) \leq -\frac{a_i}{(b_i - a_i)}e^{tb_i} + \frac{b_i}{(b_i - a_i)}e^{ta_i} = e^{g(u)}$ （$\mathbb{E}(Y_i) = 0$）

其中 $u = t(b_i - a_i), g(u) = -\gamma a_i + \log\left(1 - \gamma + \gamma e^{u}\right), \gamma = -a_i / b_i - a_i$

$g(0) = g'(0) = 0, g''(u) \leq 1/4,\ for\ u > 0$

根据 Talayor 展开，$g(u) = g(0) + ug'(0) + u^2 g''(\xi) = \frac{u^2}{2}g''(\xi) \leq \frac{1}{8}u^2 = \frac{1}{8}t^2(b_i - a_i)^2$

所以 $\mathbb{E}\left(e^{tY_i}\right) < e^{g(u)} < e^{t^2(b_i - a_i)^2/8}$，代入（1），不等式得证。

Hoeffding不等式

- 4.6 例：令 $X_1,\ldots X_n \sim Bernoulli(p)\quad n=100,\ \varepsilon=0.2$
- 则根据Chebyshev不等式，有

$$\boxed{\mathbb{P}\left(\left|\overline{X}_n - p\right| \geq \varepsilon\right) \leq \mathbb{V}\left(\overline{X}_n\right)\big/\varepsilon^2}$$

$$\mathbb{P}\left(\left|\overline{X_n} - p\right| > \varepsilon\right) \leq 0.0625.$$

- 根据Hoeffding不等式，有

$$\boxed{\mathbb{P}\left(\left|\overline{X}_n - p\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}}$$

$$\mathbb{P}\left(\left|\overline{X_n} - p\right| > .2\right) \leq 2e^{-2(100)(.2)^2} = 0.00067.$$

- 结果远远小于0.0625。

Hoeffding不等式

- 可用来计算二项分布中的参数p的置信区间
- 对给定的 $\alpha > 0$ ，令

$$\varepsilon_n = \left\{ \frac{1}{2n} \log\left(\frac{2}{\alpha} \right) \right\}^{\frac{1}{2}}$$

- 则根据Hoeffding不等式

$$\mathbb{P}\left(\left| \overline{X_n} - p \right| > \varepsilon_n \right) \leq 2e^{-2n\varepsilon_n^2} = \alpha$$

- 令 $C = \left(\overline{X_n} - \varepsilon, \overline{X_n} + \varepsilon \right)$ ，则 $\mathbb{P}(C \notin p) = \mathbb{P}\left(\left| X_n - p \right| > \varepsilon \right) \leq \alpha$
- 则 $\mathbb{P}(C \in p) \geq 1 - \alpha$ 。
- 称C为 $1-\alpha$ 置信区间。

Cauchy-Schwarze不等式

- 4.8 定理（ Cauchy-Schwarze不等式）：若X、Y是有限方差，则

$$\mathbb{E}\left(\left|XY\right|\right) \le \sqrt{\mathbb{E}\left(X^2\right)\mathbb{E}\left(Y^2\right)}$$

- 例：协方差不等式

$$Cov\left(X,Y\right) = \mathbb{E}\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right] \le \sqrt{\mathbb{E}\left(\left(X-\mu_X\right)^2\right)\mathbb{E}\left(\left(Y-\mu_Y\right)^2\right)} = \sigma_X\sigma_Y$$

$$\left(Cov\left(X,Y\right)\right)^2 \le \sigma_X^2\sigma_Y^2$$

$$-1 \le \rho = \frac{\mathrm{Cov}\left(X,Y\right)}{\sigma_X\sigma_Y} \le 1$$

Jensen不等式

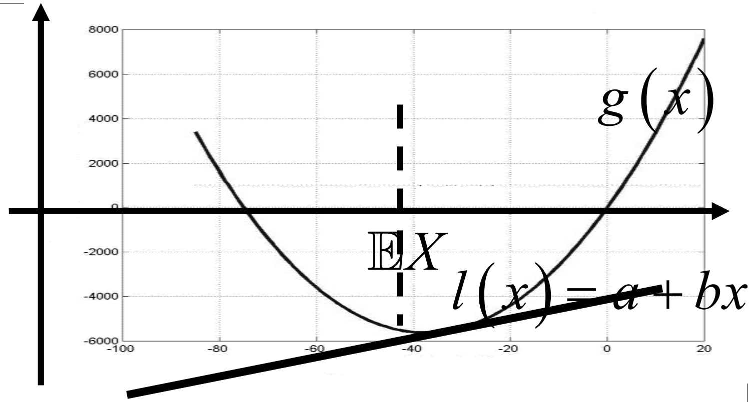
- 4.9 定理（ Jensen不等式）：如果g是凸的，则

$$g\left(X\right)\geq g\left(\mathbb{E}\left(X\right)\right)$$

- 如果g是凹的，则

$$\mathbb{E}g\left(X\right)\leq g\left(\mathbb{E}\left(X\right)\right)$$

- $g\left(x\right)=x^2 \Rightarrow \mathbb{E}\left(X^2\right)\geq\left(\mathbb{E}X\right)^2$
- $g\left(x\right)=\dfrac{1}{x} \Rightarrow \mathbb{E}\left(\dfrac{1}{X}\right)\geq\dfrac{1}{\mathbb{E}X}$
- $g\left(x\right)=\log x \Rightarrow \mathbb{E}\left(\log X\right)\leq\log\left(\mathbb{E}X\right)$



证明： $g(x)$ 为凸函数，则 $\mathbb{E}g(X)\geq g(\mathbb{E}X)$

证明：令在点 $x=\mathbb{E}X$ 处，函数 $g(x)$ 的切线为 $l(x)=a+bx$

由于 $g(x)$ 为凸函数， $g(x)$ 位于其切线之上，即

$$g(x)\geq l(x)=a+bx$$

所以 $\mathbb{E}g(X)\geq\mathbb{E}(a+bX)$ （两边同区期望，不等式仍然成立）

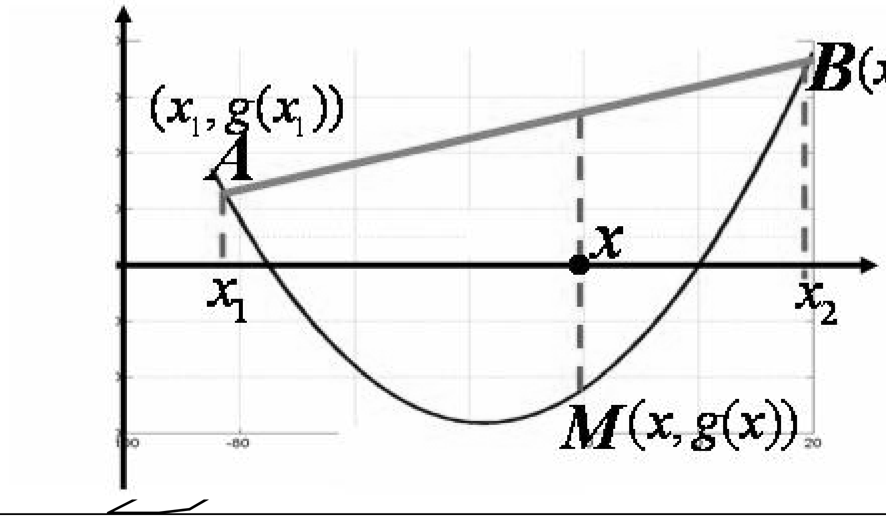
$=a+b\mathbb{E}X$ （期望的线性性质）

$=l(\mathbb{E}X)$ （l 的定义）

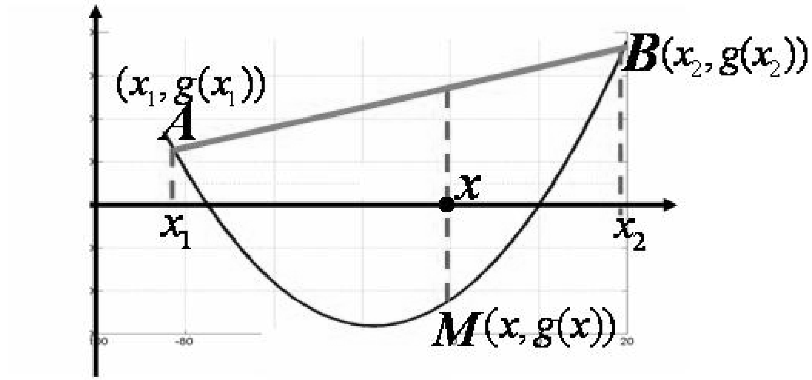
$=g(\mathbb{E}X)$ （在切点 $x=\mathbb{E}X$ 处，l 等于函数值）

凸函数

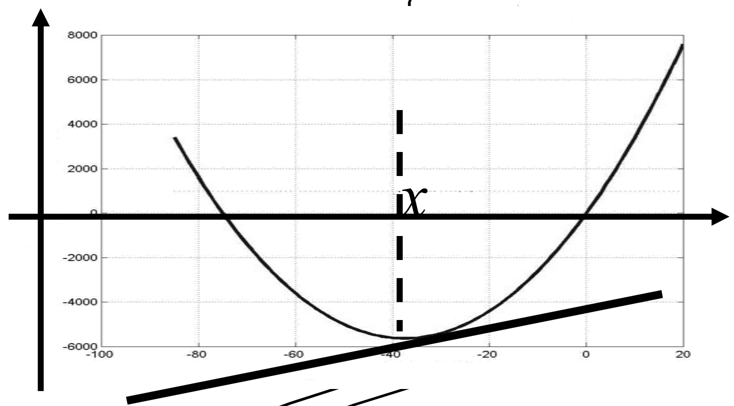
- 如果对所有的$x, y,\ 0<\lambda<1$ ，满足
  $g(\lambda x+(1-\lambda)y)\le \lambda g(x)+(1-\lambda)g(y)$
- 则函数$g(x)$为凸函数（convex），$-g(x)$为凹函数（concave）
  - 凸：装水，如 $g(x)=x^2$
  - 凹：溢出水，如 $g(x)=\log x$



凸函数



- 几何意义
  - 连接 $(a,g(a)),(b,g(b))$两点的弦，永远在 $y=g(x)$ 之上
  - 凸光滑函数上任一点的切线在曲线的下方



思考：随机变量序列的收敛性

- 随机样本：IID样本 $X_1, X_2..., X_n$ ， $X_i \sim F$
- 统计量：对随机样本概述

$$Y = T(X_1, X_2..., X_n)$$

  - Y为随机变量，Y的分布称为统计量的采样分布
  - 如：样本均值、样本方差、样本中值

- 收敛性：当样本数量n趋向无穷大时，统计量的变化
  - 大样本理论、极限定理、渐近理论