信息科学中的数学　课程讲义

周书予

2000013060@stu.pku.edu.cn

2022 年 1 月 19 日

1　High-Dimensional Space

定理 1 (Markov's Inequality). 设 x 为一非负随机变量，则对于任意 $a>0$，有

$$\Pr[x \geqslant a] \leqslant \frac{\mathbb{E}[x]}{a} \tag{1}$$

证明 1. 以连续形式为例，假设 x 的概率密度为 p。

$$\mathbb{E}[x] = \int_0^{\infty} xp(x)\mathrm{d}x \geqslant \int_a^{\infty} xp(x)\mathrm{d}x \geqslant a\int_a^{\infty} p(x)\mathrm{d}x = a\Pr[x \geqslant a] \tag{2}$$

定理 2 (Chebyshev's Inequality). 设 x 为一随机变量，则对于任意 $c>0$，有

$$\Pr[|x-\mathbb{E}[x]| \geqslant c] \leqslant \frac{\mathrm{Var}[x]}{c^2} \tag{3}$$

证明 2. 考虑随机变量 $y=|x-\mathbb{E}[x]|^2$ 非负，且 $\mathbb{E}[y]=\mathrm{Var}[x]$，故对 y 考虑 Markov's Inequality。

$$\Pr[|x-\mathbb{E}[x]| \geqslant c] = \Pr\left[y \geqslant c^2\right] \leqslant \frac{\mathbb{E}[y]}{c^2} = \frac{\mathrm{Var}[x]}{c^2} \tag{4}$$

定理 3 (大数定理). 令 $x_1,x_2,\cdots,x_n$ 为对随机变量 x 的 n 次独立随机采样，则

$$\Pr\left[\left|\frac{x_1+x_2+\cdots+x_n}{n}-\mathbb{E}[x]\right| \geqslant \varepsilon\right] \leqslant \frac{\mathrm{Var}[x]}{n\varepsilon^2} \tag{5}$$

定理 4 (体积集中在表面). 高维球的体积集中在表面。这是因为

$$\frac{\mathrm{volume}((1-\varepsilon)A)}{\mathrm{volume}(A)} = (1-\varepsilon)^d \leqslant \mathrm{e}^{-\varepsilon d} \to 0 \quad (d\to\infty) \tag{6}$$

定理 5 (高维球的体积与表面积公式). 用 $V(d)$ 和 $A(d)$ 来表示 d 维单位球的体积与表面积，则

$$V(d) = \frac{2\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2})} \qquad A(d) = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \tag{7}$$

证明 3. 用 $V(d,r)$ 和 $A(d,r)$ 表示 d 维空间中半径为 r 的球的体积与表面积，则

$$V(d) = \int_{r=0}^{1} A(d,r)\mathrm{d}r = A(d)\int_{r=0}^{1} r^{d-1}\mathrm{d}r = \frac{A(d)}{d} \tag{8}$$

于是接下来只考虑计算 $A(d)$。考虑如下积分

$$I(d) = \int_{x_1\in\mathbb{R}}\int_{x_2\in\mathbb{R}}\cdots\int_{x_d\in\mathbb{R}} \mathrm{e}^{-(x_1^2+x_2^2+\cdots+x_d^2)}\mathrm{d}x_d\cdots\mathrm{d}x_2\mathrm{d}x_1 \tag{9}$$

一方面，每个 x_i 是独立的，因此结果就是 d 个乘起来。

$$I(d) = \left[\int_{x\in\mathbb{R}} \mathrm{e}^{-x^2}\mathrm{d}x\right]^d = \sqrt{\pi}^d = \pi^{\frac{d}{2}} \tag{10}$$

另一方面，这个积分可以理解为给 d 维空间中的每个点附上了一个只与“到原点距离”有关的权重，因此可以考虑枚举“到原点距离” r 计算。

$$I(d)=\int_{r=0}^{\infty}A(d,r)\mathrm{e}^{-r^2}\mathrm{d}r=A(d)\int_{r=0}^{\infty}r^{d-1}\mathrm{e}^{-r^2}\mathrm{d}r\xlongequal{t=r^2}\frac{A(d)}{2}\int_{t=0}^{\infty}t^{\frac{d}{2}-1}\mathrm{e}^{-t}\mathrm{d}t=\frac{A(d)}{2}\Gamma(\frac{d}{2})\tag{11}$$

结合两者结果即可得到 $A(d)=\frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}$，于是 $V(d)=\frac{2\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2})}$。

定理 6 (**体积集中在赤道**). 对于 $c\geqslant 1$ 以及 $d\geqslant 3$，d 维单位球有至少 $1-\frac{2}{c}\mathrm{e}^{-c^2/2}$ 的体积满足 $|x_1|\leqslant\frac{c}{\sqrt{d-1}}$。

证明 4. 记 A 表示单位球 $x_1\geqslant\frac{c}{\sqrt{d-1}}$ 的部分，H 表示半球，需要证明

$$\frac{\text{volume}(A)}{\text{volume}(H)}\leqslant\frac{\text{upper bound volume}(A)}{\text{lower bound volume}(H)}=\frac{2}{c}\mathrm{e}^{-c^2/2}\tag{12}$$

$$\begin{aligned}\text{volume}(A)&=\int_{\frac{c}{\sqrt{d-1}}}^{1}V(d-1,\sqrt{1-x^2})\mathrm{d}x\leqslant V(d-1)\int_{\frac{c}{\sqrt{d-1}}}^{1}\mathrm{e}^{-\frac{d-1}{2}x^2}\mathrm{d}x\\&\leqslant V(d-1)\frac{\sqrt{d-1}}{c}\int_{\frac{c}{\sqrt{d-1}}}^{1}x\mathrm{e}^{-\frac{d-1}{2}x^2}\mathrm{d}x=\frac{V(d-1)}{c\sqrt{d-1}}\mathrm{e}^{-c^2/2}\end{aligned}\tag{13}$$

$$\text{volume}(H)\geqslant V\left(d-1,\sqrt{1-\frac{1}{d-1}}\right)\cdot\frac{1}{\sqrt{d-1}}\geqslant\frac{V(d-1)}{2\sqrt{d-1}}\tag{14}$$

结合两者结果即可得到结论。

定理 7 (**两两向量几近正交**). 在 d 维单位球中随机取 n 个点 $\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n$，则有 $1-O(\frac{1}{n})$ 的概率

- $|\mathbf{x}_i|\leqslant 1-\frac{2\ln n}{d}$ 对每个 i 均成立；
- $|\mathbf{x}_i\cdot\mathbf{x}_j|\leqslant\frac{\sqrt{6\ln n}}{\sqrt{d-1}}$ 对每对 $i\neq j$ 均成立。

证明 5. 根据定理 4，$\Pr[|\mathbf{x}_i|<1-\varepsilon]=(1-\varepsilon)^d\leqslant\mathrm{e}^{-\varepsilon d}$，故

$$\Pr\left[|\mathbf{x}_i|<1-\frac{2\ln n}{d}\right]\leqslant\mathrm{e}^{-(\frac{2\ln n}{d})d}=1/n^2\tag{15}$$

union bound 一下，存在一个 $\mathbf{x}_i$ 寄掉的概率不超过 $1/n$。

根据定理 6，$\Pr\left[|\mathbf{x}_i\cdot\mathbf{x}_j|>\frac{c}{\sqrt{d-1}}\right]\leqslant\frac{2}{c}\mathrm{e}^{-\frac{c^2}{2}}$，故

$$\Pr\left[|\mathbf{x}_i\cdot\mathbf{x}_j|>\frac{\sqrt{6\ln n}}{\sqrt{d-1}}\right]\leqslant\mathrm{e}^{-\frac{6\ln n}{2}}=1/n^3\tag{16}$$

union bound 一下，存在一对 $\mathbf{x}_i,\mathbf{x}_j$ 寄掉的概率不超过 $\binom{n}{2}/n^3=O(1/n)$。

定理 8 (**单位球中随机采点的方法**). **(i)**按 $p(\mathbf{x})=\frac{1}{(2\pi)^{\frac{d}{2}}}\mathrm{e}^{-\frac{\sum_{i=1}^d x_i^2}{2}}$ 的概率取向量 $\mathbf{x}$；**(ii)** 按 $\rho(r)=dr^{d-1}$ 的概率取模长 r，此时得到 $\mathbf{y}=r\frac{\mathbf{x}}{|\mathbf{x}|}$ 就是均匀随机的高维单位球中的点。

定理 9 (**Gaussian Annulus Theorem**). 对于 d 维的单位方差的高斯分布，对于 $\beta\leqslant\sqrt{d}$，有至多 $3\mathrm{e}^{-c\beta^2}$ 的概率密度分布在 $\sqrt{d}-\beta\leqslant|\mathbf{x}|\leqslant\sqrt{d}+\beta$ 之外，其中 c 是一个固定的正常数。

定理 10 (**Random Projection Theorem**). 随机取 k 个服从高斯分布的向量 $\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_k$，构造函数 $f:\mathbb{R}^d\to\mathbb{R}^k$ 满足

$$f(\mathbf{v})=(\mathbf{u}_1\cdot\mathbf{v},\mathbf{u}_2\cdot\mathbf{v},\cdots,\mathbf{u}_k\cdot\mathbf{v})\tag{17}$$

则存在常数 $c>0$ 使得对于任意 $\varepsilon\in(0,1)$ 均满足

$$\Pr\left[\left||f(\mathbf{v})|-\sqrt{k}|\mathbf{v}|\right|\geqslant\varepsilon\sqrt{k}|\mathbf{v}|\right]\leqslant 3\mathrm{e}^{-ck^2\varepsilon}\tag{18}$$

证明 6. 不妨设 $|\mathbf{v}|=1$，注意到

$$\operatorname{Var}[\mathbf{u}_i\cdot\mathbf{v}]=\operatorname{Var}\left[\sum_{j=1}^{d}u_{ij}v_j\right]=\sum_{j=1}^{d}v_j^2\operatorname{Var}[u_{ij}]=\sum_{j=1}^{d}v_j^2=1 \tag{19}$$

因此 $f(\mathbf{v})$ 也是 $\mathbb{R}^k$ 中服从高斯分布的随机向量，套用 Gaussian Annulus Theorem 即可。

定理 11 (Johnson-Lindenstrauss Lemma). 对于任意 $\varepsilon\in(0,1)$ 以及正整数 n，取 $k\geqslant\frac{3}{c\varepsilon^2}\ln n$，对于任意 $\mathbb{R}^d$ 中大小为 n 的点集 $\{\mathbf{v}_i\}$ 和随机映射 f(定义同上)，有至少 $1-\frac{3}{2n}$ 的概率，对于任意 i,j 均满足

$$(1-\varepsilon)\sqrt{k}|\mathbf{v}_i-\mathbf{v}_j|\leqslant|f(\mathbf{v}_i)-f(\mathbf{v}_j)|\leqslant(1+\varepsilon)\sqrt{k}|\mathbf{v}_i-\mathbf{v}_j| \tag{20}$$

证明 7. 根据 Random Projection Theorem，$|f(\mathbf{v}_i)-f(\mathbf{v}_j)|$ 不在 $\left[(1-\varepsilon)\sqrt{k}|\mathbf{v}_i-\mathbf{v}_j|,(1+\varepsilon)\sqrt{k}|\mathbf{v}_i-\mathbf{v}_j|\right]$ 范围中的概率不超过 $3\mathrm{e}^{-ck^2\varepsilon}\leqslant\frac{3}{n^3}$。由于 $\binom{n}{2}\leqslant\frac{n^2}{2}$，根据 union bound 可得结论。

2 Best-Fit Subspaces and Singular Value Decomposition(SVD)

定义 1 (**奇异向量与奇异值**).

$$\mathbf{v}_i = \arg\max_{|\mathbf{v}|=1, \mathbf{v}\perp\mathbf{v}_1,\cdots,\mathbf{v}_{i-1}} |A\mathbf{v}| \qquad \sigma_i = |A\mathbf{v}_i| \qquad \mathbf{u}_i = \frac{A\mathbf{v}_i}{\sigma_i} \tag{21}$$

此时可以得到 A 的奇异值分解 (Singular Value Decomposition, SVD)

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}} \tag{22}$$

定义 2 (**Frobenius norm**). 定义矩阵的 Frobenius norm 为

$$\|A\|_F = \sqrt{\sum_{j,k} a_{jk}^2} \tag{23}$$

也即根号下所有行向量长度的平方和。$\|A\|_F = \sqrt{\sum_{i=1}^{r}\sigma_i^2}$。

定义 3 (**Spectral norm**). 定义矩阵的 Spectral norm 为

$$\|A\|_2 = \max_{|\mathbf{x}|\leqslant 1} |A\mathbf{x}| \tag{24}$$

$\|A\|_2 = \sigma_1$。

定理 12 (**Greedy Algorithm Works**). 设 A 的奇异向量为 $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r$。对于任意 $1 \leqslant k \leqslant r$, 由 $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k$ 张成的空间 V_k 都是 A 的最佳 k 维近似。

证明 8. 考虑归纳。$k = 1$ 时显然成立，故尝试从 $k-1$ 维最佳推出 k 维最佳。假设 W_k 是 A 的最佳 k 维近似，由于维数是 k，故必然存在一个单位向量与 $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{k-1}$ 均垂直。不妨记 $W_k = \langle \mathbf{w}_1, \cdots, \mathbf{w}_{k-1}, \mathbf{w}_k\rangle$，其中 $\mathbf{w}_k$ 垂直于 $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{k-1}$，那么根据归纳假设，$\sum_{i=1}^{k-1} |A\mathbf{w}_i|^2 \leqslant \sum_{i=1}^{k-1} |A\mathbf{v}_i|^2$，根据 $\mathbf{v}_k$ 的定义 (选取规则)，$|A\mathbf{w}_k|^2 \leqslant |A\mathbf{v}_k|^2$，从而 $\sum_{i=1}^{k} |A\mathbf{w}_i|^2 \leqslant \sum_{i=1}^{k} |A\mathbf{v}_i|^2$，即说明 V_k 也是 A 的最佳 k 维近似。

定理 13 (**A_k 是最佳 Frobenius norm 近似**). 对于任意秩不超过 k 的矩阵 B

$$\|A - A_k\|_F \leqslant \|A - B\|_F \tag{25}$$

证明 9. $\|A - B\|_F^2$ 不小于 A 的所有行向量到 B 的行空间的距离平方和，而 A_k 恰好是后者问题中最小化这个值的 B。

定理 14 (**A_k 是最佳 Spectral norm 近似**). 对于任意秩不超过 k 的矩阵 B

$$\|A - A_k\|_2 \leqslant \|A - B\|_2 \tag{26}$$

证明 10. 考虑 $\ker B$ 与 $\langle \mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{k+1}\rangle$，由于二者维度分别为 $\geqslant d-k$ 与 $k+1$，故必然存在一单位向量 $\mathbf{z}$ 属于二者的交

$$\begin{aligned}\|A - B\|_2 \geqslant |(A-B)\mathbf{z}| = |A\mathbf{z}| &= \left|\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}} \mathbf{z}\right| = \left|\sum_{i=1}^{k+1} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}} \mathbf{z}\right| \\ &= \sqrt{\sum_{i=1}^{k+1} \sigma_i^2 (\mathbf{v}_i^{\mathrm{T}}\mathbf{z})^2} \geqslant \sigma_{k+1}\sqrt{\sum_{i=1}^{k+1} (\mathbf{v}_i^{\mathrm{T}}\mathbf{z})^2} = \sigma_{k+1}\end{aligned} \tag{27}$$

而 $\|A - A_k\|_2 = \sigma_{k+1}$，故 A_k 是最佳 Spectral norm 近似。

定理 15 (**左奇异向量两两垂直**). 左奇异向量两两垂直。

证明 11. 设 i 是最小的下标满足存在 j，使 $\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j=\delta>0$，令 $\mathbf{v}_i'=\frac{\mathbf{v}_i+\varepsilon\mathbf{v}_j}{|\mathbf{v}_i+\varepsilon\mathbf{v}_j|}=\frac{1}{\sqrt{1+\varepsilon^2}}(\mathbf{v}_i+\varepsilon\mathbf{v}_j)$，则 $A\mathbf{v}_i'=\frac{1}{\sqrt{1+\varepsilon^2}}(\sigma_i\mathbf{v}_i+\varepsilon\sigma_j\mathbf{v}_j), |A\mathbf{v}_i'|\geqslant\mathbf{u}_i^{\mathrm{T}}A\mathbf{v}_i'=\frac{1}{\sqrt{1+\varepsilon^2}}(\sigma_i+\varepsilon\sigma_j\delta)\geqslant(\sigma_i+\varepsilon\sigma_j\delta)(1-\frac{\varepsilon^2}{2})=\sigma_i+\varepsilon\sigma_j\delta-\frac{\varepsilon^2}{2}\sigma_i-\frac{\varepsilon^3}{2}\sigma_j\delta$。

当 $\varepsilon\to0$ 时，可以发现上式 $>\sigma_i$，这与 $\mathbf{v}_i$ 的选取矛盾。故左奇异向量两两垂直。

定理 16 (Power Method). A 是 $n\times d$ 的矩阵，$\mathbf{x}\in\mathbb{R}^d$ 满足 $|\mathbf{x}^{\mathrm{T}}\mathbf{v}_1|\geqslant\delta>0$，令 V 表示所有对应于 $\geqslant(1-\varepsilon)\sigma_1$ 奇异值的奇异向量张成的子空间，$\mathbf{w}$ 为通过 Power Method 迭代 $k=\frac{\ln(1/\varepsilon\delta)}{2\varepsilon}$ 轮后得到的单位向量，即

$$\mathbf{w}=\frac{(A^{\mathrm{T}}A)^k\mathbf{x}}{|(A^{\mathrm{T}}A)^k\mathbf{x}|} \tag{28}$$

则 $\mathbf{w}$ 只有不超过 ε 的分量垂直于 V。

证明 12. 设 A 的 SVD 为 $A=\sum\limits_{i=1}^{r}\sigma_i\mathbf{u}_i\mathbf{v}_i^{\mathrm{T}}$，则 $(A^{\mathrm{T}}A)^k=\sum\limits_{i=1}^{r}\sigma_i^{2k}\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}$。再设 $\mathbf{x}=\sum\limits_{i=1}^{d}c_i\mathbf{v}_i$，于是 $(A^{\mathrm{T}}A)^k\mathbf{x}=\sum\limits_{i=1}^{r}\sigma_i^{2k}c_i\mathbf{v}_i$。

考虑分别计算垂直部分长度的 upper bound 和向量总长度的 lower bound。

$$\sum_{i=m+1}^{d}(\sigma_i^{2k}c_i)^2\leqslant(1-\varepsilon)^{4k}\sigma_1^{4k}\sum_{i=m+1}^{d}c_i^2\leqslant(1-\varepsilon)^{4k}\sigma_1^{4k} \tag{29}$$

$$|(A^{\mathrm{T}}A)^k\mathbf{x}|^2=\left|\sum_{i=1}^{d}\sigma_i^{2k}c_i\mathbf{v}_i\right|^2=\sum_{i=1}^{d}(\sigma_i^{2k}c_i)^2\geqslant\sigma_1^{4k}c_1^2\geqslant\sigma_1^{4k}\delta^2 \tag{30}$$

其中 $\sigma_m\geqslant(1-\varepsilon)\sigma_1,\sigma_{m+1}<(1-\varepsilon)\sigma_1$。相除即可得到结论

$$\frac{(1-\varepsilon)^{2k}\sigma_1^{2k}}{\delta\sigma_1^{2k}}=\frac{(1-\varepsilon)^{2k}}{\delta}\leqslant\frac{\mathrm{e}^{-2k\varepsilon}}{\delta}=\varepsilon \tag{31}$$

3 Machine Learning

3.1 Perception Algorithm

定义 4 (**Perception Algorithm**). **(i)** $\mathbf{w} \leftarrow 0$，**(ii)** 每当存在 $\mathbf{x}_i$ 使得 $\mathbf{x}_i l_i \cdot \mathbf{w} \leqslant 0$，就更新 $\mathbf{w} \leftarrow \mathbf{w} + \mathbf{x}_i l_i$。

定理 17 (**Perception Algorithm 的运行时间上界**). 如果存在一个 $\mathbf{w}^*$ 满足 $(\mathbf{w}^* \cdot \mathbf{x}_i) l_i \geqslant 1$ 对于任意 i 成立，则 Perception Algorithm 可以在不超过 $r^2|\mathbf{w}^*|^2$ 步内找到一个 $(\mathbf{w} \cdot \mathbf{x}_i) l_i > 0$ 的解 $\mathbf{w}$，其中 $r = \max_i |\mathbf{x}_i|$。

证明 13. 考虑两个量：$\mathbf{w}^{\mathrm{T}}\mathbf{w}^*$ 和 $|\mathbf{w}|^2$。前者每步至少增加 1，因为

$$(\mathbf{w} + \mathbf{x}_i l_i)^{\mathrm{T}} \mathbf{w}^* = \mathbf{w}^{\mathrm{T}}\mathbf{w}^* + \mathbf{x}_i^{\mathrm{T}} l_i \mathbf{w}^* \geqslant \mathbf{w}^{\mathrm{T}}\mathbf{w}^* + 1 \tag{32}$$

后者每步至多增加 r^2，因为

$$(\mathbf{w} + \mathbf{x}_i l_i)^{\mathrm{T}}(\mathbf{w} + \mathbf{x}_i l_i) = |\mathbf{w}|^2 + 2\mathbf{x}_i^{\mathrm{T}} l_i \mathbf{w} + |\mathbf{x}_i l_i|^2 \leqslant |\mathbf{w}|^2 + |\mathbf{x}_i|^2 \leqslant |\mathbf{w}^2| + r^2 \tag{33}$$

设运行步数为 m，则由 $|\mathbf{w}||\mathbf{w}^*| \geqslant \mathbf{w}^{\mathrm{T}}\mathbf{w}^* \geqslant m, |\mathbf{w}|^2 \leqslant r^2 m$ 可以解得 $m \leqslant r^2|\mathbf{w}^*|^2$。

3.2 Kernel Function

定义 5 (**Kernel Function**). 形如 $k(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^{\mathrm{T}}\varphi(\mathbf{x}_j)$ 的函数 k 被称为 kernel function。

引理 1 (**Kernel Matrix**). 一个矩阵 K 是 kernel matrix，即存在一个函数 φ 使 $k_{ij} = \varphi(\mathbf{x}_i)^{\mathrm{T}}\varphi(\mathbf{x}_j)$，当且仅当 K 半正定。

定理 18. 设 k_1, k_2 是两个 kernel functions，则

1. 对任意 $c > 0$，ck_1 是一个 kernel function。
2. 对任意标量函数 f，$k_3(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})f(\mathbf{y})k_1(\mathbf{x}, \mathbf{y})$ 是一个 kernel function。
3. $k_1 + k_2$ 是一个 kernel function。
4. $k_1 k_2$ 是一个 kernel function。

3.3 Generalizing to New Data

定理 19. 如果训练集大小满足

$$n \geqslant \frac{1}{\varepsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta}) \tag{34}$$

则就有至少 $1 - \delta$ 的概率，每个 $h \in \mathcal{H}$ 都满足 $err_S(h) = 0 \Rightarrow err_D(h) < \varepsilon$。$S$ 表示训练集，D 表示整体分布。

证明 14. 某一个 h 寄掉的概率 $\leqslant (1-\varepsilon)^n$(相当于大小为 n 的训练集一次都没有砸中 h 的错误)，故根据 union bound，存在一个 h 寄掉的概率 $\leqslant |\mathcal{H}|(1-\varepsilon)^n \leqslant |\mathcal{H}|\mathrm{e}^{-n\varepsilon}$，$|\mathcal{H}|\mathrm{e}^{-n\varepsilon} \leqslant \delta \Rightarrow n \geqslant \frac{1}{\varepsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})$。

定理 20 (**Hoeffding bounds**). $x_1, x_2, \cdots, x_n$ 是 n 个独立随机 $\{0,1\}$ 变量满足 $\Pr[x_i = 1] = p$。令 $s = \sum_i x_i$，则对于任意 $0 \leqslant \alpha \leqslant 1$

$$\Pr\left[\frac{s}{n} > p + \alpha\right] \leqslant \mathrm{e}^{-2n\alpha^2} \qquad \Pr\left[\frac{s}{n} < p - \alpha\right] \leqslant \mathrm{e}^{-2n\alpha^2} \tag{35}$$

定理 21 (**一致收敛**). 如果训练集大小满足

$$n \geqslant \frac{1}{2\varepsilon^2}(\ln|\mathcal{H}| + \ln\frac{2}{\delta}) \tag{36}$$

则就有至少 $1 - \delta$ 的概率，每个 $h \in H$ 都满足 $|err_D(h) - err_S(h)| < \varepsilon$。

证明 15. 根据定理 20，把 $err_S(h)$ 理解成 $\frac{s}{n}$，把 $err_D(h)$ 理解成 p，可以得到某个 h 寄掉的概率 $\leqslant 2\mathrm{e}^{-2n\varepsilon^2}$，union bound 后 $2|\mathcal{H}|\mathrm{e}^{-2n\varepsilon^2} \leqslant \delta \Rightarrow n \geqslant \frac{1}{2\varepsilon^2}(\ln|\mathcal{H}| + \ln\frac{2}{\delta})$。

定理 22 (奥卡姆剃刀，Occam's razor). 本质就是定理 19 在 $\mathcal{H} = [2^b - 1]$ 时的平凡推论，即规则集合包含所有可用少于 b 个比特描述的 h。但这揭示了一件很有意思的事情：对于同一个训练样本集 S，使用越简单的规则去描述，它的置信度就越高。当然这其实也不是绝对的，因为不同主体可以有不同的定义"简单"的方式。

3.4 VC-Dimension

定义 6 (VC-Dimension). 对于一个集合系统 $(X, \mathcal{H})$，称 $\mathcal{H}$ shatter 了一个集合 $A \subseteq X$，如果 A 的每个子集都可以表示成 $A \cap h$，其中 $h \in \mathcal{H}$。$\mathcal{H}$ 的 VC-Dimension 就是最大的可被 $\mathcal{H}$ shatter 的集合大小。

命题 1 (一些 VC-Dimension 的例子).

- 边平行于坐标轴的矩形：VC-Dimension 为 4。
- 实数区间：VC-Dimension 为 2。
- 两个实数区间：VC-Dimension 为 4。
- k 个半平面：VC-Dimension 为 $2k+1$。
- 有限集：VC-Dimension 为 ∞。
- 凸多边形：VC-Dimension 为 ∞。
- d 维半平面：VC-Dimension 为 $d+1$。
- d 维球：VC-Dimension 为 $d+1$。

定义 7 (Shatter Function). 对于集合系统 $(X, \mathcal{H})$，定义其 shatter function $\pi_{\mathcal{H}}(n)$ 表示可被 $A \cap h$ 表出的 A 的子集数量的最大值，其中 A 取遍所有 n 元集合。

设 $\mathcal{H}$ 的 VC-Dimension 为 d，对于 $n \leqslant d$，有 $\pi_{\mathcal{H}}(n) = 2^n$，在这之外，$\pi_{\mathcal{H}}(n)$ 随着 n 多项式级别增长。

定理 23 (Sauer). 设 $\mathcal{H}$ 的 VC-Dimension 为 d，则 $\pi_{\mathcal{H}}(n) \leqslant \binom{n}{\leqslant d} \leqslant n^d + 1$ 对所有 n 成立。

证明 16. 考虑归纳，尝试证明

$$\pi_{\mathcal{H}}(n) = \pi_{\mathcal{H}_1}(n-1) + \pi_{\mathcal{H}_2}(n-1) \leqslant \binom{n-1}{\leqslant d} + \binom{n-1}{\leqslant d-1} = \binom{n}{\leqslant d} \tag{37}$$

定理 24 (Key Theorem). $(X, \mathcal{H})$ 是一集合系统，D 是 X 上的概率分布，S_1 包含 n 个根据 D 分布从 X 中选取的点，其中 n 满足

$$n \geqslant \max\left\{\frac{8}{\varepsilon}, \frac{2}{\varepsilon}\left[\log_2 2\pi_{\mathcal{H}}(2n) + \log_2 \frac{1}{\delta}\right]\right\} \tag{38}$$

则有至少 $1-\delta$ 的概率，$\mathcal{H}$ 中每个概率密度 $\geqslant \varepsilon$(类似于 $|h| \geqslant \varepsilon|X|$，但 X 可能是无限的) 的 h 都会满足 $h \cap S_1 \neq \varnothing$。

证明 17. 考虑事件 A：存在一个概率密度 $\geqslant \varepsilon$ 的 $h \in \mathcal{H}$ 使得 $h \cap S_1 = \varnothing$。按照与 S_1 相同的方法采样 S_2，再考虑事件 B：存在一个概率密度 $\geqslant \varepsilon$ 的 $h \in \mathcal{H}$ 使得 $h \cap S_1 = \varnothing$ 且 $|h \cap S_2| \geqslant \frac{\varepsilon}{2}n$。

先证明 $\Pr[B|A] \geqslant \frac{1}{2}$，这可以说明 $\Pr[B] \geqslant \Pr[B|A]\Pr[A] \geqslant \frac{1}{2}\Pr[A]$。考虑 $\mathbf{x} = (x_1, x_2, \cdots, x_n)$，其中 x_i 是 $\{0,1\}$ 变量表示 S_2 的第 i 次采样是否在 h 中，$\mathbb{E}[x_i] = \varepsilon$(不妨假设就是 ε)，$\mathrm{Var}[x_i] \leqslant \varepsilon$，则 $\mathbb{E}[|\mathbf{x}|^2] = n\varepsilon, \mathrm{Var}[|\mathbf{x}|^2] \leqslant n\varepsilon$，于是 $\Pr[|\mathbf{x}|^2 \geqslant \frac{\varepsilon}{2}n] \geqslant \Pr[||\mathbf{x}|^2 - \mathbb{E}[|\mathbf{x}|^2]| \leqslant \frac{\varepsilon}{2}n] \geqslant 1 - \mathrm{Var}[|\mathbf{x}|^2]\left(\frac{2}{n\varepsilon}\right)^2 \geqslant 1 - \frac{4}{n\varepsilon}$。当 $n \geqslant \frac{8}{\varepsilon}$ 时，可以得到 $\Pr[|\mathbf{x}|^2 \geqslant \frac{\varepsilon}{2}n] \geqslant \frac{1}{2}$。

于是接下来只考虑限制 $\Pr[B]$。更换 S_1, S_2 的采样顺序，考虑先进行 $2n$ 大小的采样得到 S_3，再随机选择 n 个点给 S_1。注意此时 $S_3 \cap \mathcal{H} = \{S_3 \cap h : h \in \mathcal{H}\}$ 是一个有限集合，其大小不超过 $\pi_{\mathcal{H}}(2n)$，因此我们就可以

对这个有限集合运用 union bound 了。对于每个 $h' \in S_3 \cap \mathcal{H}$，满足 $|S_1 \cap h'| = 0$ 且 $|S_2 \cap h'| \geqslant \frac{\varepsilon}{2} n$ 的概率不超过 ($|h'| \geqslant \frac{\varepsilon}{2} n$，相当于有至少 $\frac{\varepsilon}{2} n$ 个元素需要保证被划入 S_2)

$$2^{-n\varepsilon/2} \leqslant 2^{-\log_2 2\pi_{\mathcal{H}}(2n)+\log\delta} = \frac{\delta}{2\pi_{\mathcal{H}}(2n)} \tag{39}$$

从而根据 union bound，存在一个这样的 h' 的概率不超过 $\frac{\delta}{2\pi_{\mathcal{H}}(2n)} \cdot \pi_{\mathcal{H}}(2n) = \frac{\delta}{2}$，即 $\Pr[B] \leqslant \frac{\delta}{2}$。于是 $\Pr[A] \leqslant \delta$。

注 1. Key Theorem 是利用 Shatter Function 对可能为无限集的 X 给出了一个类似定理 19 的结论，其中用到了被称为 **double sampling** 的技巧。如果把 h 理解成 error，那么这个定理等价于在说：以至少 $1-\delta$ 的概率，每个 $h \in \mathcal{H}$ 都满足 $err_D(h) \geqslant \varepsilon \Rightarrow err_S(h) \geqslant 0$ 或者等价的，$err_S(h) = 0 \Rightarrow err_D(h) < \varepsilon$。

3.5 Online Learning

3.5.1 三个在线学习的例子

考虑有 n 位专家的二分类问题。

命题 2 (Q1). 若存在 perfect expert(永远回答正确)，则存在策略使出错次数不超过 $\log_2 n$。

证明 18. 使用 majority elimination 即可 (每次选取专家回答的 majority，并把出错的专家干掉，这样自己的每次出错都会使剩余专家数量减少至少一半)。

命题 3 (Q2). 采用这样的策略：所有专家初始权重均为 1，每次选取专家回答的加权 majority，并把出错的专家权重除以 2。假设最厉害的专家一共出错了 opt 次。

- 结束时总权重至少还有 $\left(\frac{1}{2}\right)^{\mathrm{opt}}$。
- 自己的每次出错会导致总权重减少至少 $\frac{1}{4}$。

从而该策略可以使得

$$\left(\frac{1}{2}\right)^{\mathrm{opt}} \leqslant n\left(\frac{3}{4}\right)^{\#\mathrm{mistake}} \Rightarrow \#\mathrm{mistake} \leqslant \frac{\mathrm{opt} + \log_2 n}{\log_2 \frac{4}{3}} \tag{40}$$

命题 4 (Q3). 采用这样的策略：所有专家初始权重均为 1，每次按权重随机一位专家的回答，并把出错的专家权重乘 $(1-\varepsilon)$。假设最厉害的专家一共出错了 opt 次。

注意每轮结束后总权重的变换总是 $w \to w(1-\varepsilon\Pr[\mathrm{mistake}])$，这是与本轮是否回答错误无关的，而 $\mathbb{E}[\#\mathrm{mistake}] = \sum \Pr[\mathrm{mistake}]$，所以考虑对后者求上界：

$$(1-\varepsilon)^{\mathrm{opt}} \leqslant n\prod(1-\varepsilon\Pr[\mathrm{mistake}]) \leqslant n\prod \mathrm{e}^{-\varepsilon\Pr[\mathrm{mistake}]} = n\mathrm{e}^{-\varepsilon\sum\Pr[\mathrm{mistake}]} \tag{41}$$

$$\mathbb{E}[\#\mathrm{mistake}] = \sum \Pr[\mathrm{mistake}] \leqslant \frac{\ln n + \mathrm{opt}\ln\frac{1}{1-\varepsilon}}{\varepsilon} \tag{42}$$

3.5.2 Boosting

定义 8 (γ-weak learner). 一个 γ-weak learner 指的是一种算法，在任何给定的带权样本集下，都能够给出一个分类器，其可以正确标识集合中权重和至少为 $(\frac{1}{2}+\gamma)\sum w_i$ 的样本。

定义 9 (Boosting). Boosting Algorithm 指的是利用一个 γ-weak learner 来得到一个 strong learner 的算法。分为如下几步：

1. 对于样本集 $S = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$，设定初始权重 $\mathbf{w} = (w_1, \cdots, w_n)$ 其中 $w_i = 1, \forall i$。
2. 重复 t_0 轮，第 t 轮将带权样本集 $(S, \mathbf{w})$ 喂给 γ-weak learner 并得到分类器 h_t，把 h_t 分错的那些元素的权重乘上 $\alpha = \frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma}$。
3. t_0 轮结束后，输出 $\mathrm{MAJ}(h_1, \cdots, h_{t_0})$ 作为最终的分类器。

定理 25. 取 $t_0 > \frac{\ln n}{\gamma^2}$，便可以得到一个 training error 为零的分类器。

证明 19. 假设最终输出的分类器错了 m 个样本。一方面，最后所有样本的权重和至少是 $m\alpha^{t_0/2}$，因为这 m 个样本需要被分错至少 $t_0/2$ 轮；另一方面，γ-weak learner 的性质保证其每次只会对不超过总权重 $(\frac{1}{2}-\gamma)$ 的部分出错，于是若记 $\mathrm{weight}(t)$ 表示第 t 轮后所有样本的总权重，则 $\mathrm{weight}(t+1) \leqslant (\alpha(\frac{1}{2}-\gamma)+(\frac{1}{2}+\gamma))\mathrm{weight}(t) = (1+2\gamma)\mathrm{weight}(t)$，从而得到 $\mathrm{weight}(t_0) \leqslant n(1+2\gamma)^{t_0}$。于是

$$m\alpha^{t_0/2} \leqslant n(1+2\gamma)^{t_0} \quad \Rightarrow \quad m \leqslant n(1-2\gamma)^{t_0/2}(1+2\gamma)^{t_0/2} = n(1-4\gamma^2)^{t_0/2} \leqslant n\mathrm{e}^{-2t_0\gamma^2} \tag{43}$$

若取 $t_0 > \frac{\ln n}{\gamma^2}$，便可得 $m < 1$，从而 training error 为零。

4 Algorithms for Massive Data Problems

4.1 Streaming

4.1.1 Picking Elements

从数据流中等概率的取一个数 a_i：维护当前已看过的数的数量 n 和选取的数 x，当出现一个新数 b 时，以 $\frac{1}{n+1}$ 的概率替换 a，同时 $n \leftarrow n+1$。

同理还可以带权随机，只需要维护权重前缀和即可。

4.1.2 Distinct Elements

命题 5 (确定性算法的空间下界). 确定性算法对于长度为 $m+1$ 的序列，需要至少 m 比特的存储空间。

证明 20. 考虑把 $\{1,2,\cdots,m\}$ 的任意非空子集喂给该确定性算法，子集有 2^m-1 种，而状态表示却只有 2^{m-1} 种，故一定产生冲突，寄。

定理 26 (用 min 估计). 假设序列中不同的元素为 $b_1,b_2,\cdots,b_d$，取一个 2-universal 的哈希函数 h，其值域为 $[0,M-1]$，取 $S=\{h(b_1),h(b_2),\cdots,h(b_d)\}$ 的最小值 min。则有至少 $\frac{2}{3}-\frac{d}{M}$ 的概率，$\frac{d}{6} \leqslant \frac{M}{\min} \leqslant 6d$。

证明 21. 先证明 $\Pr\left[\min \leqslant \frac{M}{6d}\right] \leqslant \frac{1}{6}+\frac{d}{M}$，这一部分并不依赖 pairwise independence。min 小于一个数说明存在一者小于，这个概率可以 union bound 放缩成每一者小于的概率之和

$$\Pr\left[\min \leqslant \frac{M}{6d}\right] = \Pr\left[\exists k, h(b_k) \leqslant \frac{M}{6d}\right] \leqslant \sum_{i=1}^{d} \Pr\left[h(b_i) \leqslant \frac{M}{6d}\right] \leqslant d\left(\frac{\left\lceil\frac{M}{6d}\right\rceil}{M}\right) \leqslant \frac{1}{6}+\frac{d}{M} \tag{44}$$

再证明 $\Pr\left[\min \geqslant \frac{6M}{d}\right] \leqslant \frac{1}{6}$，这就需要用到 pairwise independence。min 大于一个数说明每一者都比这个数大，用一个 $\{0,1\}$ 变量 y_i 表示 $h(b_i)$ 是否大于等于 $\frac{6M}{d}$，再记 $y=\sum_{i=1}^{d} y_i$，易得 $\mathbb{E}[y]=d\mathbb{E}[y_i]=6, \mathrm{Var}[y]=d\mathrm{Var}[y_i]=d(\mathbb{E}[y_i^2]-\mathbb{E}[y_i]^2) \leqslant d\mathbb{E}[y_i^2]=d\mathbb{E}[y_i]=\mathbb{E}[y]$，从而根据 Chebyshev's Inequality

$$\Pr\left[\min \geqslant \frac{6M}{d}\right] = \Pr\left[\forall k, h(b_k) \geqslant \frac{6M}{d}\right] = \Pr[y=0] \leqslant \Pr[|y-\mathbb{E}[y]| \geqslant \mathbb{E}[y]] \leqslant \frac{\mathrm{Var}[y]}{\mathbb{E}[y]^2} \leqslant \frac{1}{\mathbb{E}[y]} = \frac{1}{6} \tag{45}$$

4.1.3 Occurences of a Given Element

朴素做法需要 $\log n$ 的空间，因为只需要维护一个计数器。

存在一种 $\log\log n$ 空间的方法：初始记录 0，每次遇到需要数的元素，设当前记录的数是 k，就以 $1/2^k$ 的概率将这个数 $+1$，这样当记录的数字是 k 时，元素的期望出现次数就是 $1+2+\cdots+2^{k-1}=2^k-1$。

4.1.4 Majority Algorithm

命题 6 (确定性算法的空间下界). 假设元素一共有 m 种，序列长度为 n，那么确定性算法至少需要 $\Omega(\min\{n,m\})$ 的空间。

证明 22. 考虑序列前 $n/2$ 个元素组成的集合 $S \subseteq \{1,2,\cdots,m\}$，对于不同的集合 S，算法必须有不同的记录，这是因为一旦对于不同的集合 S,S' 有了相同的记录，就可以在后 $n/2$ 个数中全部写 $S \setminus S'$ 中的一个元素，这样就寄了（正确的结果是前者存在 majority 而后者不存在）。因此至少需要 $\log_2\left(\sum_{i=1}^{n/2}\binom{m}{i}\right)=\Omega(\min(n,m))$ 的空间。

定义 10 (Majority Algorithm). 维护一个数 a 和一个计数器 cnt，遇到一个新数时，如果与 a 相同，就把计数器 $+1$；否则如果计数器大于 1 就 -1，否则用新的数替换 a 并把计数器置为 1。

这个算法可能给出 false positive(没有 majority 的时候给出一个错误的 majority)，但不可能有 false negative(当 majority 存在时一定会正确地给出)。

4.1.5 Frequent Algorithm

定义 11 (Frequent Algorithm). Frequent Algorithm 是 Majority Algorithm 的升级版，维护 k 个元素以及分别的计数器，每当遇到一个新数，如果与 k 个元素中的某个相同，就将其对应的计数器 $+1$，否则若 k 个元素中存在空位，就加入该空位并把计数器置为 1，若不存在空位，则把所有数的计数器值 -1，并把计数器为 0 的元素删去。

定理 27. Frequent Algorithm 只会把一种元素少数 (under count) 至多 $\frac{n}{k+1}$ 次。当一个数没有出现在 k 个元素中时，认为这个数被数了 0 次。

证明 23. 每当触发一次“所有计数器 -1”，都会导致所有数被数的总数减少恰好 $k+1$，因此只会触发不超过 $\frac{n}{k+1}$ 次，而一种元素在一次触发中至多只会被少数 1 次，故每个数至多只会被少数 $\frac{n}{k+1}$ 次。

4.1.6 The Second Moment

设一个数据流中 s 元素出现了 f_s 次，定义该数据流的 second moment 为 $\sum_{s=1}^{m} f_s^2$，即每个元素出现次数的平方。我们希望估计这个值。取一个哈希函数 $h:\{1,2,\cdots,m\}\to\{-1,1\}$，并记 $h(s)=x_s$，此时有 $\mathbb{E}\left[\sum_{s=1}^{m} x_s f_s\right]=0$，考虑计算

$$\mathbb{E}\left[a \triangleq \left(\sum_{s=1}^{m} x_s f_s\right)^2\right] = \mathbb{E}\left[\sum_{s=1}^{m} x_s^2 f_s^2\right] + 2\mathbb{E}\left[\sum_{s<t} x_s x_t f_s f_t\right] = \sum_{s=1}^{m} f_s^2 \tag{46}$$

第二个等号要想成立，要求 $\mathbb{E}[x_s x_t]=\mathbb{E}[x_s]\,\mathbb{E}[x_t]=0$，级要求哈希函数 h 满足 pairwise independence。

我们希望进一步地限制一下方差，由于期望已经确定了，所以只需要考虑限制 $\mathbb{E}\left[a^2\right]$。如果 h 满足 4-way independence，那么

$$\begin{aligned}\mathbb{E}\left[a^2\right] &= \mathbb{E}\left[\sum_{s,t,u,v} x_s x_t x_u x_v f_s f_t f_u f_v\right] \leqslant \binom{4}{2}\mathbb{E}\left[\sum_{s<t} x_s^2 x_t^2 f_s f_t\right] + \mathbb{E}\left[\sum_{s} x_s^4 f_s^4\right] \\ &= 6\sum_{s<t} f_s^2 f_t^2 + \sum_{s=1}^{m} f_s^4 \leqslant 3\left(\sum_{s=1}^{m} f_s^2\right)^2 = 3\mathbb{E}\left[a\right]^2\end{aligned} \tag{47}$$

4.2 Hash Functions

我们说的 pairwise independence(或者 2-universal)，指的是对于一个哈希函数族 H，它满足任意 $x,y\in\{1,2,\cdots,m\}$(定义域) 满足 $x\neq y$ 和任意 $w,z\in\{0,1,\cdots,M-1\}$(值域)，都有

$$\Pr\left[h(x)=w \wedge h(y)=z\right] = \frac{1}{M^2} \tag{48}$$

随机性来自于 h 从 H 中的选取。

如果进一步要求 k-way independence，那么要求就改为对于任意 $\binom{m}{k}$ 种定义选取和 M^k 种值域选取，概率都是一样的 $1/M^k$。

4.3 Sampling & Sketching

4.3.1 Sketching Matrix Multiplication

假设需要 sketch 的是 $m\times n$ 的矩阵 A 和 $n\times p$ 的矩阵 B 的乘积。记 α_k 表示 A 的第 k 个列向量，β_k 表示 B 的第 k 个行向量，则

$$AB = \sum_{k=1}^{n} \alpha_k \beta_k^{\mathrm{T}} \tag{49}$$

考虑以 $\{p_k\}$ 的概率采样 $X=\frac{1}{p_k}\alpha_k\beta_k^{\mathrm{T}}$，这样不论 p 怎么取，用 X 对 AB 的估计都是无偏的，即

$$\mathbb{E}[X]=\sum_{k=1}^{n}p_k\frac{1}{p_k}\alpha_k\beta_k^{\mathrm{T}}=\sum_{k=1}^{n}\alpha_k\beta_k^{\mathrm{T}}=AB \tag{50}$$

现在我们对“方差”感兴趣。更形式化的说，我们关注的是这个量

$$\mathbb{E}\left[\|AB-X\|_F^2\right] \tag{51}$$

我们希望适当地选取 p 来最小化这个值。一方面，可以找出使这个值最小化的精确结果

$$\begin{aligned}\mathbb{E}\left[\|AB-X\|_F^2\right]&=\sum_{i,j}\mathbb{E}\left[x_{ij}^2\right]-\mathbb{E}\left[x_{ij}\right]^2=\left(\sum_{i,j,k}p_k\left(\frac{1}{p_k}a_{ik}b_{kj}\right)^2\right)-\|AB\|_F^2\\&=\sum_k\frac{1}{p_k}\left(\sum_i a_{ik}^2\right)\left(\sum_j b_{kj}^2\right)-\|AB\|_F^2\\&=\sum_k\frac{1}{p_k}|\alpha_k|^2|\beta_k|^2-\|AB\|_F^2\end{aligned} \tag{52}$$

取 $p_k\propto|\alpha_k||\beta_k|$ 即 $p_k=\frac{|\alpha_k||\beta_k|}{\sum_j|\alpha_j||\beta_j|}$，可以使上式取到最小值。

另一方面，可以估计一下“方差”的上界 (取上述最佳的 p_k 时)。

$$\mathbb{E}\left[\|AB-X\|_F^2\right]\leqslant\left(\sum_k|\alpha_k||\beta_k|\right)^2\leqslant\left(\sum_k|\alpha_k|^2\right)\left(\sum_k|\beta_k|^2\right)=\|A\|_F^2\|B\|_F^2 \tag{53}$$

增加采样次数可以减少“方差”。形式化的，有如下结论：

定理 28. 设 A 是 $m\times n$ 矩阵，B 是 $n\times p$ 矩阵，则矩阵乘积 AB 可以被 CR 来估计，其中 C,R 分别是 $m\times s,s\times p$ 的矩阵，其生成方式为：按照 $\{p_k\}$ 随机采样 $k_1,\cdots,k_s$，并以 $\frac{\alpha_{k_1}}{\sqrt{sp_{k_1}}},\cdots,\frac{\alpha_{k_s}}{\sqrt{sp_{k_s}}}$ 作为 C 的列向量，以 $\frac{\beta_{k_1}}{\sqrt{sp_{k_1}}},\cdots,\frac{\beta_{k_s}}{\sqrt{sp_{k_s}}}$ 作为 R 的行向量，此时有

$$\begin{aligned}\mathbb{E}\left[CC^{\mathrm{T}}\right]&=AA^{\mathrm{T}}\\\mathbb{E}\left[R^{\mathrm{T}}R\right]&=B^{\mathrm{T}}B\\\mathbb{E}\left[\|AB-CR\|_F^2\right]&\leqslant\frac{\|A\|_F^2\|B\|_F^2}{s}\end{aligned} \tag{54}$$

4.3.2 Sketching Matrices

这个问题是比前一个问题 (sketch 矩阵乘积) 要难的。可能听起来有点奇怪，但可以给出一个 intuitive 的解释：假设需要 sketch 一个 $m\times n$ 的矩阵 A，将其写成 $A=AI$ 并考虑套用前述方法去 sketch 等式右边的矩阵乘积。然而 $\|I\|_F^2=n$ 很大，我们只能得到 $\mathbb{E}\left[\|A-X\|_F^2\right]\leqslant\frac{n}{s}\|A\|_F^2$ 的上界。然而如果用全零矩阵去 sketch，得到的 error 也只有 $\|A\|_F^2$。换句话说，上述做法想要做得比 0 矩阵好，就需要 $s>n$，而这是无意义的，因为有这个容量已经可以把 A 完整地表示出来了。

我们考虑的事情是找到一个 “pseudo-identity matrix” $\hat{I}$，使得 $A\approx A\hat{I}$，然后用前面的方法去 sketch 矩阵乘积 $A\hat{I}$。构造方法是这样的：

1. 先按照 length square 采样 A 的行向量 α_k，采样 r 次，得到一个 $r\times n$ 的矩阵 R。

   注意此时 $\mathbb{E}\left[R^{\mathrm{T}}R\right]=A^{\mathrm{T}}A$，且 $\mathbb{E}\left[\|A^{\mathrm{T}}A-R^{\mathrm{T}}R\|_F^2\right]\leqslant\frac{\|A\|_F^4}{r}$，这是因为这个采样恰好符合前面的形式。

2. 构造 $\hat{I}=R^{\mathrm{T}}(RR^{\mathrm{T}})^{-1}R$。$RR^{\mathrm{T}}$ 可能不可逆，此时可以用 SVD 来构造 RR^{T} 的“伪逆”：设 $RR^{\mathrm{T}}=\sum_{i=1}^{r}\sigma_i\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}$ 是 RR^{T} 的 SVD，则取 $\hat{I}=R^{\mathrm{T}}\left(\sum_{i=1}^{r}\frac{1}{\sigma_i}\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}\right)R$。

记 R 行向量张成的空间为 V_R，可以验证 $\forall v \in V_R, \hat{I}v = v$，而 $\forall v \perp V_R, \hat{I}v = 0$。

$\|\hat{I}\|_F^2$ 可以限制住，因为 $\|\hat{I}\|_F^2 = \mathrm{rank}\hat{I} \leqslant r$。从而对矩阵乘积 $A\hat{I}$ 的估计也可以有限制

$$\mathbb{E}\left[\|X - A\hat{I}\|_2^2\right] \leqslant \mathbb{E}\left[\|X - A\hat{I}\|_F^2\right] \leqslant \frac{\|A\|_F^2\|\hat{I}\|_F^2}{s} \leqslant \frac{r}{s}\|A\|_F^2 \tag{55}$$

而 $\|A - A\hat{I}\|_2^2$ 也可以被限制住，这是因为

$$\begin{aligned}
\|A - A\hat{I}\|_2^2 &= \max_v |(A - A\hat{I})v|^2 \\
&= \max_{v \perp V_R} |Av|^2 \\
&= \max_{v \perp V_R} v^{\mathrm{T}} A^{\mathrm{T}} A v \\
&= \max_{v \perp V_R} v^{\mathrm{T}} (A^{\mathrm{T}} A - R^{\mathrm{T}} R) v \\
&\leqslant \|A^{\mathrm{T}} A - R^{\mathrm{T}} R\|_2 \\
&\leqslant \|A^{\mathrm{T}} A - R^{\mathrm{T}} R\|_F \\
&\leqslant \frac{\|A\|_F^2}{\sqrt{r}}
\end{aligned} \tag{56}$$

(由于 $\hat{I}, R$ 的定义是包含随机的，所以严格来说以上的所有东西都需要加上 $\mathbb{E}[]$。)

从而根据 Spectral norm 的三角不等式，$\|X - A\|_2^2$ 也可以被限制住。形式化地，我们知道 $c \leqslant a + b \Rightarrow c^2 \leqslant (a+b)^2 = a^2 + b^2 + 2ab \leqslant 2a^2 + 2b^2$，所以

$$\mathbb{E}\left[\|X - A\|_2^2\right] \leqslant 2\mathbb{E}\left[\|X - A\hat{I}\|_2^2\right] + 2\mathbb{E}\left[\|A - A\hat{I}\|_2^2\right] \leqslant 2\|A\|_F^2 \left(\frac{1}{\sqrt{r}} + \frac{r}{s}\right) \tag{57}$$

5 Random Graphs

定义 12. 一张随机图 $G(n,p)$ 就是一张 n 点完全图，其中每条边独立地以 p 的概率出现。

推论 1 (关于存在性与不存在性的证明技巧).

如果想要证明一个性质在图中**不存在**，可以考虑使用 Markov's Inequality

$$\Pr[X \geqslant 1] \leqslant \mathbb{E}[X] \tag{58}$$

如果想要证明一个性质在图中**存在**，可以考虑使用 Chebyshev's Inequality

$$\Pr[X=0] \leqslant \Pr[|X-\mathbb{E}[X]| \geqslant \mathbb{E}[X]] \leqslant \frac{\operatorname{Var}[X]}{\mathbb{E}[X]^2} \tag{59}$$

由于分析时前者只用到了期望而后者用到了方差，故通常会把前者称为 first moment method，后者称为 second moment method。

定义 13 (相变阈值). 相变阈值 (threshold for phase transitions) 是针对于图的某个性质 $\mathcal{P}$ (比如说，存在 K_3) 的关于 n 的函数 $r(n)$，满足

$$\Pr[G(n,p(n)) \text{ satisfies } \mathcal{P}] \to \begin{cases} 0, & p(n)/r(n) \to 0 \\ 1, & p(n)/r(n) \to \infty \end{cases} \quad (n \to \infty) \tag{60}$$

有些时候甚至可以得到精确的常数，即只要偏离该常数一点就会导致概率收敛至 0 或 1，此时称这个阈值为 sharp threshold，同时也有更精确的形式

$$\Pr[G(n,p(n)) \text{ satisfies } \mathcal{P}] \to \begin{cases} 0, & p(n) < r(n) \\ 1, & p(n) > r(n) \end{cases} \quad (n \to \infty) \tag{61}$$

定理 29 (三角形). “图中存在 K_3” 的相变阈值是 $r(n) = \Theta(n^{-1})$。

证明 24. 对于 $G(n,p)$ 分析。设图中 K_3 的数量为 X，则显然 $\mathbb{E}[X] = \binom{n}{3}p^3 = \Theta(n^3p^3)$，当 $p = o(n^{-1})$ 时，$\mathbb{E}[X] \to 0$ 说明 $\Pr[X \geqslant 1] \to 0$，于是 $r(n) = \Omega(n^{-1})$。

考虑求 $\operatorname{Var}[X]$，主要问题在于求 $\mathbb{E}[X^2]$，也就是枚举两个 K_3 计算同时存在的概率再求和。这里两个 K_3 有三种情况：要么边不交 (完全独立，由 Chebyshev's Inequality 放缩成 $\mathbb{E}[X]^2$)，要么只交一条边 (4 个点 5 条边)，要么完全重合 ($\mathbb{E}[X]$)，因此

$$\mathbb{E}[X^2] \leqslant \mathbb{E}[X]^2 + \Theta(n^4p^5) + \mathbb{E}[X], \qquad \frac{\operatorname{Var}[X]}{\mathbb{E}[X]^2} \leqslant \frac{\mathbb{E}[X] + \Theta(n^4p^5)}{\mathbb{E}[X]^2} = \Theta(n^{-3}p^{-3}) + \Theta(n^{-2}p^{-1}) \tag{62}$$

当 $p = \omega(n^{-1})$ 时，$\frac{\operatorname{Var}[X]}{\mathbb{E}[X]^2} \to 0$ 说明 $\Pr[X=0] \to 0$，于是 $r(n) = O(n^{-1})$。

定理 30 (4-clique). “图中存在 K_4” 的相变阈值是 $r(n) = \Theta(n^{-2/3})$。

证明 25. 设图中 K_4 数量为 X，$\mathbb{E}[X] = \Theta(n^4p^6)$，当 $p = o(n^{-2/3})$ 时 $\mathbb{E}[X] \to 0$，于是 $r(n) = \Omega(n^{-2/3})$。

两个 K_4 相交有 4 种情况：(i)边不交 ($\mathbb{E}[X]^2$)，(ii)交两个点 (6 个点 11 条边)，(iii)交三个点 (5 个点 9 条边)，(iv)完全重合 ($\mathbb{E}[X]$)，故

$$\begin{aligned} \mathbb{E}[X^2] &\leqslant \mathbb{E}[X]^2 + \Theta(n^6p^{11}) + \Theta(n^5p^9) + \mathbb{E}[X] \\ \frac{\operatorname{Var}[X]}{\mathbb{E}[X]^2} &\leqslant \Theta(n^{-2}p^{-1}) + \Theta(n^{-3}p^{-3}) + \Theta(n^{-4}p^{-6}) \end{aligned} \tag{63}$$

当 $p = \omega(n^{-2/3})$ 时，$\frac{\operatorname{Var}[X]}{\mathbb{E}[X]^2} \to 0$ 说明 $\Pr[X=0] \to 0$，于是 $r(n) = O(n^{-2/3})$。

定理 31 (风筝). “图中存在 kite” 相变阈值是 $r(n) = \Theta(n^{-2/3})$。一个 kite 是指一个 K_4 伸出去一个点。

注 2. 注意并不是 $\Theta(n^{-5/7})$，虽然 kite 有 5 个点 7 条边没错，但 kite 包含一个 K_4 作为子图，而 K_4 是需要至少 $\Theta(n^{-2/3})$ 的阈值的。可以证明大于这个值也能使 kite 的出现概率收敛到 1。

证明 26. 风筝数量的期望 $\mathbb{E}[\#\text{kite}]=\Theta(n^5p^7)$，说明 $r(n)=\Omega(n^{-5/7})$，但实际上可以更强，因为要出现风筝必然要出现 K_4，而后者的阈值是 $\Theta(n^{-2/3})$，因此应该有 $r(n)=\Omega(n^{-2/3})$。

两个风筝相交的情况很多，但考虑如果相交了 a 个点和 b 条边，那么就会导致 $\mathrm{Var}[\#\text{kite}]$ 的上界中出现 $\Theta(n^{10-a}p^{14-b})$ 一项，于是 $\Pr[\#\text{kite}=0]\leqslant\frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2}$ 中就有 $\Theta(n^{-a}p^{-b})$ 一项。希望分析得到前者 $\to 0$，那么势必需要让 $p=\omega(n^{-a/b})$。于是考虑取 $-a/b$ 的最大值即 a/b 的最小值，这恰好是图中最大密度子图的密度的倒数。

在这个问题中，风筝的最大密度子图的密度是 $\frac{3}{2}$(有一个 K_4)，故 $r(n)=O(n^{-2/3})$。

综上，$r(n)=\Theta(n^{-2/3})$。

推论 2 (* 自己瞎编的，不是书上写的). “图中存在模式子图 T”的相变阈值是 $\Theta(n^{-1/\rho})$，其中 ρ 是 T 的最大密度子图的密度。

定理 32 (直径至多为 2). “图的直径不超过 2”有 sharp threshold，是 $\sqrt{\frac{2\ln n}{n}}$。

证明 27. 只考虑对于 $G\left(n,p(n)=c\sqrt{\frac{\ln n}{n}}\right)$ 分析，其他量级的 $p(n)$ 只需要根据单调性即可推广。

X 表示图中距离大于 2 的点对数量，一对点 i,j 的距离大于 2，当且仅当边 ij 不存在，且任意 $k\in[1,n]\setminus\{i,j\}$，边 ik 与 jk 不同时存在，因此

$$\begin{aligned}\mathbb{E}[X]&=\binom{n}{2}(1-p)(1-p^2)^{n-2}\\&=\binom{n}{2}\left(1-c\sqrt{\frac{\ln n}{n}}\right)\left(1-c^2\frac{\ln n}{n}\right)^{n-2}\\&=\Theta\left(n^2\mathrm{e}^{-c^2\ln n}\right)=\Theta(n^{2-c^2})\end{aligned}\tag{64}$$

当 $c>\sqrt{2}$ 时，$\mathbb{E}[X]\to 0$，得到 $r(n)\leqslant\sqrt{\frac{2\ln n}{n}}$。

考虑分析 $\mathbb{E}[X^2]$，即枚举两对点 $(i,j),(i',j')$ 距离均大于 2 的概率。有三种情况：(i)点不交，可以用 $\mathbb{E}[X]^2$ 来 bound，(ii)有一个点相交，此时对于确定的某三个点 i,j,k，i 到 j,k 的距离均大于 2 的概率是 $(1-p)^2\left((1-p)+p(1-p)^2\right)^{n-3}$，代入 $p=c\sqrt{\frac{\ln n}{n}}$ 后可分析到 $\Theta(n^{-2c^2})$，(iii)两对点重合，这部分就是 $\mathbb{E}[X]$。综上，可以得到

$$\frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2}\leqslant\frac{\Theta(n^{3-2c^2})+\Theta(n^{2-c^2})}{\Theta(n^{4-2c^2})}\tag{65}$$

当 $c<\sqrt{2}$ 时，$\frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2}\to 0$，于是 $r(n)\geqslant\sqrt{\frac{2\ln n}{n}}$。

综上，$r(n)=\sqrt{\frac{2\ln n}{n}}$。