哈爾濱工業大學（深圳）
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

随 机 算 法

第2章 矩与离差

定理 2.1 **期望的线性性**

对于任意一组有限个具有有限期望的离散型随机变量$X_1, X_2, \dots, X_n$，有

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

引理 2.2

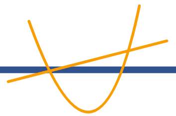
对于任意的常数c和离散型随机变量X，有

$$E[cX] = cE[X]$$

凸函数定义: $\forall x_1, x_2,\ x_1 \neq x_2$

$\forall \lambda,\ 0 \le \lambda \le 1$

$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$



定理 2.3 **詹森不等式**

如果f是一个凸函数，那么

$$E[f(X)] \ge f(E[X])$$

凸函数引理: 若f二次可微, f凸 $\Leftrightarrow f'' \ge 0$

证明: $f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(\xi)}{2}(x-x_0)^2$

$f(x) \ge f(x_0) + f'(x_0)(x-x_0)$

$E[f(x)] \ge E[f(x_0)] + f'(x_0)E[(x-x_0)]$

令 $x_0 = E[x]$, 则

$E[f(x)] \ge \underbrace{E[f(E[x])]}_{\Downarrow \text{常数}} + f'(x_0)\underbrace{[E[x] - E[x_0]]}_{\text{等于}0}$

$E[f(x)] \ge f(E[X])$

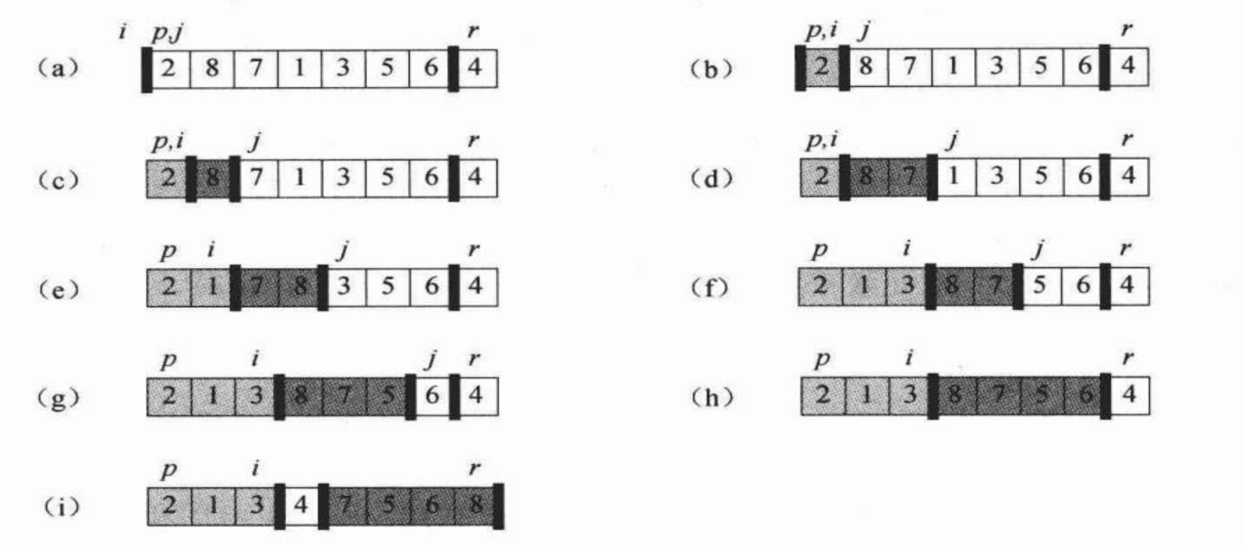
注释：

更一般的例子： $E[X^2] \ge (E[X])^2$ 。

2.1 理论回顾

■ 应用：快速排序的期望运行时间

输入是n个不同的数$x_1, x_2, \ldots, x_n$的列表，选择一个基准元素x，将其他每个元素与x进行比较，进而分成两个子表：小于x的元素和大于x的元素。快速排序法是对这些子表的递归排序。



算法 2.1 快速排序

输入： 全序总体上n个不同元素的列表$S=\{x_1,x_2,\ldots,x_n\}$.

输出： 排序后的S的元素.

1. 如果S只有一个或零个元素，返回S；否则继续.
2. 选择S中一个元素作为基准元素，称为x.
3. 为了将其他元素分成两个子列表，S中的每个其他元素与x作比较；

a) S_1是S中所有比x小的元素

b) S_2是S中所有比x大的元素

最差比较次数 $(n-1)+(n-2)+\cdots+1=\frac{n(n-1)}{2}$

4. 对S_1和S_2进行快速排序.
5. 返回列表S_1，x，S_2.

2.1 理论回顾

- 把算法改成随机选取基准，将快速排序变成一种随机化算法。

①随机选取基准 x_i

②逐一比较，分为子集 S_1，S_2

③对 S_1，S_2 进行排序

期望次数为 $2n\ln n+O(n)$，期望是关于基准的随机选取。

- 可能性是保持原有的确定性算法，用列表的第一个元素作为基准，但考虑输入的概率模型。

期望次数为 $2n\ln n+O(n)$，期望是关于输入的随机选取。

定理 2.4

假设在随机快速排序法中，每一次都是从所有可能中独立且随机地选取基准的，那么对于任意的输入，随机快速排序法所做比较的期望次数为$2n\ln n+O(n)$

n个数，期望比较次数$E(n)$

$n=0,1$时，$E(n)=0$，$n\geq 2$时

$$E(n)=\frac{1}{n-1}\sum_{K=1}^{n-1}\left[E(K)+E(n-K)+n\right]$$

$$=\frac{2}{n-1}\sum_{K=1}^{n-1}E(K)+n$$

记$S(n)=\sum_{K=1}^{n}E(K)$

$$S(n)-S(n-1)=\frac{2}{n-1}S(n-1)+n$$

$$S(n)=\frac{n+1}{n-1}S(n-1)+n$$

$$\frac{1}{n(n+1)}S(n)=\frac{1}{n(n-1)}S(n-1)+\frac{1}{n+1}$$

记$t(n)=\frac{1}{n(n+1)}S(n)$

$$t(n)=t(n-1)+\frac{1}{n+1}$$

$t(1)=0$；

$t(2)=\frac{1}{3}$；

$t(3)=\frac{1}{3}+\frac{1}{4}$；

……

$$t(n-1)=\sum_{i=1}^{n}\frac{1}{i}-\frac{1}{2}-1$$

$$=H(n)-\frac{3}{2}$$

$$\boxed{\begin{array}{l}\text{其中}H(n)\text{为调和级数}\\ \ln n\leq H(n)\leq \ln n+1\\ H(n)=\ln n+O(1)\end{array}}$$

$$t(n)=\frac{1}{n+1}+H(n)-\frac{3}{2}$$

$$=H(n+1)-\frac{3}{2}$$

$\Downarrow$

$$S(n)=n(n+1)t(n)$$

$$E(n)=S(n)-S(n-1)$$

$$=n(n+1)t(n)-n(n-1)t(n-1)$$

$$\Downarrow\ t(n)=t(n-1)+\frac{1}{n+1}$$

$$=2nt(n-1)+n$$

$$=2nH(n)-2n$$

$$=2n\ln n+\underbrace{2nO(1)-2n}$$

$$=2n\ln n+O(n)$$

定理 2.5

假设在随机快速排序法中，每次选取子列表中第一个元素作为基准，如果输入是在其所有可能排列中均匀随机选取的，那么确定性快速排序法所做比较的期望次数为$\boldsymbol{2n \ln n + O(n)}$。

定理 2.6 马尔可夫不等式

设 X 是只取非负值的随机变量，那么对所有$\boldsymbol{a} > \mathbf{0}$，有

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

构造 $I(x) = \mathbb{1}(x \geq a)$，一定有 $I(x) \leq \frac{x}{a}$

$E(I) = \mathbb{P}(x \geq a) \leq E(\frac{x}{a}) = \frac{E(x)}{a}$

注释：

重要性在于不知道**X**的分布($\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{p}_{\boldsymbol{k}}$)情况下，通过$\boldsymbol{E}[\boldsymbol{X}]$估计事件$\{\boldsymbol{X} \geq \boldsymbol{\varepsilon}\}$的概率上限。

定义 2.1

- 一个随机变量 $\boldsymbol{X}$ 的 k 阶矩为 $\boldsymbol{E}[\boldsymbol{X}^k]$.
- 一个随机变量 $\boldsymbol{X}$ 的方差和标准差分别定义为

标准差

$$\mathbf{Var}[X] = \boldsymbol{E}[(X - \boldsymbol{E}[X])^2] = \boldsymbol{E}[X^2] - (\boldsymbol{E}[X])^2 \text{和 } \sigma[\boldsymbol{X}] = \sqrt{\mathbf{Var}[X]}$$

- 两个随机变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的协方差为

$$\mathbf{Cov}(X,Y) = \boldsymbol{E}[(X - \boldsymbol{E}[X])(Y - \boldsymbol{E}[Y])]$$

定理 2.7

对任意两个随机变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$，有

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}(X,Y)$$

定理 2.8 切比雪夫不等式

$|X-E[X]|$ 也是一个 r.v.

$\mathbb{P}(|X-E[X]|\geq a)$

$=\mathbb{P}((X-E(X))^2\geq a^2)$

马尔科夫不等式

$\leq \frac{E[(X-E(X))^2]}{a^2}=\frac{Var[X]}{a^2}$

对任意的$a>0$，有

$$Pr(|X-E[X]|\geq a)\leq \frac{\mathrm{Var}[X]}{a^2}$$

推论 2.9 直接由切比雪夫不等式可得

对任意的 $t>1$，有 $Pr(|X-E[X]|\geq t\cdot\sigma[X])\leq \frac{1}{t^2}$

$$Pr(|X-E[X]|\geq t\cdot E[X])\leq \frac{\mathrm{Var}[X]}{t^2(E[X])^2}$$

2.1 理论回顾

■ 问题描述

假设有n种不同的赠券，每盒麦片内附有其中的一种赠券，每种赠券获取机率相同，而且无限供应.

➢ 假定每盒麦片中的赠券是从n中可能中独立且随机选取的

➢ 假定收集赠券时不与其他人合作

在拥有每种赠券至少一张之前，需要购买多少盒麦片？

2.1 理论回顾

- 令 X 表示收集到每种赠券至少一张所需要购买的麦片盒数，求收集n张赠券的时间X的期望为 $\mathbf{E}[X]$。

- 如果X_i表示恰有$i-1$种不同的赠券时所购买的盒数，那么得到$X=\sum_{i=1}^{n} X_i$，当得到恰好$i-1$种赠券时，再得到一张新赠券的概率是 $p_i=1-\frac{i-1}{n}$，因此

$$E[X_i]=\frac{1}{p_i}=\frac{n}{n-i+1} \Longrightarrow \mathbf{E}[X]=\mathbf{E}[\textstyle\sum_{i=1}^{n} X_i]=\sum_{i=1}^{n} E[X_i]=n\boxed{\sum_{i=1}^{n}\frac{1}{i}}$$

称为调和数 H_n

X_i 服从几何分布 $Ge(p_i)$，$p_i=\frac{n-i+1}{n}$

$$D(X_i)=\frac{1-p_i}{p_i^2}=\frac{n(i-1)}{(n-i+1)^2}$$

$$D(X)=\sum_{i=1}^{n} D(X_i)=\sum_{i=1}^{n}\frac{n(i-1)}{(n-i+1)^2}$$

$$\le\sum_{i=1}^{n}\frac{n^2}{(n-i+1)^2}=n^2\sum_{i=1}^{n}\frac{1}{i^2}$$

$$\le\frac{\pi^2}{6}n^2 \qquad \sum_{i=1}^{\infty}\frac{1}{i^2}=\frac{\pi^2}{6}$$

引理 2.10 调和数$H_n=\sum_{i=1}^{n}\frac{1}{i}$满足$H_n=\ln n+O(1)$.

由马尔可夫不等式得 $Pr(X\geq 2nH_n)\leq\frac{E[X]}{2nH_n}=\frac{1}{2}$

引理2.11 参数为p的几何随机变量的方差是$\frac{1-p}{p^2}$. 期望为$\frac{1}{p}$

利用切比雪夫不等式，得到 切比雪夫不等式可以给出更高的精度

$\mathbb{P}(X\geq 2nH_n)\Leftrightarrow Pr(|X-nH_n|\geq nH_n)\leq\frac{\mathrm{Var}[X]}{(nH_n)^2}=\frac{\pi^2}{6(H_n)^2}=O\left(\frac{1}{\ln^2 n}\right)$

$\ln n\leq H_n\leq \ln n+1$

定义 2.2 中位数 随机变量的中位数

设 X 是随机变量，X 的中位数定义为满足下列条件的m值

$$Pr(X \le m) \ge {}^{1}/_{2} \text{ 和 } Pr(X \ge m) \ge {}^{1}/_{2}$$

$x_1, x_2, \cdots, x_{2k+1} \longrightarrow x_{k+1}$ 中位数是 x_{k+1}

$x_1, x_2, \cdots, x_{2k} \longrightarrow (x_k, x_{k+1})$ 中位数是区间 (x_k, x_{k+1}) 任意数

定理 2.12 对于具有有限期望$E[X]$和有限中位数m的任意随机变量X

1. 期望$E[X]$是使下列表达式最小的c的值 $E\left[(X-c)^2\right]$
2. 中位数m是使下列表达式最小的c的值 $E[|X-c|]$

证明：1. $E[(X-c)^2]$

$= E[X^2-2cX+c^2]$

$= E[X^2]-2cE[X]+c^2$

开口向上抛物线，当且仅当 $c=E[X]$ 时取最小

2、记住就行

定理 2.13 如果X是随机变量，且具有有限标准差σ，期望μ和中位数m，则 $|\mu - m| \le \sigma$

证明期末考

证明：$|\mu - m|$

$= |E(X) - m|$

$= |E(X-m)|$

$\le E(|X-m|)$ 詹森不等式，选取$|X|$为凸函数

$\le E(|X-\mu|)$ 中位数性质

$\le \sqrt{E[(X-\mu)^2]} = \sigma$ 詹森不等式，选取$\sqrt{|X|^2}$为凸函数

■ 计算中位数的随机化算法

目的：找到两个元素$\boldsymbol{d}$和$\boldsymbol{u}$，它们依S的排序是彼此接近的，且中位数$\boldsymbol{m}$于它们之间。

即：1. $\boldsymbol{d} \leq \boldsymbol{m} \leq \boldsymbol{u}$

C的规模要比较小

2.对C $= \{s \in S: d \leq s \leq u\}$，$|C| = \boldsymbol{O}(n/\log n)$ (在$\boldsymbol{d}$和$\boldsymbol{u}$之间的元素总数是少的)

主要思想：从S中有放回的抽样，得到一个含有$\lceil n^{3/4} \rceil$个元素的多重集合R。以大概率保证集合C包含中位数$\boldsymbol{m}$，固定$\boldsymbol{d}$和$\boldsymbol{u}$分别为R的排序的第$\left\lfloor \frac{1}{2} n^{3/4} - \sqrt{n} \right\rfloor$个和第$\left(\left\lceil \frac{1}{2} n^{3/4} + \sqrt{n} \right\rceil\right)$个元素。保证a. 集合C足够大，以大概率包含$\boldsymbol{m}$；b.集合C足够小，以大概率用次线性时间排序。

算法 2.2 随机化中位数算法

输入： 一个全序总体上n个元素的集合S.

输出： S的中位数元素，用m表示.

1. 独立地、均匀随机地，有放回地从S中取出$\lceil n^{3/4} \rceil$个元素组成一个（多重）集合R.
2. 对集合R排序.
3. 设d为排序集合R中第$\left(\left\lfloor \frac{1}{2} n^{3/4} - \sqrt{n} \right\rfloor\right)$个最小元素.
4. 设u为排序集合R中第$\left(\left\lceil \frac{1}{2} n^{3/4} + \sqrt{n} \right\rceil\right)$个最小元素.
5. 将集合S中每个元素与d和u比较，计算集合$C = \{x \in S: d \le x \le u\}$及数$\ell_d = |\{x \in S: x < d\}|$和 $\ell_u = |\{x \in S: x > u\}|$.
6. 如果$\ell_d > n/2$或$\ell_u > n/2$,则输出FAIL.
7. 如果$|C| \le 4n^{3/4}$，则对集合C排序；否则，输出FAIL. 对C进行控制，保障运行时间不过长
8. 输出排序集合C中的第$(\lfloor n/2 \rfloor - \ell_d + 1)$个元素.

2.3 应用

定理 2.14 随机化中位数算法以线性时间结束，而且如果它输出的不是FAIL，则它输出的是输入集合S的正确中位数元素

 证明在下一页

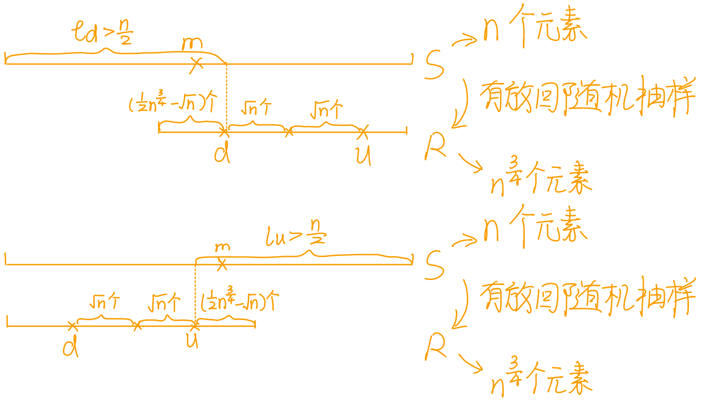
定理 2.15 随机化中位数算法失败的概率的界为$n^{-1/4}$

$$\mathbb{P}\{Fail\} = \mathbb{P}\{\ell_d > \tfrac{n}{2}\} + \mathbb{P}\{\ell_u > \tfrac{n}{2}\} + \mathbb{P}\{|C| > 4n^{\frac{3}{4}}\} \le n^{-\frac{1}{4}}$$

- 中位数的蒙特卡罗随机化算法可以通过重复运行直到成功为止而转化为Las Vegas算法，也就是说把它转化为Las Vegas算法意味着运行时间是可变的，尽管期望运行时间是线性的。

中位数算法失败

- $\ell_d > \frac{n}{2} \iff d > m \iff$ R中元素 $\le$ 中位数m的个数小于 $\frac{1}{2}n^{\frac{3}{4}} - \sqrt{n}$
- $\ell_u > \frac{n}{2} \iff u < m \iff$ R中元素 $\ge$ 中位素m的个数小于 $\frac{1}{2}n^{\frac{3}{4}} - \sqrt{n}$
- $|C| > 4n^{\frac{3}{4}}$



 定理2.15证明

■ 考虑下面三个事件

$$\mathcal{E}_1: \ Y_1 = |\{r \in R | r \le m\}| < \frac{1}{2}n^{3/4} - \sqrt{n}$$ $\Leftrightarrow$ 算法在第6步的 $\ell_d > \frac{n}{2}$ 失败

$$\mathcal{E}_2: \ Y_2 = |\{r \in R | r \ge m\}| < \frac{1}{2}n^{3/4} - \sqrt{n}$$ $\Leftrightarrow$ 算法在第6步的 $\ell_u > \frac{n}{2}$ 失败

$$\mathcal{E}_3: \ |C| > 4n^{3/4}$$ $\Leftrightarrow$ 算法在第七步失败

$$Pr(\mathcal{E}_1) \le \frac{1}{4}n^{-\frac{1}{4}} \qquad Pr(\mathcal{E}_2) \le \frac{1}{4}n^{-1/4} \qquad Pr(\mathcal{E}_3) \le \frac{1}{2}n^{-1/4}$$

证明在下一页

$$\mathbb{P}\{Fail\} = \mathbb{P}\{\ell_d > \tfrac{n}{2}\} + \mathbb{P}\{\ell_u > \tfrac{n}{2}\} + \mathbb{P}\{|C| > 4n^{\frac{3}{4}}\}$$

$$= \mathbb{P}\{\mathcal{E}_1\} + \mathbb{P}\{\mathcal{E}_2\} + \mathbb{P}\{\mathcal{E}_3\} \le n^{-\frac{1}{4}}$$

证明:

① $\mathbb{P}(\mathcal{E}_1)\le \frac{1}{4}n^{-\frac{1}{4}}$

定义随机变量X_i: $X_i=\begin{cases}1, & 第i次抽样\le 中位数m\\ 0, & 其它\end{cases}$

$\mathbb{P}\{X_i=1\}=\dfrac{\frac{(n-1)}{2}+1}{n}=\frac{1}{2}+\frac{1}{2n}$, 事件 $\mathcal{E}_1 \Leftrightarrow Y_1=\sum_{i=1}^{n^{\frac{3}{4}}} < \frac{1}{2}n^{\frac{3}{4}}-\sqrt{n}$

Y_1是伯努利试验的和, Y_1服从二项分布, $Y_1\sim B(n^{\frac{3}{4}}, \frac{1}{2}+\frac{1}{2n})$

$Var(Y_1)= n^{\frac{3}{4}}(\frac{1}{2}+\frac{1}{2n})(\frac{1}{2}-\frac{1}{2n})=\frac{1}{4}n^{\frac{3}{4}}-\frac{1}{4}n^{-\frac{5}{4}}<\frac{1}{4}n^{\frac{3}{4}}$

由切比雪夫不等式:

$\mathbb{P}(\mathcal{E}_1)=\mathbb{P}\{Y_1<\frac{1}{2}n^{\frac{3}{4}}-\sqrt{n}\}\le \mathbb{P}\{|Y_1-E(Y_1)|>\sqrt{n}\}\le \dfrac{Var[Y_1]}{n}<\frac{1}{4}n^{-\frac{1}{4}}$

---

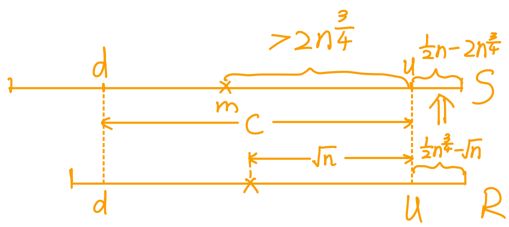
② $\mathbb{P}(\mathcal{E}_2)\le \frac{1}{4}n^{-\frac{1}{4}}$

证明同①

---

③ $\mathbb{P}(\mathcal{E}_3)\le \frac{1}{2}n^{-\frac{1}{2}}$

若$\mathcal{E}_3$发生,则以下两事件至少有一个发生:



$\mathcal{E}_{3,1}$: C中至少$2n^{\frac{3}{4}}$个元素大于中位数

$\mathcal{E}_{3,2}$: C中至少$2n^{\frac{3}{4}}$个元素大于中位数

考虑$\mathcal{E}_{3,1}$发生的概率的界:

若C中至少$2n^{\frac{3}{4}}$个元素大于中位数, 依S从小到大的排序, u的次序$\ge \frac{1}{2}n+2n^{\frac{3}{4}}$

这样R中至少有$\frac{1}{2}n-\sqrt{n}$次抽样是在S的$\frac{1}{2}n-2n^{\frac{3}{4}}$个最大元素中进行的

定义随机变量X_i: $X_i=\begin{cases}1, & 第i次抽样是在S的\frac{1}{2}n-2n^{\frac{3}{4}}个最大元素中\\ 0, & 其它\end{cases}$

$\mathbb{P}\{X_i=1\}=\frac{1}{2}-2n^{-\frac{1}{4}}$

令 $X=\sum_{i=1}^{n^{\frac{3}{4}}}X_i \sim B(n^{\frac{3}{4}}, \frac{1}{2}-2n^{-\frac{1}{4}})$

$E[X]=\frac{1}{2}n^{\frac{3}{4}}-2\sqrt{n}$

$Var[X]= n^{\frac{3}{4}}(\frac{1}{2}-2n^{-\frac{1}{4}})(\frac{1}{2}+2n^{-\frac{1}{4}})=\frac{1}{4}n^{\frac{3}{4}}-4n^{\frac{1}{4}}<\frac{1}{4}n^{\frac{3}{4}}$

由切比雪夫不等式:

$\mathbb{P}\{\mathcal{E}_{3,1}\}=\mathbb{P}\{X\ge\frac{1}{2}n^{\frac{3}{4}}-\sqrt{n}\}\le\mathbb{P}\{|X-E[X]|\ge\sqrt{n}\}\le\dfrac{Var[X]}{n}<\frac{1}{4}n^{-\frac{1}{4}}$

同理: $\mathbb{P}\{\mathcal{E}_{3,2}\}<\frac{1}{4}n^{-\frac{1}{4}}$

从而: $\mathbb{P}\{\mathcal{E}_3\}\le\mathbb{P}\{\mathcal{E}_{3,1}\}+\mathbb{P}\{\mathcal{E}_{3,2}\}<\frac{1}{2}n^{-\frac{1}{4}}$