

# 测度集中

从独立试验到相依随机过程

## 指数尾界、逐步揭示与泊松化

期望给出了随机算法分析的起点，却未必揭示典型行为。指数矩控制独立试验之和，条件期望把逐步揭示信息的过程变成鞅。本讲用这些思想分析球入箱、放大随机算法的成功概率、严格论证 Bloom 过滤器，并将泊松启发式转化为定理。

高级算法课程教学团队

南京大学

第四讲 · 随机算法与随机化技术

## 学习目标与阅读建议

本讲研究怎样从随机量的期望出发，得到具有明确失败概率的保证。读完本讲，应当能够：

1. 推导乘性 Chernoff 界与加性 Hoeffding 界，并根据问题选择合适的形式；
2. 用多数表决或中位数放大随机算法的成功概率；
3. 用 McDiarmid 不等式分析非线性统计量，并严格分析 Bloom 过滤器的错误概率；
4. 用 Doob 鞅与条件取值区间解释并证明有界差分方法；
5. 用泊松化证明最大负载下界与优惠券收集时间的极限。

第一次阅读可依次学习第 1–4 节：独立求和与成功概率放大；有界差分；Bloom 过滤器。McDiarmid 不等式在第 3.1 节陈述，在第 5.5 节证明。先通过应用理解定理的用途，再由第 5 节的条件期望与鞅解释其机制。偏好先读证明的读者，可在第 2 节之后先读第 5 节，查阅定理 3.1 的陈述，再返回各项应用。

第 6 节介绍泊松化、最大负载下界与优惠券收集。这条补充路线使用指数矩与初等条件化，不依赖鞅集中定理。第 7 节总结方法选择。附录 A 整理初等估计，附录 B 给出负相依性的选读推广。所有应用及主要集中定理的完整证明均保留在正文中。

全文中  $\ln$  表示自然对数， $[n] = \{1, \dots, n\}$ 。球入箱结论中的“以高概率”默认指概率为  $1 - O(1/n)$ ；另有明确失败概率时，以该概率为准。渐近阈值针对充分大的  $n$ ，需要时会明确整数取整。每个应用的记号与随机性假设均在相应部分引入。

## 目录

|                                   |    |
|-----------------------------------|----|
| 学习目标与阅读建议                         | i  |
| 1 从期望到高概率保证                       | 1  |
| 2 独立求和：Chernoff、Hoeffding 与成功概率放大 | 2  |
| 3 有界差分与非线性统计量                     | 6  |
| 4 Bloom 过滤器：严格的错误分析               | 9  |
| 5 有界差分为何有效：鞅与条件揭示                 | 11 |
| 6 泊松化与占用阈值                        | 16 |
| 7 如何选择集中性论证                       | 21 |
| 8 习题                              | 21 |
| A 初等概率与分析估计                       | 23 |
| B 负相依性与指数矩界                       | 24 |
| 文献说明                              | 25 |

## 1 从期望到高概率保证

期望描述的是对随机性取平均后的结果。算法保证通常还需要回答：较大的误差、严重过载的位置或异常长的运行，会以多大的概率出现？集中不等式控制的是偏差事件的概率，而不只是随机量的平均大小。一个典型目标是建立如下形式的界：

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq \delta,$$

将偏差阈值  $t > 0$  与允许的失败概率  $0 < \delta < 1$  联系起来。其他问题可能需要单侧尾界，或更精细的极限分布。

### 球入箱模型

共有  $n \geq 2$  个箱子。每个球相互独立、均匀地选择  $[n]$  中的一个落点。投入整数  $m \geq 1$  个球时，我们关心最大负载和空箱数；不断投入新球时，我们关心全部箱子首次被覆盖的时刻。

把  $m$  个球相互独立、均匀地投入  $n$  个箱子。令  $B_j \in [n]$  表示第  $j$  个球的落点， $X_i = \sum_{j=1}^m \mathbf{1}\{B_j = i\}$  表示第  $i$  个箱子的负载。落点  $B_1, \dots, B_m$  相互独立，但负载  $X_1, \dots, X_n$  并不独立，因为其和恒为  $m$ 。

对称性与期望的线性性给出  $\mathbb{E}X_i = m/n$ 。这回答的是固定箱子的平均负载，而非最大负载。特别地， $\max_i \mathbb{E}X_i = m/n$  并不是  $\mathbb{E} \max_i X_i$  的表达式。收集优惠券问题问随机映射何时首次成为满射，最大负载问题则问最大的原像集合有多大。

下面反复出现三个统计量：最大负载  $M = \max_i X_i$ 、满足  $X_i = 0$  的空箱个数，以及所有箱子首次非空的优惠券收集时间。它们各有适合的表示方式。固定箱子的负载是独立指示变量之和；空箱数虽然自然写成相依指示变量之和，却也是独立投球落点的函数；全部覆盖则是涉及所有负载的联合事件。

并合界提供了从单个对象到同时保证的简单途径。对事件  $A_1, \dots, A_n$ ,

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathbb{P}(A_i).$$

因为并事件的指示变量不超过各事件指示变量之和，所以不需要独立性。特别地，若每个箱子的失败概率至多为  $n^{-2}$ ，总失败概率就至多为  $1/n$ 。因此，即使只要求多项式小的失败概率，指数尾界也非常有用。

### 先选择随机性的表示，再选择不等式

独立求和时分解指数矩；面对独立输入的非线性函数时，控制单个输入能造成的变化。后面将用条件揭示解释这一原理，并推广到相依增量。若固定总数妨碍独立性，可以先将总数随机化，再把结论转移回来。

## 2 独立求和：Chernoff、Hoeffding 与成功概率放大

### 2.1 对指数函数使用 Markov 不等式

对非负随机变量  $Z$  及  $a > 0$ , 由  $Z \geq a \mathbf{1}\{Z \geq a\}$  得到  $\mathbb{P}(Z \geq a) \leq \mathbb{E}Z/a$ 。真正有用的选择常常不是  $Z = X$ , 而是  $Z = e^{\lambda X}$ 。定义矩母函数  $M_X(\lambda) = \mathbb{E}e^{\lambda X}$ , 其定义域取为期望有限的参数集合。对  $\lambda > 0$ ,

$$\mathbb{P}(X \geq a) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda a}) \leq e^{-\lambda a} M_X(\lambda). \quad (1)$$

分析下尾时取  $\lambda < 0$ , 此时  $X \leq a$  蕴含  $e^{\lambda X} \geq e^{\lambda a}$ 。指数函数把加法统计量变成乘积, 而独立性可以把乘积的期望分解。参数  $\lambda$  一直保留到最后再优化。

对有界  $X$ , Taylor 展开可以逐项取期望:  $M_X(\lambda) = \sum_{r \geq 0} \lambda^r \mathbb{E}[X^r]/r!$ , 这解释了“矩母函数”名称的来源。对无界随机变量, 必须检查相应的可积性及交换求和与期望的条件。

### 2.2 Bernoulli 矩母函数与乘性尾界

设  $X_1, \dots, X_N$  为相互独立、取值于  $\{0, 1\}$  的随机变量, 记  $p_i = \mathbb{P}(X_i = 1)$ 、 $X = \sum_i X_i$  及  $\mu = \mathbb{E}X = \sum_i p_i$ 。这些成功概率不必相同。取值于  $\{0, 1\}$  的变量称为 Bernoulli 试验; 第 6 节介绍的泊松变量则取非负整数值。

对任意实数  $\lambda$ ,

$$\begin{aligned} \mathbb{E}e^{\lambda X_i} &= 1 + p_i(e^\lambda - 1) \leq \exp(p_i(e^\lambda - 1)), \\ \mathbb{E}e^{\lambda X} &= \prod_{i=1}^N \mathbb{E}e^{\lambda X_i} \leq \exp(\mu(e^\lambda - 1)). \end{aligned} \quad (2)$$

不等式  $1 + u \leq e^u$  对正负  $u$  都成立, 因此这个矩母函数估计可以同时服务于上尾与下尾。若  $\mu = 0$ , 则所有变量几乎必然为零, 集中性问题平凡。

**定理 2.1 (乘性 Chernoff 界)。** 在上述假设下, 若  $\mu > 0$ , 则

$$\mathbb{P}(X \geq (1 + \delta)\mu) \leq \left( \frac{e^\delta}{(1 + \delta)^{1 + \delta}} \right)^\mu, \quad \delta > 0, \quad (3)$$

$$\mathbb{P}(X \leq (1 - \delta)\mu) \leq \left( \frac{e^{-\delta}}{(1 - \delta)^{1 - \delta}} \right)^\mu, \quad 0 < \delta < 1. \quad (4)$$

**证明。** 对上尾, 将指数 Markov 不等式与式 (2) 结合, 得到

$$\mathbb{P}(X \geq (1 + \delta)\mu) \leq \exp\left(\mu(e^\lambda - 1 - \lambda(1 + \delta))\right), \quad \lambda > 0.$$

指数中函数的一阶导数为  $\mu(e^\lambda - (1 + \delta))$ , 二阶导数为正, 故在  $\lambda = \ln(1 + \delta)$  处达到最小值, 给出式 (3)。对下尾取  $\lambda < 0$ , 同理有

$$\mathbb{P}(X \leq (1 - \delta)\mu) \leq \exp\left(\mu(e^\lambda - 1 - \lambda(1 - \delta))\right).$$

此时最优参数为  $\lambda = \ln(1 - \delta) < 0$ , 代入即得式 (4)。

□

### 指数矩方法的基本步骤

选择一个在坏事件上取大值的指数函数，利用概率结构控制其期望，再优化指数参数。并合界或算法重复运行的论证属于后续步骤，不能替代矩母函数估计。

**便于使用的形式。** 利用附录 A 证明的初等对数估计，可简化上面的指数。对  $0 < \delta \leq 1$ ,

$$\mathbb{P}(X \geq (1 + \delta)\mu) \leq e^{-\mu\delta^2/3}, \quad \mathbb{P}(X \leq (1 - \delta)\mu) \leq e^{-\mu\delta^2/2}. \quad (5)$$

当  $\delta = 1$  时，下尾结论可由极限得到，也可直接使用  $\mathbb{P}(X = 0) = \prod_i (1 - p_i) \leq e^{-\mu}$ 。对阈值  $a > \mu$ ，取  $\delta = a/\mu - 1$  得到

$$\mathbb{P}(X \geq a) \leq e^{-\mu} \left(\frac{e\mu}{a}\right)^a \leq 2^{-a} \quad \text{当 } a \geq 2e\mu. \quad (6)$$

## 2.3 最大负载上界

回到  $m$  个球独立、均匀投入  $n \geq 2$  个箱子的模型。固定箱子的负载满足  $X_i \sim \text{Bin}(m, 1/n)$ ，均值为  $\mu = m/n$ 。对实数阈值  $L > \mu$ ，先用完整的 Chernoff 界，再对箱子并合，得到

$$\mathbb{P}(M \geq L) \leq n\mathbb{P}(X_1 \geq L) \leq ne^{-\mu} \left(\frac{e\mu}{L}\right)^L. \quad (7)$$

**定理 2.2 (最大负载上界)。** 以  $1 - O(1/n)$  的概率，

$$M = \begin{cases} O(\ln n / \ln \ln n), & m = n, \\ O(m/n), & m \geq n \ln n. \end{cases}$$

**证明。** 当  $m = n$  时，取整数  $L = \lceil e \ln n / \ln \ln n \rceil$ 。当  $n \rightarrow \infty$  时，

$$L(\ln L - 1) + 1 = (e + o(1)) \ln n \geq 2 \ln n$$

对充分大的  $n$  成立。于是  $\mathbb{P}(X_i \geq L) \leq e^{L-1}/L^L \leq n^{-2}$ ，再用并合界得到失败概率至多为  $1/n$ 。

当  $m \geq n \ln n$  时，取  $L = \lceil 2em/n \rceil$ 。由式 (6)， $\mathbb{P}(M \geq L) \leq n2^{-L} \leq 1/n$ ，因为  $L \geq 2e \ln n$ 。若  $L > m$ ，事件不可能发生。□

对于远大于均值  $\mu = 1$  的负载，小相对偏差形式 (5) 并不合适。完整指数保留了  $L \ln L$  项，正是它产生了  $\ln n / \ln \ln n$  尺度。

**初等替代证明：两次并合界。** 记  $M = \max_{1 \leq i \leq n} X_i$ 。若箱子  $i$  的负载至少为整数  $L \geq 1$ ，就必定存在一个含  $L$  个球的子集，其中所有球都落入该箱子。因此

$$\begin{aligned} \mathbb{P}(M \geq L) &\leq n\mathbb{P}(X_1 \geq L) \\ &\leq n \binom{m}{L} n^{-L} \leq n \left(\frac{em}{nL}\right)^L, \quad 1 \leq L \leq m. \end{aligned} \quad (8)$$

第一次并合针对箱子，第二次并合针对球的子集。独立性用于断言：固定的  $L$  个球都落入固定箱子的概率为  $n^{-L}$ 。当  $L > m$  时，事件不可能发生。最后一步使用阶乘估计  $L! \geq (L/e)^L$ 。

这个计数界不使用指数矩，也能给出相同的量级。当  $m = n$  时， $L = \lceil 3 \ln n / \ln \ln n \rceil$  使  $(e/L)^L \leq n^{-2}$  对充分大的  $n$  成立。当  $m \geq n \ln n$  时， $L = \lceil e^2 m / n \rceil$  使  $em/(nL) \leq e^{-1}$  且  $ne^{-L} \leq 1/n$ 。两个证明都抓住了同一个罕见事件：许多个指定的球同时选择一个箱子。

## 2.4 Hoeffding 引理与加性集中

**引理 2.3 (Hoeffding 引理)。** 若  $Y \in [a, b]$  几乎必然成立，且  $\mathbb{E}Y = 0$ ，则对任意  $\lambda \in \mathbb{R}$ ，

$$\mathbb{E}e^{\lambda Y} \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right). \quad (9)$$

*证明。* 当  $a = b$  时结论显然。否则记  $d = b - a$  及  $p = -a/d \in [0, 1]$ 。由凸性，区间内的  $e^{\lambda y}$  不超过连接端点函数值的弦。对该弦取期望，得到

$$\mathbb{E}e^{\lambda Y} \leq (1-p)e^{\lambda a} + pe^{\lambda b} = e^{-p\lambda d}(1-p + pe^{\lambda d}).$$

对  $0 < p < 1$ ，定义  $h(s) = \ln(1-p + pe^s) - ps$ 。有  $h(0) = h'(0) = 0$ ，且

$$h''(s) = \frac{p(1-p)e^s}{(1-p + pe^s)^2} = q_s(1-q_s) \leq \frac{1}{4}, \quad q_s = \frac{pe^s}{1-p + pe^s}.$$

带积分余项的 Taylor 公式对正负  $s$  都给出  $h(s) \leq s^2/8$ 。代入  $s = \lambda d$  并取指数即可。端点情形  $p = 0, 1$  对应退化分布，同样满足结论。 □

**对称区间的特例。** 对中心化的  $Y \in [-c, c]$  ( $c > 0$ )，同样的弦估计有一个特别简单的形式：

$$\mathbb{E}e^{\lambda Y} \leq \frac{e^{\lambda c} + e^{-\lambda c}}{2} = \sum_{r \geq 0} \frac{(\lambda c)^{2r}}{(2r)!} \leq \sum_{r \geq 0} \frac{(\lambda c)^{2r}}{2^r r!} = e^{\lambda^2 c^2 / 2}.$$

其中的级数比较使用  $(2r)! \geq 2^r r!$ 。这是 Azuma 不等式背后的有界增量指数矩估计。由于区间宽度为  $2c$ ，它与一般 Hoeffding 引理一致。

**定理 2.4 (Hoeffding 不等式)。** 设  $X_1, \dots, X_N$  相互独立，且  $X_i \in [a_i, b_i]$  几乎必然成立。记  $X = \sum_i X_i$ 、 $V = \sum_i (b_i - a_i)^2$ 。若  $V > 0$ ，则对任意  $t > 0$ ，

$$\mathbb{P}(X - \mathbb{E}X \geq t) \leq e^{-2t^2/V}, \quad \mathbb{P}(X - \mathbb{E}X \leq -t) \leq e^{-2t^2/V}, \quad (10)$$

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq 2e^{-2t^2/V}. \quad (11)$$

若  $V = 0$ ，则  $X$  为常数。

*证明。* 将各变量中心化： $Y_i = X_i - \mathbb{E}X_i$ 。中心化不改变取值区间的宽度。由引理 2.3 及独立性，

$$\mathbb{E}e^{\lambda \sum_i Y_i} = \prod_i \mathbb{E}e^{\lambda Y_i} \leq e^{\lambda^2 V / 8}.$$

指数 Markov 不等式把上尾界为  $\exp(-\lambda t + \lambda^2 V/8)$ , 其中  $\lambda > 0$ 。取  $\lambda = 4t/V$  得到上尾界。对  $-\sum_i Y_i$  使用同一论证, 或取  $\lambda = -4t/V$ , 可得下尾界。最后并合两侧事件。

□

对 Bernoulli 变量, 各区间宽度均为 1, 故  $V = N$ 。与乘性 Chernoff 界不同, 加性界的指数取决于取值范围的总预算, 而不取决于均值。两者各有用途, 不能在所有参数范围内相互替代。

## 2.5 多数表决与中位数放大

**判定算法。** 设  $f: \{0,1\}^* \rightarrow \{0,1\}$  为判定问题, 随机算法  $\mathcal{A}$  对每个固定输入  $x$  都满足

$$\mathbb{P}(\mathcal{A}(x) = f(x)) \geq \frac{1}{2} + p, \quad 0 < p \leq \frac{1}{2}.$$

固定目标错误概率  $0 < \delta < 1$ 。

### 多数表决放大

取满足  $r \geq \ln(1/\delta)/(2p^2)$  的奇数  $r$ 。用  $r$  个相互独立的随机种子运行  $\mathcal{A}(x)$ , 返回这  $r$  个答案中的多数答案。

令  $I_j$  表示第  $j$  次是否正确,  $S = \sum_j I_j$ 。则  $\mathbb{E}S \geq r(1/2 + p)$ , 未必等于这个下界。多数表决失败蕴含  $S \leq r/2 \leq \mathbb{E}S - rp$ , 故由 Hoeffding 不等式,

$$\mathbb{P}(\text{多数答案错误}) \leq e^{-2rp^2} \leq \delta \quad \text{当 } r \geq \frac{\ln(1/\delta)}{2p^2}. \quad (12)$$

必要时向上取最近的奇数。这个推导要求各次运行使用新的独立随机性。反复使用同一个随机种子, 可能完全无法降低错误概率。

**数值近似。** 现在假设, 每次独立运行输出的实数以至少  $1/2 + p$  的概率落在固定目标区间  $I_x = [a_x, b_x]$  内。

### 中位数放大

取满足  $r \geq \ln(1/\delta)/(2p^2)$  的奇数  $r$ 。独立运行算法得到  $r$  个数值输出, 返回其中位数。

如果超过半数输出都在  $I_x$  内, 那么无论错误输出多么极端, 中位数必定仍在区间内。沿用上述成功指示变量可得

$$\mathbb{P}(\text{中位数} \notin I_x) \leq e^{-2rp^2}. \quad (13)$$

对非负量  $f(x)$ , 取  $I_x = [(1-\epsilon)f(x), (1+\epsilon)f(x)]$ , 便得到乘性近似保证。对于可能为负的量, 应使用端点顺序明确的区间, 例如  $[f(x) - \epsilon|f(x)|, f(x) + \epsilon|f(x)|]$ 。证明的关键是落入区间的次数, 而不是近似值的算术平均。

投票问题给出了相同偏差尺度的另一个应用: 在  $N$  个相互独立的公平随机投票之外, 增加  $b$  个投向同一方的票, 则该方未获得严格多数的概率至多为  $e^{-b^2/(2N)}$ 。因为失败要求随机票中支持该方的票数比  $N/2$  至少少  $b/2$ 。所以在这个理想化模型中,  $\sqrt{N \ln N}$  量级的额外票数对应多项式小的失败概率。

## 2.6 次高斯变量：共同的表达形式

**定义 2.5.** 若中心化随机变量  $Y$  满足下式，就称其为方差因子为  $v \geq 0$  的次高斯变量，记为  $Y \in \mathcal{G}(v)$ ：

$$\mathbb{E}e^{\lambda Y} \leq e^{\lambda^2 v/2} \quad \text{对任意 } \lambda \in \mathbb{R}. \quad (14)$$

方差因子是矩母函数上界中的参数，并不声称  $\text{Var}(Y) = v$ 。Hoeffding 引理说明，每个中心化且取值于  $[a, b]$  的变量都属于  $\mathcal{G}((b-a)^2/4)$ 。“对任意实数  $\lambda$ ”是定义的一部分。

**推论 2.6.** 若中心化变量  $Y_i \in \mathcal{G}(v_i)$  相互独立，则  $\sum_i Y_i \in \mathcal{G}(\sum_i v_i)$ 。对  $v = \sum_i v_i > 0$  及  $t > 0$ ,

$$\mathbb{P}\left(\sum_i Y_i \geq t\right) \leq e^{-t^2/(2v)}, \quad \mathbb{P}\left(\sum_i Y_i \leq -t\right) \leq e^{-t^2/(2v)}.$$

*证明.* 由独立性 将各边缘矩母函数界相乘，再在  $\lambda = t/v$  处优化  $-\lambda t + \lambda^2 v/2$ 。对变量取负得到另一侧尾界。

□

这个定义把前面使用的高斯型指数矩估计统一起来。接下来不再要求统计量是一个和，而是问：独立输入的一般函数，能否具有同样形式的集中性？第 3 节给出一个容易检查的条件。理解其应用之后，第 5 节再用条件指数矩解释其中的机制。附录 B 则给出另一种基于负相依性的推广，使矩母函数乘积不等式仍然适用。

## 3 有界差分与非线性统计量

### 3.1 从独立求和到独立输入的函数

Hoeffding 不等式控制独立有界输入之和。其取值范围预算有一个简单解释：改变输入  $i$ ，至多使和改变  $b_i - a_i$ 。下面的定理把这一原理推广到一般函数；函数可以是非线性的，自然的求和项之间也可以存在依赖。

**定理 3.1 (McDiarmid 不等式).** 设  $X = (X_1, \dots, X_N)$  的各坐标相互独立，其中  $X_i$  取值于集合  $\Omega_i$ 。设  $f: \prod_{i=1}^N \Omega_i \rightarrow \mathbb{R}$  为可测函数。假设确定的常数  $c_i \geq 0$  满足：对每个  $i$ 、每个  $x \in \prod_j \Omega_j$  及每个  $x'_i \in \Omega_i$ ，都有

$$|f(x_1, \dots, x_i, \dots, x_N) - f(x_1, \dots, x'_i, \dots, x_N)| \leq c_i. \quad (15)$$

记  $C = \sum_{i=1}^N c_i^2$ 。若  $C > 0$ ，则对任意  $t > 0$ ,

$$\begin{aligned} \mathbb{P}(f(X) - \mathbb{E}f(X) \geq t) &\leq \exp(-2t^2/C), \\ \mathbb{P}(f(X) - \mathbb{E}f(X) \leq -t) &\leq \exp(-2t^2/C). \end{aligned} \quad (16)$$

因此，

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp(-2t^2/C). \quad (17)$$

若  $C = 0$ ，则  $f$  在乘积定义域上为常数，任意正偏差的概率均为零。

假设控制的是：固定其他坐标后，把一个坐标从任意一个取值改成任意另一个取值，函数值至多改变  $c_i$ 。这就是逐坐标的 Lipschitz 条件，各  $c_i$  不必相同。这个定理也称为有界差分不等式 [2]。

我们先讨论它的用途。完整证明放在第 5.5 节，届时通过逐一揭示独立输入，解释为什么各坐标变化界的平方可以相加。

**它如何推广 Chernoff–Hoeffding 界。** 对独立变量  $X_i \in [a_i, b_i]$  之和，取

$$f(x_1, \dots, x_N) = \sum_{i=1}^N x_i, \quad c_i = b_i - a_i.$$

McDiarmid 不等式就给出常数完全相同的 Hoeffding 不等式。对独立 Bernoulli 输入，各  $c_i = 1$ ，从而得到单侧加性界  $e^{-2t^2/N}$ 。因此，推广的是从独立输入的和到独立输入的函数，而不是从独立输入推广到任意相依输入。

完整的乘性 Chernoff 界仍是另一项工具：它保留了均值  $\mu$  的信息，而有界差分预算只记录坐标变化。因此，McDiarmid 推广了加性 Chernoff–Hoeffding 界，但并不包含所有依赖均值的 Chernoff 估计。

#### 偏差由什么控制

找出相互独立的随机选择，把统计量写成这些选择的函数，再控制替换每个选择时能造成的变化。输入  $i$  对总预算  $C = \sum_i c_i^2$  贡献  $c_i^2$ 。若  $C > 0$  且  $0 < \delta < 1$ ，偏差达到  $\sqrt{(C/2) \ln(2/\delta)}$  的概率至多为  $\delta$ 。这项计算既不要求求和项独立，也不要求统计量具有加性表示。

### 3.2 相依指示变量与独立随机选择

**空箱数。** 在  $m$  个球投入  $n$  个箱子的模型中，令  $Y$  为最终空箱数。固定箱子为空的概率为  $(1 - 1/n)^m$ ，因而

$$\mathbb{E}Y = n(1 - 1/n)^m. \quad (18)$$

空箱指示变量之间存在依赖，但球的落点相互独立。将空箱数写为

$$Y = f(B_1, \dots, B_m) = n - |\{B_1, \dots, B_m\}|.$$

把一个球从箱子  $u$  移到箱子  $v$ 。若  $u \neq v$ ，移出可能新产生一个空箱，移入可能消除一个空箱；若两种变化同时发生，它们在总数中相互抵消。因此  $Y$  的净变化绝对值至多为 1，而不是 2。对  $m$  个坐标取  $c_j = 1$ ，由式 (17) 得到

$$\mathbb{P}(|Y - n(1 - 1/n)^m| \geq t) \leq 2e^{-2t^2/m}. \quad (19)$$

例如，代入  $t = \sqrt{cm \ln n/2}$ ，失败概率至多为  $2n^{-c}$ 。这是加性估计。当期望空箱数很小时，该加性窗口可能远大于均值；围绕均值集中并不自动等价于相对误差很小。

**重叠模式的出现次数。** 设  $X_1 \dots X_n$  为字母表  $\Sigma$  上的均匀随机字符串，其中  $|\Sigma| = q \geq 1$ 。固定模式  $\pi \in \Sigma^k$ ， $1 \leq k \leq n$ 。令  $Y$  统计所有出现次数，包括相互重叠的出现：

$$Y = \sum_{j=1}^{n-k+1} \mathbf{1}\{X_j X_{j+1} \dots X_{j+k-1} = \pi\}, \quad \mathbb{E}Y = \frac{n-k+1}{q^k}.$$

每个字符参与至多  $k$  个候选窗口。改变它至多影响  $k$  个指示变量，因此  $Y$  是  $n$  个独立字符的函数，坐标变化界满足  $c_i \leq k$ 。故

$$\mathbb{P}\left(\left|Y - \frac{n-k+1}{q^k}\right| \geq t\right) \leq 2e^{-2t^2/(nk^2)}. \quad (20)$$

窗口的重叠使我们不能把这些指示变量当作独立变量。这个证明完全绕开了对其相关结构的分析。

### 3.3 非加性统计量

**超立方体上到集合的距离。** 固定非空集合  $S \subseteq \{0,1\}^n$ ，均匀随机选取  $X \in \{0,1\}^n$ 。定义 Hamming 距离  $H(x,y) = \sum_i |x_i - y_i|$ ，以及到集合的距离

$$D = f_S(X) = \min_{y \in S} H(X,y).$$

改变一位比特，到每个固定  $y$  的距离都至多改变 1，所以这些距离的最小值也至多改变 1。由各比特独立，

$$\mathbb{P}(|D - \mathbb{E}D| \geq t) \leq 2e^{-2t^2/n}, \quad \mathbb{P}\left(|D - \mathbb{E}D| \geq \sqrt{cn \ln n/2}\right) \leq 2n^{-c}. \quad (21)$$

因此，在这个加性尺度上，大多数点到  $S$  的距离差不多。论证并未确定  $\mathbb{E}D$ ，也没有断言大多数点离  $S$  很近。条件  $S \neq \emptyset$  保证最小值有定义。

**赋范空间中的随机符号。** 在任意赋范空间中固定单位向量  $u_1, \dots, u_n$ ，独立、均匀选取符号  $\varepsilon_i \in \{-1,1\}$ 。令  $Z = \|\sum_i \varepsilon_i u_i\|$ 。翻转第  $i$  个符号会使向量和改变  $2u_i$ ，由反三角不等式，范数至多改变 2。于是

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq 2e^{-t^2/(2n)}, \quad \mathbb{P}(|Z - \mathbb{E}Z| \geq s\sqrt{n}) \leq 2e^{-s^2/2}. \quad (22)$$

这里不需要内积或欧氏几何结构。再次强调，不等式控制的是波动，而非均值的位置。

### 3.4 选择独立坐标：随机图染色

随机图  $G \sim G(n,p)$  的顶点集为  $[n]$ ，每条可能边以概率  $p$  独立出现。色数  $\chi(G)$  是对顶点进行正常染色所需的最少颜色数；正常染色要求相邻顶点颜色不同。即使不计算平均色数，也能证明集中性。关键是把随机边按顶点组织成块，而不是将每条边都当作一个单独坐标。

**定理 3.2 (色数的集中性)。** 对  $n \geq 2$ 、 $0 \leq p \leq 1$ 、 $G \sim G(n,p)$  及任意  $s > 0$ ，

$$\mathbb{P}(|\chi(G) - \mathbb{E}\chi(G)| \geq s\sqrt{n}) \leq 2e^{-2s^2}. \quad (23)$$

**证明。** 将边指示变量划分为独立的块  $B_i$ ，其中  $B_i$  包含顶点  $i$  与  $1, \dots, i-1$  之间的边。每条边恰好属于一个块，因此各块相互独立，并共同确定  $G$ 。

固定其他块，比较  $B_i$  的两种取值。对应的图  $G, G'$  仅在与顶点  $i$  相接的边上不同。删除  $i$  后，两者变成同一个图  $H$ 。由于

$$\chi(H) \leq \chi(G) \leq \chi(H) + 1, \quad \chi(H) \leq \chi(G') \leq \chi(H) + 1,$$

两图的色数之差至多为 1：先最优地染色  $H$ ，必要时为  $i$  分配一个新颜色。因此，每个独立块的坐标变化界为 1。对这些独立块应用 McDiarmid 不等式，取  $C \leq n$  及  $t = s\sqrt{n}$ ，即得式 (23)。

□

以每条边为坐标，同样可以得到单位变化界，但预算为  $\binom{n}{2}$ ，而非至多  $n$ 。按顶点分组的表示更好，是因为改变同一顶点处的全部相关边，色数仍至多改变一。两种计算都未确定平均色数。第 5.6 节将把这两种表示解释为边揭示鞅和顶点揭示鞅；Shamir 与 Spencer 的工作使用了这一视角 [6]。

## 4 Bloom 过滤器：严格的错误分析

### 4.1 概率模型与精确条件公式

#### 理想哈希模型中的近似成员查询

设  $U$  为有限的键空间。固定由  $n \geq 1$  个不同键组成的集合  $S \subseteq U$ ，以及数组长度  $m \geq 2$  和整数  $k \geq 1$ 。集合与待查询的非成员均在随机性确定前固定；目标是控制这一次查询的假阳性概率。

选择相互独立、均匀随机的函数  $h_1, \dots, h_k : U \rightarrow [m]$ 。

#### 共享数组的 Bloom 过滤器

将  $A[1], \dots, A[m]$  全部初始化为零。对每个插入的键  $y \in S$ ，将所有  $j \in [k]$  对应的  $A[h_j(y)]$  设为一。查询  $x$  时，当且仅当全部  $k$  个被查位  $A[h_j(x)]$  都为一，才回答“是”。

已插入的键不会发生假阴性。在抽取哈希函数之前固定一个非成员  $x \notin S$ 。在这一理想随机函数模型下，全部  $kn$  个插入位置相互独立且均匀； $k$  个查询位置也相互独立、均匀，并与插入形成的数组独立。这些性质强于两两独立。

令  $W = \sum_{a=1}^m A[a]$  表示置为 1 的比特数，并记

$$q_m = \frac{\mathbb{E}W}{m} = 1 - \left(1 - \frac{1}{m}\right)^{kn}. \quad (24)$$

先固定一个已经生成的数组  $A$ 。各查询位置独立，每次命中 1 的概率为  $W/m$ ，所以条件假阳性概率为  $(W/m)^k$ 。去掉条件时，只需对可能的数组取平均。占用比特数同为  $w$  的数组具有相同的条件错误概率，故按  $w$  分组得到

$$\begin{aligned} \mathbb{P}(\text{假阳性} \mid A) &= \left(\frac{W}{m}\right)^k, \\ p_{\text{FP}} &= \sum_{w=0}^m \left(\frac{w}{m}\right)^k \mathbb{P}(W = w) = \mathbb{E} \left[ \left(\frac{W}{m}\right)^k \right]. \end{aligned} \quad (25)$$

这只是全概率公式所给出的有限加权平均，并未假设各占用比特相互独立。

### 条件独立不等于无条件独立

这  $k$  个测试共享同一个随机数组。把  $\mathbb{E}[(W/m)^k]$  替换为  $(\mathbb{E}W/m)^k$ ，就忽略了这一共同随机性。事实上，对  $k \geq 1$ ，由凸性有  $\mathbb{E}[(W/m)^k] \geq (\mathbb{E}W/m)^k$ 。在控制占用比特数的波动之前，熟悉的  $(1 - e^{-kn/m})^k$  仅是启发式表达式。

## 4.2 有限规模的集中界

由第 3.2 节的空箱数论证， $W$  关于每一个插入位置都是 1-Lipschitz 的，共有  $kn$  个独立插入位置。对  $t > 0$ ，由 McDiarmid 不等式的上尾界，

$$\mathbb{P}(W > \mathbb{E}W + t) \leq \exp\left(-\frac{2t^2}{kn}\right).$$

在其补事件上， $(W/m)^k \leq \min\{1, q_m + t/m\}^k$ ；在例外事件上，该量至多为 1。将式 (25) 中的期望按这两个事件分解，就得到如下有限规模结论。

**定理 4.1 (Bloom 过滤器的假阳性界)。** 在上述理想模型下，对固定集合及固定非成员，

$$q_m^k \leq p_{\text{FP}} \leq \min\left\{1, q_m + \frac{t}{m}\right\}^k + \exp\left(-\frac{2t^2}{kn}\right) \quad (t > 0). \quad (26)$$

对任意  $0 < \eta < 1$ ，取  $t = \sqrt{(kn/2) \ln(1/\eta)}$ ，即可使例外项至多为  $\eta$ 。这给出了不依赖渐近范围的有限规模误差预算。

例如，取  $t = \sqrt{50kn \ln(2n)}$ ，例外概率至多为  $(2n)^{-100} \leq n^{-100}$ ，从而

$$p_{\text{FP}} \leq \min\left\{1, q_m + \frac{\sqrt{50kn \ln(2n)}}{m}\right\}^k + n^{-100}. \quad (27)$$

这是明确的概率保证，而非独立性近似。式 (26) 中的下界说明，用平均占用比例代替随机占用比例，往往会低估错误概率。

## 4.3 渐近公式与 $k$ 的选择

固定  $c > 0$  和整数  $k \geq 1$ ，取  $m = \lceil cn \rceil$ 。则  $q_m \rightarrow 1 - e^{-k/c}$ ，且  $\sqrt{50kn \ln(2n)}/m = o(1)$ 。上下界夹逼给出

$$p_{\text{FP}} = (1 - e^{-k/c})^k + o(1). \quad (28)$$

若要明确记录误差，有  $q_m = 1 - e^{-k/c} + O(1/n)$ ，且  $u \mapsto u^k$  在  $[0, 1]$  上为  $k$ -Lipschitz。因此本论证给出的上界误差为  $O_{c,k}(\sqrt{\ln n/n}) + n^{-100}$ 。不能在有限  $n$  的不等式中直接删去这个占用波动项。

为选择哈希函数个数，考虑连续松弛  $F_c(k) = (1 - e^{-k/c})^k$ 。令  $u = e^{-k/c} \in (0, 1)$ ，则  $\ln F_c(k) = -c \ln u \ln(1 - u)$ 。乘积  $\ln u \ln(1 - u)$  在  $u = 1/2$  处最大（附录 A 给出微积分验证），故松弛最优解为

$$k = c \ln 2, \quad F_c(k) = e^{-c(\ln 2)^2} = (0.6185\dots)^c. \quad (29)$$

实际数据结构必须采用整数  $k$ 。比较  $c \ln 2$  相邻的正整数，选择  $F_c(k)$  较小者；当  $c \ln 2 < 1$  时取  $k = 1$ 。式 (29) 给出的是连续松弛的最优值，并非对任意整数取整都成立的精确保证。

**保证的适用范围**

集合与非成员查询必须独立于随机哈希函数而预先固定。这个证明并未对自适应选择的查询，或仅两两独立的哈希族建立相同结论。数组占用  $m$  比特，一次查询需要  $k$  次哈希求值；完全随机函数是模型假设，而不是短种子的实现方案。

以上分析只使用了 McDiarmid 不等式的结论、占用比特数的单坐标变化界，以及初等条件化。下面转向这些非线性集中界背后的共同机制。

## 5 有界差分为何有效：鞅与条件揭示

前面的应用把集中性归结为一个条件：单个独立输入能造成多大的变化？即使统计量是非线性的，为何这个条件仍然足够？逐一揭示输入，并跟踪最终函数值的条件估计，就能将总变化写成一系列小增量之和，且每个增量的条件均值为零。这不仅将证明定理 3.1，也会解释在相依情形下如何继续证明集中性。

独立求和时，指数矩可以直接分解。新的增量未必独立，因此要用逐步条件化替代分解。先建立完成这一操作所需的平均规则。

### 5.1 条件期望与已揭示的信息

对可积离散随机变量  $Z$  以及满足  $\mathbb{P}(E) > 0$  的事件  $E$ ，定义

$$\mathbb{E}[Z | E] = \sum_z z \mathbb{P}(Z = z | E) = \frac{\mathbb{E}[Z \mathbf{1}_E]}{\mathbb{P}(E)}. \quad (30)$$

对离散  $X$ ，在每个具有正概率的取值上定义  $g(x) = \mathbb{E}[Z | X = x]$ ，再置  $\mathbb{E}[Z | X] = g(X)$ 。在  $X$  的支撑集外如何定义  $g$ ，都不影响这个随机变量的几乎必然意义。以元组  $(X_1, \dots, X_i)$  为条件时，只是把标签改为向量。

一个人口例子可以区分函数及其随机求值。从有限总体中均匀选一个人，令  $Z$  为身高， $X$  为国家标签。则  $g(x)$  是国家  $x$  内的平均身高， $g(X)$  是被抽中者所属国家的平均身高。等价地，可以按人口比例抽取国家，再返回该国的平均身高。均匀抽取国家通常会给出另一个分布。

对于有限信息划分  $\mathcal{P}$ ， $\mathbb{E}[Z | \mathcal{P}]$  在每个样本点处取其所在划分块内  $Z$  的平均值。一般记号写作  $\mathbb{E}[Z | \mathcal{F}]$ ，其中  $\mathcal{F}$  为相应的  $\sigma$ -代数。以下恒等式均在几乎必然意义下成立，并假定涉及的随机量及乘积可积。

**线性性与提出已知因子。** 条件期望具有线性性。对常数  $a, b$  及可积的  $U, V$ ，

$$\mathbb{E}[aU + bV | X] = a\mathbb{E}[U | X] + b\mathbb{E}[V | X].$$

若一旦知道  $X$  就知道  $h(X)$ ，则

$$\mathbb{E}[h(X)Z | X] = h(X)\mathbb{E}[Z | X]. \quad (31)$$

验证时，固定条件  $X = x$ 。因子  $h(X)$  变成常数  $h(x)$ ，可以提出定义条件均值的有限和或绝对收敛级数。这里不需要  $X$  与  $Z$  独立。特别地，

$$\mathbb{E}[h(X) | X] = h(X), \quad \mathbb{E}[X | X] = X.$$

若  $Z$  与  $X$  独立, 其条件分布等于原分布, 故  $\mathbb{E}[Z | X] = \mathbb{E}Z$ 。这是独立性的另一种应用, 并非仅由线性性推出。

**全期望公式与塔式法则。**

**命题 5.1。** 对可积的  $Z$  及离散标签  $X, U$ ,

$$\mathbb{E}[\mathbb{E}[Z | X]] = \mathbb{E}Z, \quad (32)$$

$$\mathbb{E}[\mathbb{E}[Z | X, U] | U] = \mathbb{E}[Z | U]. \quad (33)$$

更一般地, 若  $\mathcal{G} \subseteq \mathcal{F}$ , 则

$$\mathbb{E}[\mathbb{E}[Z | \mathcal{F}] | \mathcal{G}] = \mathbb{E}[Z | \mathcal{G}]. \quad (34)$$

**证明。** 对全期望公式, 按照条件标签求和:

$$\begin{aligned} \mathbb{E}[\mathbb{E}[Z | X]] &= \sum_x \sum_z z \mathbb{P}(Z = z | X = x) \mathbb{P}(X = x) \\ &= \sum_z z \sum_x \mathbb{P}(Z = z, X = x) = \mathbb{E}Z. \end{aligned}$$

对塔式法则, 在每个具有正概率的事件  $U = u$  内进行同样的计算。按照  $X | U = u$  对更细的条件均值再取平均, 就得到  $Z | U = u$  的均值。用划分解释, 就是先在细块内部平均, 再对组成同一粗块的细块求加权平均, 每个块的权重是其条件概率。

□

在上述人口例子中, 可以先固定一个性别, 再按该群体中各国人口所占比例, 对各国在该群体内的平均身高加权。结果就是给定该性别时的平均身高。这仍是同一加权平均规则, 只不过限制在一个条件总体内。

方向非常重要: 外层条件保留的是较粗的信息。若  $\mathcal{F}_{i-1} \subseteq \mathcal{F}_i$ , 用旧信息对细化后的估计  $\mathbb{E}[Z | \mathcal{F}_i]$  再平均, 就会返回原来的估计。下面的随机过程正是由这一观察引出的。

把塔式法则与式 (31) 结合, 就得到一个常用的迭代期望恒等式:

$$\mathbb{E}[\mathbb{E}[f(X)g(X, Y) | X]] = \mathbb{E}[f(X)\mathbb{E}[g(X, Y) | X]]. \quad (35)$$

在指数矩论证中, 已知因子就是过去部分和的指数, 剩下的条件平均只涉及新增量。

## 5.2 Doob 鞅: 逐步细化估计

可以把条件期望理解为信息随时间逐步揭示。滤过  $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \dots$  记录每个时刻已知的信息。在本讲的有限情形下, 它可看作样本空间上一系列越来越细的划分。若第  $i$  步掌握的信息已经确定某变量的值, 就称该变量为  $\mathcal{F}_i$ -可测。

设  $Z = f(X_1, \dots, X_N)$  可积, 此处不要求输入独立。令  $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$ , 其中  $\mathcal{F}_0$  为平凡信息, 并定义

$$Y_i = \mathbb{E}[Z | \mathcal{F}_i], \quad Y_0 = \mathbb{E}Z, \quad Y_N = Z. \quad (36)$$

第  $i$  个值是在已揭示前缀的条件下，对尚未揭示的坐标取平均，而不是简单地把未知坐标代为零。初始时尚未观察任何随机性；最后一步时， $f$  的值已完全确定。

**定义 5.2 (鞅)。** 若可积随机变量序列  $Y_0, Y_1, \dots$  满足  $Y_i$  为  $\mathcal{F}_i$ -可测，且

$$\mathbb{E}[Y_i | \mathcal{F}_{i-1}] = Y_{i-1} \quad (i \geq 1), \quad (37)$$

就称其为关于  $(\mathcal{F}_i)$  的鞅。它每一步增量的条件期望为零。将等号换成  $\leq$  得到上鞅，换成  $\geq$  得到下鞅。

这里， $\sigma(X_0, \dots, X_i)$  表示所列变量确定的全部信息。取  $\mathcal{F}_i = \sigma(X_0, \dots, X_i)$ ，就得到关于某个底层随机序列所揭示信息的鞅。取  $\mathcal{F}_i = \sigma(Y_0, \dots, Y_i)$ ，则得到用过程自身历史定义的版本。这里明确加入可积性，以保证条件期望有定义。

**命题 5.3 (Doob 构造)。** 式 (36) 定义的序列是关于  $(\mathcal{F}_i)$  的鞅。

*证明。* 每个  $Y_i$  均由已揭示的前缀确定。由塔式法则 (34)，

$$\mathbb{E}[Y_i | \mathcal{F}_{i-1}] = \mathbb{E}[\mathbb{E}[Z | \mathcal{F}_i] | \mathcal{F}_{i-1}] = \mathbb{E}[Z | \mathcal{F}_{i-1}] = Y_{i-1}.$$

□

这个证明没有使用独立性。证明第 3 节的有界差分定理时，独立性将用于控制每一步细化估计的条件取值范围。

**把非线性统计量写成增量之和。** 令  $D_i = Y_i - Y_{i-1}$ ，则

$$Z - \mathbb{E}Z = Y_N - Y_0 = \sum_{i=1}^N D_i, \quad \mathbb{E}[D_i | \mathcal{F}_{i-1}] = 0.$$

第一个等式来自望远镜求和，条件均值零则来自鞅性质。以空箱数为例， $Y_i$  是看到前  $i$  个落后对最终空箱数的条件估计，而不是当前的空箱数。我们重新得到一个和，但其增量可以相互依赖。接下来的集中论证将使用条件均值与条件取值范围，而不再要求增量独立。

**公平增量与自适应选择。** 基本例子是独立公平符号  $\varepsilon_i \in \{-1, 1\}$  的部分和  $Y_i = \sum_{j \leq i} \varepsilon_j$ 。由于  $\varepsilon_i$  独立于过去且均值为零， $\mathbb{E}[Y_i | \varepsilon_1, \dots, \varepsilon_{i-1}] = Y_{i-1}$ 。更一般地，若有界的下注量  $b_i$  由过去的信息决定，则  $\sum_{j \leq i} b_j \varepsilon_j$  仍为鞅。这就是公平游戏的直观含义：选择可以适应历史，但下一步收益的条件期望仍为零。

#### 加倍下注并不消除风险

考虑一种每次输钱后将赌注加倍的策略。连续输  $r$  次后首次获胜，净收益为  $2^r - \sum_{j=0}^{r-1} 2^j = 1$ ；但随着  $r$  增长，所需赌注没有上界。任意长的连续失败都有正概率。赌本有限时，策略可能在下一次获胜前就无法继续；这个形式上的净收益计算，并不意味着可以无风险获利。

公平随机游走也具有 Markov 性，但鞅的定义不要求 Markov 性。鞅条件只约束条件均值，而不约束下一状态的完整条件分布。

### 5.3 条件区间形式的 Azuma-Hoeffding 不等式

即使条件均值为零，条件取值区间也未必以零为中点。Hoeffding 引理保留的信息是这个区间的宽度。

**定理 5.4 (Azuma-Hoeffding 不等式：条件区间形式)。** 设  $(Y_i)_{i=0}^N$  为关于  $(\mathcal{F}_i)$  的鞅，记  $D_i = Y_i - Y_{i-1}$ 。假设存在  $\mathcal{F}_{i-1}$ -可测的端点  $a_i, b_i$  以及确定的常数  $w_i \geq 0$ ，使得几乎必然有

$$a_i \leq D_i \leq b_i, \quad b_i - a_i \leq w_i.$$

记  $W = \sum_{i=1}^N w_i^2$ 。若  $W > 0$ ，则对任意  $t > 0$ ，

$$\begin{aligned} \mathbb{P}(Y_N - Y_0 \geq t) &\leq e^{-2t^2/W}, \\ \mathbb{P}(Y_N - Y_0 \leq -t) &\leq e^{-2t^2/W}, \\ \mathbb{P}(|Y_N - Y_0| \geq t) &\leq 2e^{-2t^2/W}. \end{aligned} \tag{38}$$

若  $W = 0$ ，则每个增量几乎必然为零。

*证明。* 由鞅性质， $\mathbb{E}[D_i | \mathcal{F}_{i-1}] = 0$ 。给定过去信息后，区间端点已经确定。对条件分布应用 Hoeffding 引理，对任意实数  $\lambda$  得到

$$\mathbb{E}[e^{\lambda D_i} | \mathcal{F}_{i-1}] \leq \exp(\lambda^2 w_i^2 / 8). \tag{39}$$

端点可以依赖于过去，但宽度上界  $w_i$  是确定的。

令  $S_i = \sum_{j=1}^i D_j = Y_i - Y_0$ ，并置  $S_0 = 0$ 。增量不必独立，因此不能将无条件边缘矩母函数直接相乘。正确做法是利用  $S_{i-1}$  在第  $i-1$  步已经确定这一事实，由塔式法则和提出已知因子的规则得到

$$\begin{aligned} \mathbb{E}e^{\lambda S_i} &= \mathbb{E}[\mathbb{E}[e^{\lambda S_{i-1}} e^{\lambda D_i} | \mathcal{F}_{i-1}]] \\ &= \mathbb{E}[e^{\lambda S_{i-1}} \mathbb{E}[e^{\lambda D_i} | \mathcal{F}_{i-1}]] \\ &\leq e^{\lambda^2 w_i^2 / 8} \mathbb{E}e^{\lambda S_{i-1}}. \end{aligned} \tag{40}$$

从  $i = N$  递推到 1，得到

$$\mathbb{E}e^{\lambda(Y_N - Y_0)} \leq \exp\left(\frac{\lambda^2}{8} \sum_{i=1}^N w_i^2\right) = e^{\lambda^2 W / 8}.$$

对  $\lambda > 0$ ，指数 Markov 不等式将上尾概率界为  $\exp(-\lambda t + \lambda^2 W / 8)$ 。最优参数为  $\lambda = 4t / W$ ，给出  $e^{-2t^2 / W}$ 。对取负后的过程使用同一论证可得下尾界，最后将两侧概率相加。若  $W = 0$ ，每个条件区间宽度均为零；再由条件均值为零，每个增量都只能为零。□

#### 统一的条件指数矩论证

更一般地，若对确定的  $v_i \geq 0$  及任意实数  $\lambda$ ，有  $\mathbb{E}[e^{\lambda D_i} | \mathcal{F}_{i-1}] \leq e^{\lambda^2 v_i / 2}$ ，同一迭代就说明  $\sum_i D_i$  为方差因子  $\sum_i v_i$  的次高斯变量。当总因子为正时，其单侧尾界为  $e^{-t^2 / (2 \sum_i v_i)}$ 。条件区间定理使用的是  $v_i = w_i^2 / 4$ 。

## 5.4 有界增量的 Azuma 不等式

**推论 5.5 (Azuma 不等式)。** 设  $(Y_i)_{i=0}^N$  为鞅, 且对确定的  $c_i \geq 0$  及  $i \geq 1$ , 有  $|Y_i - Y_{i-1}| \leq c_i$  几乎必然成立。记  $C = \sum_i c_i^2 > 0$ , 则对任意  $t > 0$ ,

$$\mathbb{P}(Y_N - Y_0 \geq t) \leq e^{-t^2/(2C)}, \quad \mathbb{P}(Y_N - Y_0 \leq -t) \leq e^{-t^2/(2C)}, \quad (41)$$

$$\mathbb{P}(|Y_N - Y_0| \geq t) \leq 2e^{-t^2/(2C)}. \quad (42)$$

*证明.* 给定过去信息后,  $D_i = Y_i - Y_{i-1}$  位于区间  $[-c_i, c_i]$ , 其宽度为  $2c_i$ 。在定理 5.4 中取  $w_i = 2c_i$  即可。从指数矩看, 此时的条件估计是

$$\mathbb{E}[e^{\lambda D_i} \mid \mathcal{F}_{i-1}] \leq e^{\lambda^2 c_i^2 / 2}. \quad (43)$$

所以方差因子是  $\sum_i c_i^2$ , 而不是  $\sum_i c_i^2 / 4$ 。□

绝对增量界与条件区间宽度界是不同的假设。保留实际条件宽度, 可使指数改善四倍; 理解这一区别是正确推导 McDiarmid 不等式的关键。

## 5.5 McDiarmid 不等式的证明

现在完成第 3.1 节暂缓的证明。只剩下一个关键步骤: 将改变单个输入时的函数变化界, 转化为单个 Doob 增量的条件区间宽度界。若仅用  $|D_i| \leq c_i$  代入通常的 Azuma 不等式, 指数会损失四倍; 条件区间形式则保留了所需信息。

*定理 3.1 的证明.* 构造 Doob 鞅  $M_i = \mathbb{E}[f(X) \mid X_1, \dots, X_i]$ , 其中  $M_0 = \mathbb{E}f(X)$ 、 $M_N = f(X)$ 。只需控制增量  $D_i = M_i - M_{i-1}$  的条件取值区间。

固定已经揭示的前缀  $x_{<i} = (x_1, \dots, x_{i-1})$ 。对坐标  $i$  的可能取值  $a$ , 令

$$g_i(a) = \mathbb{E}[f(x_{<i}, a, X_{>i})], \quad X_{>i} = (X_{i+1}, \dots, X_N).$$

独立性保证: 无论  $a$  取何值, 这个平均都使用同一个未来分布。对任意  $a, b \in \Omega_i$ ,

$$|g_i(a) - g_i(b)| \leq \mathbb{E}[|f(x_{<i}, a, X_{>i}) - f(x_{<i}, b, X_{>i})|] \leq c_i.$$

因此,  $g_i$  的可能取值位于宽度至多为  $c_i$  的区间内。给定前缀后,  $M_i = g_i(X_i)$ , 而  $M_{i-1}$  是它的条件均值。减去均值只平移区间, 不改变其宽度。所以 Doob 增量的条件均值为零, 条件取值区间宽度至多为  $c_i$ ; 等价地, 条件 Hoeffding 引理给出

$$\mathbb{E}[e^{\lambda D_i} \mid X_1, \dots, X_{i-1}] \leq e^{\lambda^2 c_i^2 / 8}. \quad (44)$$

在定理 5.4 中取  $w_i = c_i$ , 即得两个单侧界及其并合。若  $C = 0$ , 逐一改变坐标便知  $f$  为常数。□

### 证明揭示了什么

对同一个独立未来取平均, 保留了单坐标变化界; 减去条件均值, 不会扩大所得区间。再由条件 Hoeffding 引理及反复条件化, 各区间宽度的平方得以相加。这就证明了前面的各项有界差分应用, 而无须假设其表面的求和项相互独立。

## 5.6 从独立到相依：平均有界差分

设  $Z = f(X_1, \dots, X_N)$  可积，输入坐标可以相依。若确定的  $c_i \geq 0$  满足：几乎必然有

$$|\mathbb{E}[f(X) \mid X_1, \dots, X_i] - \mathbb{E}[f(X) \mid X_1, \dots, X_{i-1}]| \leq c_i, \quad (45)$$

就可以将 Doob 构造与 Azuma 不等式结合。记  $C = \sum_i c_i^2 > 0$ ，得到

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2 \exp\left(-\frac{t^2}{2C}\right). \quad (46)$$

这称为平均有界差分方法。“平均”指每个差分内部的条件平均；对差分的上界必须几乎必然成立，而不是只在再取期望后成立。单侧尾概率的上界为  $e^{-t^2/(2C)}$ 。

输入相依时，仅有逐坐标的 Lipschitz 性质，并不能推出式 (45)。揭示一个坐标可能改变全部未揭示坐标的条件分布。例如，令  $X_1 = \dots = X_N = B$ ，其中  $B$  为公平比特， $f(x) = \sum_i x_i$ 。改变  $f$  的一个坐标至多使其改变 1，但揭示  $X_1$  就会使条件估计从  $N/2$  变成 0 或  $N$ 。第一个 Doob 增量的绝对值为  $N/2$ 。

**重新理解边揭示与顶点揭示。** 对顶点集为  $[n]$  的随机图  $G$  及可积统计量  $f(G)$ ，将所有可能边编号，其指示变量记为  $I_1, \dots, I_m$ ，其中  $m = \binom{n}{2}$ 。边揭示鞅为

$$Y_j = \mathbb{E}[f(G) \mid I_1, \dots, I_j], \quad 0 \leq j \leq m.$$

也可以用  $G[i]$  表示  $[i]$  上的诱导子图。由于这些诱导子图相互嵌套，揭示的信息逐步增加， $\mathbb{E}[f(G) \mid G[i]]$  ( $0 \leq i \leq n$ ) 就是顶点揭示鞅。取  $G[0]$  为空图，初始值为  $\mathbb{E}f(G)$ 。这两种构造都不要求边相互独立。

在  $G(n, p)$  中，第 3.4 节的独立块  $B_i$  满足：前  $i$  个块恰好揭示  $G[i]$ 。因此，前面的图染色论证控制的正是顶点揭示增量的条件区间宽度。换一种随机图分布后，Doob 序列仍然是鞅，但必须针对该分布重新建立增量界或条件区间界，不能直接沿用独立块的证明。

## 6 泊松化与占用阈值

### 6.1 将总数随机化以获得独立性

最大负载的上界只需要固定箱子的尾界与一次并合。下界问的是另一个问题：所有箱子的负载都很小，是否不太可能？若能把各箱子的概率相乘，论证会很方便；但恒等式  $\sum_i X_i = m$  妨碍了独立性。

泊松化改变实验本身：不再固定投球次数，而是先选择一个适当的随机总数，再独立、均匀地投入这么多个球。此时各箱子的负载变为独立的泊松变量。将其总数条件化为  $m$ ，便精确还原原始联合分布。关键是判断：去掉这一条件后，所关心事件的概率会改变多少。

### 6.2 泊松变量与精确条件恒等式

泊松随机变量  $Z \sim \text{Pois}(\mu)$  ( $\mu > 0$ ) 的概率质量函数为

$$\mathbb{P}(Z = j) = e^{-\mu} \frac{\mu^j}{j!}, \quad j = 0, 1, 2, \dots$$

直接求和，对任意实数  $\lambda$  都得到精确矩母函数

$$\mathbb{E}e^{\lambda Z} = e^{-\mu} \sum_{j \geq 0} \frac{(\mu e^\lambda)^j}{j!} = \exp(\mu(e^\lambda - 1)). \quad (47)$$

这恰好是分析 Bernoulli 试验之和时使用的矩母函数上界。在零点求导得到  $\mathbb{E}Z = \mu$ 、 $\text{Var } Z = \mu$ 。独立泊松变量的矩母函数相乘，还说明其和仍服从泊松分布，参数为各均值之和。

**命题 6.1 (泊松尾界)。** 对实数阈值  $a > 0$ ,

$$\mathbb{P}(Z \geq a) \leq e^{-\mu} \left( \frac{e\mu}{a} \right)^a, \quad a > \mu, \quad (48)$$

$$\mathbb{P}(Z \leq a) \leq e^{-\mu} \left( \frac{e\mu}{a} \right)^a, \quad 0 < a < \mu. \quad (49)$$

*证明.* 把式 (47) 代入指数 Markov 不等式，最小化  $\mu(e^\lambda - 1) - \lambda a$ 。驻点为  $\lambda = \ln(a/\mu)$ ：上尾时为正，下尾时为负。当  $a = 0$  时，下尾概率恰为  $e^{-\mu}$ 。

□

与式 (5) 相同的简化给出：对  $0 < t \leq \mu$ ,

$$\mathbb{P}(Z \geq \mu + t) \leq e^{-t^2/(3\mu)}, \quad \mathbb{P}(Z \leq \mu - t) \leq e^{-t^2/(2\mu)}. \quad (50)$$

**条件化后的联合分布。** 令  $X = (X_1, \dots, X_n)$  为恰好  $m$  次均匀投球的负载向量。以下定理将它与独立泊松负载精确联系起来。

**定理 6.2 (精确泊松条件化)。** 对任意  $\lambda > 0$ ，设  $Y_1, \dots, Y_n$  为独立的  $\text{Pois}(\lambda)$  变量， $S = \sum_i Y_i$ ，则

$$(X_1, \dots, X_n) \stackrel{d}{=} (Y_1, \dots, Y_n \mid S = m). \quad (51)$$

*证明.* 对和为  $m$  的非负整数  $x_1, \dots, x_n$ ，负载向量的多项分布概率为

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \frac{m!}{x_1! \cdots x_n!} n^{-m}.$$

由于  $S \sim \text{Pois}(n\lambda)$ ，对应的泊松条件概率为

$$\frac{\prod_{i=1}^n (e^{-\lambda} \lambda^{x_i} / x_i!)}{e^{-n\lambda} (n\lambda)^m / m!} = \frac{m!}{x_1! \cdots x_n! n^m}.$$

两个分布在总和不为  $m$  的向量上概率都为零。因此相等的是整个联合分布，而不只是各边缘分布。

□

等价的构造是先抽取  $S \sim \text{Pois}(m)$ ，再独立、均匀地投入  $S$  个球。其无条件负载为独立的  $\text{Pois}(m/n)$  变量；给定  $S = m$  后，就还原为原始实验。

### 6.3 一般的泊松化比较界

**定理 6.3** (损失因子为  $e\sqrt{m}$  的比较界)。设  $m \geq 1$ ,  $X$  为  $m$  个球投入  $n$  个箱子的负载向量,  $Y_i$  为独立的  $\text{Pois}(m/n)$  变量。对任意非负函数  $f$ ,

$$\mathbb{E}f(X) \leq e\sqrt{m} \mathbb{E}f(Y). \quad (52)$$

特别地, 泊松化模型中概率为  $q$  的事件, 在固定总数模型中的概率至多为  $e\sqrt{m}q$ 。

*证明.* 利用非负性, 只保留总数为  $m$  的一项:

$$\mathbb{E}f(Y) = \sum_{s \geq 0} \mathbb{E}[f(Y) \mid S = s] \mathbb{P}(S = s) \geq \mathbb{E}f(X) \mathbb{P}(S = m).$$

由附录 A 的阶乘估计  $m! \leq e\sqrt{m}(m/e)^m$ ,

$$\mathbb{P}(S = m) = e^{-m} \frac{m^m}{m!} \geq \frac{1}{e\sqrt{m}}.$$

整理即得结论。令  $f$  为事件指示函数, 便得到概率版本。 □

这个定理给出的是单侧比较, 而非断言所有事件概率都渐近相等。当坏事件在泊松化模型中已经极不可能时, 它尤其有效。对于非平凡的常数概率极限, 需要更精细的论证; 后面将对优惠券收集给出这样的证明。

### 6.4 最大负载下界

取  $m = n$ , 则泊松化后的各负载为独立的  $\text{Pois}(1)$  变量。对整数  $L \geq 1$ ,

$$\begin{aligned} \mathbb{P}\left(\max_i Y_i < L\right) &= \mathbb{P}(Y_1 < L)^n \\ &\leq (1 - \mathbb{P}(Y_1 = L))^n = \left(1 - \frac{1}{eL!}\right)^n \leq \exp\left(-\frac{n}{eL!}\right). \end{aligned} \quad (53)$$

独立性使“全部箱子的负载都小”变得易于分析。取  $L = \lfloor \ln n / \ln \ln n \rfloor$ 。为估计阶乘项, 记  $A = \ln n$ ,  $B = \ln \ln n$ ,  $D = \ln \ln \ln n$ 。对充分大的  $n$ ,  $L \leq A/B$  且  $\ln L \leq B - D$ 。仅用初等界  $L! \leq L^L$  即可得到

$$\ln \frac{n}{eL!} \geq A - 1 - L \ln L \geq \frac{AD}{B} - 1.$$

由于  $(AD/B)/B \rightarrow \infty$ , 对充分大的  $n$  有  $n/(eL!) \geq 3 \ln n$ 。因此式 (53) 至多为  $n^{-3}$ 。应用定理 6.3 得到

$$\mathbb{P}\left(\max_i X_i < L\right) \leq e\sqrt{n} n^{-3} \leq \frac{1}{n}.$$

**定理 6.4** (典型最大负载)。将  $n$  个球独立、均匀投入  $n$  个箱子时,

$$\max_i X_i = \Theta\left(\frac{\ln n}{\ln \ln n}\right) \quad \text{以概率 } 1 - O(1/n). \quad (54)$$

下界来自刚刚证明的泊松化论证，上界来自定理 2.2。另一个直接推论是：当  $m \geq n \ln n$  时，确定性不等式  $\max_i X_i \geq m/n$  与上界结合，给出以高概率成立的  $\Theta(m/n)$ 。

## 6.5 优惠券收集：期望、尾界与极限分布

**期望与初等上尾界。** 不断投入新球，令  $T$  为所有  $n$  个箱子首次都非空的时刻。等价地，每次购买独立、均匀地获得  $n$  种优惠券中的一种， $T$  是集齐所有种类所需的购买次数。令  $T_i$  表示已有  $i-1$  个非空箱子后，到首次命中新箱子为止的投球次数，包含最后一次成功。

**命题 6.5。** 收集优惠券所需时间满足

$$T = \sum_{i=1}^n T_i, \quad T_i \sim \text{Geom}\left(\frac{n-i+1}{n}\right), \quad \mathbb{E}T = nH_n = n \ln n + O(n), \quad (55)$$

其中  $H_n = \sum_{i=1}^n 1/i$ ，几何分布计数直到首次成功为止的试验次数。

**证明。** 当恰有  $i-1$  个非空箱子时，每次新投球命中新箱子的概率为  $p_i = (n-i+1)/n$ ，平均等待时间为  $1/p_i$ （见附录 A）。对期望求和可得  $\mathbb{E}T = \sum_i n/(n-i+1) = nH_n$ 。用  $1/x$  的积分比较即可得到  $\ln n \leq H_n \leq 1 + \ln n$ 。期望的线性性本身不需要各等待时间独立。

□

期望并不能保证某一次运行会在该时刻附近结束。下面的上尾界有一个初等证明，无须分析等待时间之和。

**命题 6.6 (优惠券收集时间的上尾)。** 对任意整数  $m \geq 0$  及  $n \geq 2$ ,

$$\mathbb{P}(T > m) \leq n \left(1 - \frac{1}{n}\right)^m \leq ne^{-m/n}. \quad (56)$$

因此，对  $c > 0$  有  $\mathbb{P}(T > \lceil n \ln n + cn \rceil) \leq e^{-c}$ 。

**证明。** 固定箱子在  $m$  次投球中始终未被命中的概率为  $(1 - 1/n)^m$ 。事件  $T > m$  等价于至少有一个箱子始终为空。对  $n$  个箱子使用并合界，再利用  $1 - x \leq e^{-x}$  即可。

□

事件  $T > m$  恰好表示“投完  $m$  个球后尚未覆盖全部箱子”。期望给出平均完成时间，上尾界给出高置信度的停止时间保证；两者本身都不能确定  $n \ln n$  附近  $O(n)$  窗口内的波动分布。

**精确波动窗口。** 固定  $c \in \mathbb{R}$ ，令  $m = \lfloor n \ln n + cn \rfloor$ ，并取充分大的  $n$  使  $m > 0$ 。对于独立变量  $Y_i \sim \text{Pois}(m/n)$ ，全部箱子非空的概率恰为

$$\mathbb{P}(Y_i > 0 \forall i) = \left(1 - e^{-m/n}\right)^n = \left(1 - \frac{e^{-c+o(1)}}{n}\right)^n \rightarrow e^{-e^{-c}}. \quad (57)$$

这只是泊松化模型的计算。为将极限转移到恰好  $m$  个球，令  $q_n(s)$  表示  $s$  次均匀投球覆盖全部  $n$  个箱子的概率。用同一条无限投球序列耦合全部实验，可知  $q_n(s)$  关于整数  $s$  单调不减。精确条件化给出

$$\mathbb{P}(Y_i > 0 \forall i) = \mathbb{E}[q_n(S)], \quad S = \sum_i Y_i \sim \text{Pois}(m). \quad (58)$$

取整数窗口半径  $a = \lceil \sqrt{2m \ln m} \rceil$ 。由于  $m \sim n \ln n$ ，故  $a = o(n)$ ，且对充分大的  $n$  有  $a < m$ 。由式 (50)，

$$\eta_n := \mathbb{P}(|S - m| > a) \leq e^{-a^2/(3m)} + e^{-a^2/(2m)} \leq m^{-2/3} + m^{-1} = o(1). \quad (59)$$

**引理 6.7** (覆盖概率随投球次数缓慢变化)。对整数  $s, r \geq 0$ ,

$$0 \leq q_n(s+r) - q_n(s) \leq r/n. \quad (60)$$

*证明*. 用相同的前  $s$  次投球耦合两个实验。若此时已经覆盖全部箱子，两者的覆盖指示变量相同。否则，选定一个特定空箱，例如编号最小的空箱。给定前  $s$  次投球后，它就是一个固定箱子。若接下来  $r$  次投球使覆盖变得完整，这个特定箱子必定被命中。由并合界，该条件概率至多为  $r/n$ 。再取平均即可。  $\square$

选定一个特定空箱至关重要。当空箱很多时，新球命中“某个空箱”的概率未必至多为  $r/n$ ；但完成全部覆盖一定要求命中我们选定的那个箱子。

在  $|S - m| \leq a$  上，单调性使  $q_n(S)$  与  $q_n(m)$  都位于  $q_n(m-a)$  和  $q_n(m+a)$  之间。由引理 6.7，该区间宽度至多为  $2a/n$ 。窗口外，两者都在  $[0, 1]$  中。因此

$$|\mathbb{E}q_n(S) - q_n(m)| \leq \frac{2a}{n} + \eta_n = o(1). \quad (61)$$

结合式 (57)–(61)，得到所需极限。

**定理 6.8** (优惠券收集的极限分布)。对任意固定实数  $c$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\frac{T - n \ln n}{n} \leq c\right) = e^{-e^{-c}}. \quad (62)$$

波动窗口位于  $n \ln n$  附近，宽度量级为  $n$ 。其上尾形式为  $1 - e^{-e^{-c}}$ 。由引理 6.7，在阈值处切换严格与非严格不等式，覆盖概率至多改变  $1/n$ ，不影响极限。

### 泊松化的两种不同用途

对最大负载，独立性使坏事件概率极小，损失  $e\sqrt{m}$  的一般比较界仍能保留结论。对优惠券收集的极限，则必须精确保留常数量级的概率；总数集中、单调性及缓慢变化的耦合共同保证，去掉条件只产生  $o(1)$  误差。

## 7 如何选择集中性论证

选择集中不等式，首先要看随机量的表示方式，而不是定理的名称。下表总结本讲已经建立的几类论证。

| 随机结构      | 需要控制的量                        | 方法                 |
|-----------|-------------------------------|--------------------|
| 独立指示变量之和  | 总均值与完整的乘性指数；相对偏差很大时，不宜过早简化指数。 | Chernoff（第 2 节）    |
| 独立有界变量之和  | 取值区间宽度的平方和。                   | Hoeffding（第 2.4 节） |
| 独立随机选择的函数 | 每个坐标改变时的影响；必要时将输入组织成合适的独立块。   | McDiarmid（第 3 节）   |
| 条件均值为零的增量 | 条件取值区间的宽度，或增量绝对值的上界。          | 鞅集中（第 5 节）         |
| 总数固定的负载   | 独立泊松负载，以及转回固定总数模型时所需的比较误差。    | 泊松化（第 6 节）         |

**使偏差尺度与问题相匹配。** 加性窗口未必相对于均值很小。对于罕见模式或期望空箱数很小的情形，有效的有界差分估计仍可能过于粗糙。另一方面，围绕  $\mathbb{E}\chi(G)$  或  $\mathbb{E}\|\sum_i \varepsilon_i u_i\|$  的集中界，并不一定确定该期望的位置。控制波动与估计中心是两个不同任务。

**分别核算每一项误差。** 单侧尾界没有额外的因子 2；双侧结论需要把两侧概率相加。同时控制多个对象时，还需要对这些对象使用并合界。在 Bloom 过滤器中，占用波动与例外概率是不同的误差项。在去泊松化中，精确的条件分布恒等式本身，并不能说明条件概率可以直接换成无条件概率。

### 共同的证明步骤

确定坏事件，选择适当的随机性表示，再控制指数矩或独立乘积。转回原始统计量时，保留每个结论的假设和常数。最有效的改进常常是更好的随机表示，而不是更一般的不等式。

## 8 习题

以下习题按照正文的讲解顺序组织：先应用集中界，再检验其鞅证明及概率假设。

### 独立求和与成功概率放大

- 调整最大负载的失败概率。** 当  $m = n$  且  $\alpha > 0$  固定时，证明任意固定  $K > \alpha + 1$  都满足：对充分大的  $n$ ， $\mathbb{P}(M \geq \lceil K \ln n / \ln \ln n \rceil) \leq n^{-\alpha}$ 。说明  $\alpha + 1$  中额外的 1 从何而来。  
提示：对  $n(e/L)^L$  取对数。
- 多数表决与中位数是不同的估计器。** 某算法正确的概率至少为  $2/3$ 。按本讲的界，将错误概率降至  $10^{-6}$  需要多少次独立运行？构造一组数值输出，说明即使严格多数位于目标区间内，其算术平均仍可能不在区间内。

提示：取  $p = 1/6$ ，使用奇数次重复，并允许一个任意大的离群值。

### 有界差分与 Bloom 过滤器

3. **模式匹配的精确坐标预算。** 对每个位置  $i$ ，计算包含它的长度为  $k$  的窗口数  $r_i$ 。用  $\sum_i r_i^2$  替换式 (20) 中的分母预算  $nk^2$ 。讨论为何在模式极罕见时，这个改进后的加性估计仍可能很弱。

提示：窗口起点属于  $[1, n - k + 1] \cap [i - k + 1, i]$ 。

4. **边坐标与顶点块。** 以每条边为一个坐标应用 McDiarmid 不等式，再把所得色数波动尺度与顶点分块界 (23) 比较。验证：两图若仅在与某个顶点相接的边上不同，其色数之差至多为一。

提示：比较染色前，先在两图中都删除该顶点。

5. **有限规模的 Bloom 过滤器计算。** 取  $m = 10n$ 、 $k = 7$ 。写出式 (27) 的有限规模上界，不把  $(1 - 1/m)^{kn}$  替换为指数。分别指出占用波动、指数近似误差及例外概率。

提示：前两种误差不同，只有最后一种是  $n^{-100}$ 。

### 鞅与有界差分的证明

6. **条件取值区间的宽度。** 在定理 5.4 中，解释为何区间端点可以依赖于过去，却必须在下一增量发生前已知。对  $B \sim \text{Bernoulli}(p)$  ( $0 < p < 1$ )，计算  $B - p$  的精确取值范围，将其宽度与最大绝对值的两倍比较。利用区间宽度界，为独立有界变量之和重新推导 Hoeffding 指数。
7. **为什么不能省略独立性。** 令  $B$  为公平比特， $X_1 = \dots = X_n = B$ 。函数  $f(x) = \sum_i x_i$  关于每个坐标都是 1-Lipschitz 的。计算第一个非平凡 Doob 增量，以及  $\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq n/2)$ 。准确指出独立输入证明中的哪一步失效。

提示：揭示一个坐标就揭示了全部坐标。

### 泊松化与占用阈值

8. **高置信度的收集时间。** 给定  $0 < \delta < 1$ ，求一个明确的整数投球次数，使全部覆盖的概率至少为  $1 - \delta$ 。将其量级与  $\mathbb{E}T$  比较，并解释为何这个单侧保证不能确定  $(T - n \ln n)/n$  的极限分布。
9. **常数稍弱但更简单的负载下界。** 取  $L = \lfloor \ln n / (2 \ln \ln n) \rfloor$ ，利用  $L! \leq L^L$  证明：对充分大的  $n$ ，泊松化模型中所有负载均小于  $L$  的概率至多为  $\exp(-\sqrt{n}/e)$ 。再把这一结论转移到固定总数实验。
10. **去泊松化究竟需要什么。** 证明：只要整数窗口  $a_n = o(n)$  满足  $\mathbb{P}(|S - m| > a_n) = o(1)$ ，就足以推出优惠券收集的极限。解释为何一般的  $e\sqrt{m}$  比较界不能单独恢复该极限，以及为什么缓慢变化的证明中不能把“一个特定空箱”替换为“任意一个空箱”。

## A 初等概率与分析估计

本附录补全正文论证中使用的初等估计。

### A.1 几何等待时间与阶乘不等式

若  $G$  统计成功概率为  $p > 0$  的独立 Bernoulli 试验直到首次成功所需的次数，则对  $r \geq 1$ ， $\mathbb{P}(G \geq r) = (1-p)^{r-1}$ 。由恒等式  $G = \sum_{r \geq 1} \mathbf{1}\{G \geq r\}$ ，对非负项逐项取期望，得到

$$\mathbb{E}G = \sum_{r \geq 1} \mathbb{P}(G \geq r) = \sum_{r \geq 1} (1-p)^{r-1} = 1/p.$$

对任意整数  $r \geq 1$ ，本讲需要的阶乘估计为

$$(r/e)^r \leq r! \leq e\sqrt{r}(r/e)^r. \quad (63)$$

对下界，由  $\ln x$  的单调性， $\ln r! \geq \int_1^r \ln x dx = r \ln r - r + 1 \geq r \ln r - r$ 。对上界，由  $\ln x$  的凹性，其图像不低于各区间上的弦，故

$$\int_1^r \ln x dx \geq \sum_{j=1}^{r-1} \frac{\ln j + \ln(j+1)}{2} = \ln r! - \frac{1}{2} \ln r.$$

整理并取指数即可。这些初等不等式已足以证明损失  $e\sqrt{m}$  的比较界，不需要完整的 Stirling 渐近公式。另外， $r! \leq r^r$  常常是更简便的上界。

### A.2 简化 Chernoff 指数

定义  $h_+(\delta) = (1+\delta) \ln(1+\delta) - \delta$  及  $h_-(\delta) = \delta + (1-\delta) \ln(1-\delta)$ 。精确的上尾与下尾界分别为  $e^{-\mu h_+(\delta)}$  和  $e^{-\mu h_-(\delta)}$ 。

对  $x \geq 0$ ，有  $\ln(1+x) \geq 2x/(2+x)$ ：两者之差在零点为零，导数为  $x^2/((1+x)(2+x)^2) \geq 0$ 。特别地，对  $0 \leq x \leq 1$ ， $\ln(1+x) \geq 2x/3$ 。由  $h_+(0) = 0$  及  $h'_+(x) = \ln(1+x)$ ，

$$h_+(\delta) \geq \int_0^\delta \frac{2x}{3} dx = \frac{\delta^2}{3} \quad (0 \leq \delta \leq 1).$$

同理，对  $0 \leq x < 1$ ， $h'_-(x) = -\ln(1-x) \geq x$ ，所以

$$h_-(\delta) \geq \int_0^\delta x dx = \frac{\delta^2}{2} \quad (0 \leq \delta < 1).$$

这就是 Bernoulli 与泊松小偏差界中共同使用的不等式。端点  $\delta = 1$  可用连续性或直接计算零次成功的概率处理。

### A.3 Bloom 过滤器松弛最优值的一元计算

为补全第 4 节的优化，令  $R(u) = \ln u \ln(1-u)$ ，其中  $0 < u < 1$ 。有

$$R'(u) = \frac{(1-u) \ln(1-u) - u \ln u}{u(1-u)}.$$

分子记为  $H(u)$ , 则  $H(0+) = H(1/2) = 0$ , 且在  $(0, 1/2)$  上  $H''(u) = (2u - 1)/(u(1 - u)) < 0$ . 它严格凹, 因此在两个零端点之间为正。故  $R$  在  $1/2$  之前递增; 由对称性  $R(u) = R(1 - u)$ , 之后递减。其唯一最大值为  $(\ln 2)^2$ 。再由  $\ln F_c(k) = -cR(e^{-k/c})$ , 松弛最优参数为  $k = c \ln 2$ 。

## B 负相依性与指数矩界

独立求和的证明使用指数矩的乘积法则。其实等式并非必需: 只要联合指数矩不超过边缘指数矩的乘积, 就足以完成论证。负相依性正是保证这一不等式的一种联合条件。

**定义 B.1 (负相依性)。** 若变量族  $(X_i)_{i \in I}$  对任意不相交子集  $A, B \subseteq I$  以及期望存在的逐坐标单调不减函数  $f, g$  都满足

$$\mathbb{E}[f(X_A)g(X_B)] \leq \mathbb{E}f(X_A) \mathbb{E}g(X_B), \quad (64)$$

则称其为负相依的。这约束的是不同坐标组上单调函数的乘积期望。

对负相依变量族, 只要所需指数矩有限, 就有

$$\mathbb{E} \prod_{i=1}^N e^{\lambda X_i} \leq \prod_{i=1}^N \mathbb{E} e^{\lambda X_i} \quad (\lambda \in \mathbb{R}). \quad (65)$$

当两个函数都单调不增时, 也有相同的协方差不等式: 对  $-f$  与  $-g$  应用定义即可。若  $f_i$  为单坐标上的非负函数, 且都单调不减, 则任一前缀组的乘积仍单调不减。对因子数归纳可得

$$\mathbb{E} \prod_i f_i(X_i) \leq \prod_i \mathbb{E} f_i(X_i).$$

当所有因子都单调不增时, 同样可以归纳。取  $f_i(x) = e^{\lambda x}$  即证式 (65):  $\lambda > 0$  时各因子都递增,  $\lambda < 0$  时各因子都递减。对无界因子, 先使用非负截断, 再取极限。

一个补充检验可以区分负相依性与两两独立: 从四个比特向量 001, 010, 100, 111 中均匀选择。每对坐标都独立, 且每个比特公平, 但  $\mathbb{P}(X_1 = X_2 = X_3 = 1) = 1/4 > 1/8 = \prod_i \mathbb{P}(X_i = 1)$ 。因此仅有两两独立时, 连三个单坐标递增函数的乘积都未必满足乘积不等式。

**独立求和集中界的推广。** 在第 2 节的各项证明中, 用式 (65) 替换独立乘积等式即可。因此, 乘性 Chernoff 界适用于负相依的 Bernoulli 变量; Hoeffding 不等式适用于负相依的有界变量, 仍使用相同的区间宽度; 方差因子分别为  $v_i$  的中心化负相依次高斯变量之和, 具有方差因子  $\sum_i v_i$ 。逐坐标中心化保留负相依性, 因为平移自变量不改变单调性。

这些推广针对求和及其乘积不等式, 并不取消 McDiarmid 定理的独立坐标假设。仅有两两协方差非正, 或仅有两两独立, 都不能建立联合乘积不等式。若没有为实际使用的变量族证明负相依性, 算法应用中的独立性假设仍不可省略。

## 文献说明

Hoeffding [1]研究有界随机变量之和的指数尾界；McDiarmid [2]系统介绍独立输入函数的有界差分方法；Azuma [3]研究有界鞅差分的集中性。条件取值区间的论证说明，这些不等式可以放在同一个指数矩框架下理解。

Bloom [4]提出了近似成员查询的数据结构。Bose 等 [5]分析其有限规模的假阳性概率，以及条件独立与无条件独立之间的区别。Shamir 与 Spencer [6]用鞅方法研究随机图色数的集中性。

## 参考文献

- [1] Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi:10.1080/01621459.1963.10500830.
- [2] Colin McDiarmid. On the Method of Bounded Differences. In J. Siemons (ed.), *Surveys in Combinatorics*, 1989, pages 148–188. Cambridge University Press, 1989. doi:10.1017/CBO9781107359949.008.
- [3] Kazuoki Azuma. Weighted Sums of Certain Dependent Random Variables. *Tohoku Mathematical Journal*, 19(3):357–367, 1967. doi:10.2748/tmj/1178243286.
- [4] Burton H. Bloom. Space/Time Trade-offs in Hash Coding with Allowable Errors. *Communications of the ACM*, 13(7):422–426, 1970. doi:10.1145/362686.362692.
- [5] Prosenjit Bose, Hua Guo, Evangelos Kranakis, Anil Maheshwari, Pat Morin, Jason Morrison, Michiel Smid, and Yihui Tang. On the False-Positive Rate of Bloom Filters. *Information Processing Letters*, 108(4):210–213, 2008. doi:10.1016/j.ipl.2008.05.018.
- [6] Eli Shamir and Joel Spencer. Sharp Concentration of the Chromatic Number on Random Graphs  $G_{n,p}$ . *Combinatorica*, 7(1):121–129, 1987. doi:10.1007/BF02579208.