第二章 独立随机变量和的集中

Roman Vershynin

本章向读者介绍集中不等式这一丰富的课题. 在 2.1 节说明为什么需要学习本章内容之后, 我们将在后面几节证明一些基本的集中不等式: 2.2 节和 2.6 节证明 Hoeffding (霍夫丁) 不等式, 2.3 节证明 Chernoff (切尔诺夫) 不等式, 2.8 节证明 Bernstein (伯恩斯坦) 不等式. 本章的另一个目标介绍两类重要的分布: 2.5 节中的次高斯分布和 2.7 节中的次指数分布. 这些类别形成了一个自然的" 栖息地", 在其中, 许多高维概率的结果及其应用得到了发展. 我们也将在 2.2 节和 2.4 节中分别给出集中不等式在随机算法中的两个快速应用. 本章的内容在后面还有更多的应用.

2.1 集中不等式的由来

集中不等式量化了随机变量 X 如何偏离它的均值 μ. 它们通常给出 $X-\mu$ 尾分布的双侧边界形式. 例如:

$$\mathbb{P}\left\{|X-\mu|>t\right\} \leqslant \varepsilon \;\; (\varepsilon<1) \tag{2.1}$$

最简单的集中不等式是切比雪夫不等式 (Coronary 1.3). 它具有一般性, 但往往不够有力. 让我们用二项分布的例子来说明这一点.

Question 2.1. 抛一枚硬币 N 次, 问至少得到 $\frac{3N}{4}$ 次正面朝上的概率是多少?

设 S_N 为正面朝上的次数, 那么

$$\mathbb{E}S_N=\frac{N}{2},\quad Var\left(S_N\right)=\frac{N}{4} \tag{2.2}$$

切比雪夫不等式界定的至少得到 $\frac{3N}{4}$ 次正面朝上的概率为

$$\mathbb{P}\left\{S_N\geqslant\frac{3}{4}N\right\}\leqslant\mathbb{P}\left\{\left|S_N-\frac{N}{2}\geqslant\frac{N}{4}\right|\right\}\leqslant\frac{4}{N} \tag{2.3}$$

因此其概率关于 N 至少线性地收敛于零.

这是正确的递减速度, 还是我们应该期待更快的递减速度? 让我们用中心极限定理来处理同一问题. 为了做到了这一点, 我们把 S_N 表示为独立随机变量的和:

$$S_N=\sum_{i=1}^{N}X_i \tag{2.4}$$

其中, X_i 是相互独立的, 服从参数为 $\frac{1}{2}$ 的伯努利分布的随机变量, $\mathbb{P}\left\{X_i=0\right\}=\mathbb{P}\left\{X_i=1\right\}=\frac{1}{2}$ (这里 X_iI 表示第 i 次抛硬币出现的结果, $X_i=1$ 表示正面朝上). 棣莫弗–拉普拉斯中心极限定理指出, 正面朝上的次数的标准化数分布

$$Z_N=\frac{S_N-\frac{N}{2}}{\sqrt{\frac{N}{4}}} \tag{2.5}$$

依分布收敛于标准正态分布 $N\left(0,1\right)$, 因此, 我们可以推断出, 当 N 是一个很大的数时, 我们有

$$\mathbb{P}\left\{S_N\geqslant\frac{3}{4}N\right\}=\mathbb{P}\left\{Z_N\geqslant\sqrt{\frac{1}{4}N}\right\}\approx\mathbb{P}\left\{g\geqslant\sqrt{\frac{1}{4}N}\right\} \tag{2.6}$$

其中, $g\sim N\left(0,1\right)$. 为了明白这个量关于 N 是如何递减的, 我们现在引入正态分布尾分布的一个很好的界.

Claim 2.1 (正态分布的尾分布)**.** 设 $g \sim N(0,1)$. 则对任意 $t > 0$, 都有

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \leqslant \mathbb{P}\{g \geqslant t\} \leqslant \frac{1}{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \tag{2.7}$$

特别地, 如果 $t \geq 1$, 则尾分布的上界为密度函数

$$\mathbb{P}\{g \geqslant t\} \leqslant \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \tag{2.8}$$

证明. 为了获得尾分布的一个上界, 先考虑

$$\mathbb{P}\{g \geqslant t\} \leqslant \frac{1}{\sqrt{2\pi}} \int_t^{\infty} e^{-\frac{x^2}{2}} dx \tag{2.9}$$

作变量替换 $x = t + y$, 得到

$$\mathbb{P}\{g \geqslant t\} = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{t^2}{2}} e^{-ty} \underbrace{e^{-\frac{y^2}{2}}}_{\leqslant 1 \; if \; y \geqslant 0} dy \leqslant \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \int_0^{\infty} e^{-ty} dy \tag{2.10}$$

因为最后一个积分等于 $\frac{1}{t}$, 所以得到尾分布的上界.

同时, 由于 $1 - 3x^{-4} \leqslant 1$, 下界来自下面的恒等式

$$\int_t^{\infty} (1 - 3x^{-4}) e^{-\frac{x^2}{2}} dx = \left(\frac{1}{t} - \frac{1}{t^3}\right) e^{-\frac{t^2}{2}} \tag{2.11}$$

故得证.

□

先回到 2.3 式, 我们可以看到至少有 $\frac{3}{4}N$ 次正面朝上的概率小于

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{N}{8}} \tag{2.12}$$

这个量关于 N 呈指数快速地递减到零, 这比切比雪夫不等式得出的 2.3 中的线性衰减要好的多.

遗憾的是 2.12 没有严格遵循中心极限定理. 虽然 2.6 中的正态密度函数是近似有效的, 但近似误差

不可忽略，并且误差递减得太慢，甚至比 N 的线性递减还要慢，这可以从下面的中心极限定理的精确定量版本中看到．

Theorem 2.1 (Berry-Esseen 中心极限定理)**.** *在 Lindeberg-Lévy 定理中，对任意 N 和任意 $t \in \mathbb{R}$，有*

$$|\mathbb{P}\{Z_N \geqslant t\} - \mathbb{P}\{g \geqslant t\}| \leqslant \frac{\rho}{\sqrt{N}} \tag{2.13}$$

其中 $\rho = \frac{\mathbb{E}|X_1-\mu|}{\sigma^3}$，且 $g \sim N(0,1)$．

Remark 1 (Lindeberg-Lévy 中心极限定理)**.** 设随机变量 $X_1, X_2, \ldots$ 是独立同分布的序列，其均值为 μ，方差为 σ^2．考虑随机变量之和 $S_N = X_1 + \cdots + X_N$，并对其进行标准化以获得具有零均值和单位方差的随机变量，即

$$Z_N : \frac{S_N - \mathbb{E}S_N}{\sqrt{Var(S_N)}} = \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^{N} (X_i - \mu) \tag{2.14}$$

则当 $N \to \infty$ 时，有

$$Z_N \overset{\text{依分布}}{\to} N(0,1) \tag{2.15}$$

By Thm 2.1, 2.6 中的近似误差的阶为 $\frac{1}{\sqrt{N}}$，这不满足呈指数递减的结果 2.12.

我们能使用中心极限定理改进所涉及的近似误差吗？一般来说，并不能．如果 N 是偶数的话，那么恰好得到 $\dfrac{N}{2}$ 从正面向上的概率是

$$\mathbb{P}\left\{S_N = \frac{N}{2}\right\} = 2^{-N}\binom{N}{\frac{N}{2}} \sim \frac{1}{\sqrt{N}} \tag{2.16}$$

最后的估计可以用 Stirling 公式得到．

斯特林公式:

(https://en.wikipedia.org/wiki/Stirling%27s_approximation)

$$\sqrt{2\pi}\, n^{n+\frac{1}{2}}\, e^{-n} \leqslant n! \leqslant e\, n^{n+\frac{1}{2}}\, e^{-n} \tag{2.17}$$

因此 $\mathbb{P}\{g=0\} = 0$，所以这时近似误差的阶数必须是 $\frac{1}{\sqrt{N}}$.

让我们总结前面的结论．中心极限定理通过正态分布来逼近独立随机变量之和 $S_N = X\_1 + \cdots + X_N$．正态分布很好，因为它的尾分布很轻，呈指数递减．但与之相应的是，中心极限定理的近似误差递减得太慢，甚至比线性得递减还要慢，这个巨大得误差是证明具有指数递减尾分布得随机变量 S_N

集中特性得一个障碍.

为了解决这个问题, 我们将探讨替代得, 直接得, 绕过中心极限定理得集中方法。

Homework 2.1 (截断正态分布)**.** 设 $g \sim N(0,1)$, 求证: 对于所有的 $t \geq 1$, 都有

$$\mathbb{E}g^2 1_{\{g>t\}} = \frac{t}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} + \mathbb{P}\{g > t\} \leqslant \left(t + \frac{1}{t}\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \tag{2.18}$$

2.2 Hoeffding 不等式

2.3 Chernoff 不等式

2.4 应用: 随机图的度数

2.5 次高斯分布

2.6 广义 Hoeffding 不等式和辛钦不等式

2.7 次指数分布

2.8 Bernstein 不等式

2.9 后注

2.10 参考文献

[1] R. Durrett , Probablity: Theory and Examples, *Cambridge Series in Statistical and Probabilistic Mathematics*, 2010, Vol. 31.

[2] P. Billingsley , Probability and Measure, *Wiley Series in Probability and Mathematical Statistics*, 1995.