概率论、费曼图以及随机矩阵理论（一）：中心极限定理与随机游走

陈童

中心极限定理可以说是概率论中最为重要的定理之一。它解决的是N个独立随机变量的和在N很大时的概率分布问题。中心极限定理说，不管什么随机变量，在满足一些合理的条件以后这样的概率分布一定会趋向于一个高斯分布。这样一个普适性的结果是非常令人吃惊的，它首次揭示了系统在大N极限下可能会展现出普遍性的规律。这在整个数学和物理学上都影响深远。我们这里主要讨论这个重要结果的数学证明。为简单起见，这里我们将仅仅处理N个独立同分布随机变量的情形。

1 高尔顿钉板实验

英国生物统计学家高尔顿是达尔文的表弟，高尔顿钉板实验是他设计来直观演示中心极限定理的。

高尔顿在木板上钉了很多钉子，这些钉子分成了N层，每一层的相邻两颗钉子之间间距相等，上一层钉子的水平位置都恰好落在下一层的相邻两颗钉子正中间。在实验时，从上方放入一个小圆球，小圆球的直径小于任意两颗钉子之间的几何距离，但是大于相邻两层钉子之间的水平距离。因此，小圆球在下落的过程中总会碰到钉子，它碰到钉子以后可以往左边滚下来，也可以往右边滚下来，概

率均为1/2。然后再碰到下一层的钉子，又以1/2的概率往左滚或往右滚，如此不断进行下去，直到经过N层钉子以后，小圆球最终会落到下方的格子中。

在第i层，小圆球往左滚还是往右滚是随机的，往右滚则它的水平位移增加1，往左滚则减少1，我们可以定义小圆球在第i层的水平位移为随机变量X_i, 小圆球往右滚$X_i = 1$，往左滚则$X_i = -1$。很显然，经过N层钉子以后，小圆球最终会落在哪个方格子中由总水平位移$X = X_1 + X_2 + ... + X_N$决定。$X$当然也是一个随机变量，它的分布情况可以通过将实验重复很多很多次，然后统计下方各格子中小圆球的数目，由这个数目的分布情况反映出来。

网上有这个实验的视频，从视频中可以清楚地看到，最终小圆球的数目分布趋向于一个正态分布(当N很大时)，因此也就是说，随机变量X最终趋向于一个正态分布。这就是中心极限定理的一个最直观例子。下面我们开始讨论中心极限定理的证明。

2 随机变量的特征

在证明中心极限定理之前，我们先简单地讨论一个有用的数学概念，即一个随机变量的特征。假设有一个随机变量X，其概率分布为$p(x)$，则X的特征$\widehat{p}(k)$定义为

$$\widehat{p}(k) = \langle e^{ikX} \rangle = \int dx e^{ikx} p(x). \tag{1}$$

很显然，$\widehat{p}(0) = 1$。更一般的，$\widehat{p}(k)$在0处关于ik泰勒展开的n阶项系

数就是随机变量X的n极矩M_n，$M_n = \langle X^n \rangle = \int dx x^n p(x)$，即

$$\widehat{p}(k) = 1 + \sum_{n=1}^{+\infty} \frac{1}{n!}(ik)^n M_n. \tag{2}$$

一个更有用的概念是所谓的累积量(cumulants)，我们这里也称之为集团n极矩K_n，它由下式给出

$$\widehat{p}(k) = \exp\left(\sum_{n=1}^{+\infty} \frac{1}{n!}(ik)^n K_n\right), \tag{3}$$

人们也常常将K_n记为$K_n = \langle X^n \rangle_c$。

由n极矩和集团n极矩的两个定义式子可以得出

$$\begin{aligned} M_1 &= K_1, \\ M_2 &= K_2 + K_1^2, \\ M_3 &= K_3 + 3K_1K_2 + K_1^3, \\ M_4 &= K_4 + 4K_3K_1 + 3K_2^2 + 6K_2K_1^2 + K_1^4 .... \end{aligned} \tag{4}$$

实际上，如果我们将K_m表示成一个m个点的集团，然后将任意n个点完全分解成一些这样的集团，那么M_n就可以看成是n个点所有可能集团分解的和，如图(1)所示。 很明显，(1)中的图形正好对应等式(4)。其实，$K_1 = M_1 = \langle X \rangle$就是通常所说的平均值，我们也记为$\mu$，$K_2 = M_2 - M_1^2 = \langle X^2 \rangle - \langle X \rangle^2$就是我们通常所说的方差$\sigma^2$。因此，根据累积量，或者说集团$n$极矩的定义，我们有

$$\widehat{p}(k) = \exp\left(ik\mu - \frac{1}{2}k^2\sigma^2 + ...\right). \tag{5}$$

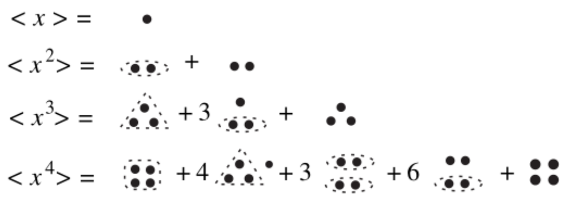

Figure 1: 图中每一个未被圈起来的单独黑点自身是一个单点集团，每一个被虚线圈起来的多点是一个多点集团。

3　中心极限定理的证明与随机游走

假设有N个独立的随机变量$X_i, i = 1, 2, ..., N$，每一个的概率分布都是同一个函数$p(x_i)$, 其平均值为μ, 方差为σ^2。这里我们以小写的x_i来表示随机变量X_i的取值。则随机变量$X = \sum_i X_i$的概率分布$q(x)$可以表示成，

$$q(x) = \int [\prod_i dx_i] \delta(x - \sum_i x_i) p(x_1) p(x_2) ... p(x_N). \tag{6}$$

利用$\delta(x) = \int \frac{dk}{2\pi} e^{-ikx}$, 我们就有

$$\begin{aligned} q(x) &= \int \frac{dk}{2\pi} e^{-ikx} \int [\prod_i dx_i] e^{ik \sum_i x_i} p(x_1) p(x_2) ... p(x_N) && (7) \\ &= \int \frac{dk}{2\pi} e^{-ikx} [\int dx' e^{ikx'} p(x')]^N && (8) \\ &= \int \frac{dk}{2\pi} e^{-ikx} \widehat{p}^N(k), && (9) \end{aligned}$$

式中$\widehat{p}(k) = \int dx' e^{ikx'} p(x')$。利用(5)式我们容易将上面的结果重写成，

$$q(x) = \int \frac{dk}{2\pi} e^{-ik(x-N\mu)-N\frac{1}{2}\sigma^2 k^2 + NO(k^3)}. \tag{10}$$

假设我们将Nk看成一个新变量，则容易看出在$N \to \infty$的时候式中的$O(k^3)$可以忽略。由此就可以得到

$$q(x) \to \frac{1}{\sqrt{2\pi N}\sigma} \exp(-\frac{1}{2}\frac{(x-N\mu)^2}{N\sigma^2}). \tag{11}$$

这就证明了中心极限定理。从最终结果我们可以看到随机变量X的平均值为$N\mu$, 而方差为$N\sigma^2$。

下面我们给出中心极限定理的一个简单解释。假设我们考虑一个醉汉在x轴上随机游走，我们将上文的随机变量X_i看成是醉汉第i步这一步所产生的位移，由于这是一个醉汉，因此这个位移是随机变量。μ就是醉汉平均每一步前进的距离。那么随机变量X就是醉汉在N步中的总位移，中心极限定理告诉我们，在N很大时，这个总位移趋向于一个高斯分布，这个高斯分布的方差是$N\sigma^2$。

关于方差的这个结果其实可以用下面的方法容易地得到，下面为了简单起见我们不妨假设醉汉醉得很厉害，以致于$\mu = 0$。我们记醉汉在i步中的总位移为$X(i)$, 显然

$$X^2(i+1) = (X(i) + X_{i+1})^2 = X^2(i) + X_{i+1}^2 + 2X(i)X_{i+1}. \tag{12}$$

然后我们将上面这个方程两边取期望值，由于醉汉是随机游走，因此X_{i+1}与$X(i)$统计无关，因此$\langle X(i)X_{i+1}\rangle = \langle X(i)\rangle\langle X_{i+1}\rangle = 0$，因此

$$\langle X^2(i+1)\rangle = \langle X^2(i)\rangle + \langle X_{i+1}^2\rangle = \langle X^2(i)\rangle + \sigma^2. \tag{13}$$

将这个式子进行递推就可以得到我们需要的结论。当然，从这个推导可以看出，关于方差的这个结果不需要大N极限，这一点也可以从我们在上文中那个更加数学化的证明方法中看出来。实际上，只有最终的分布是一个高斯分布这一点才是依赖于大N极限的。