10 • 概率基础

概率空间

一次随机试验的所有可能结果组成样本空间 Ω。我们需要给某些结果集合赋概率，这些集合叫事件，组成事件族 $\mathcal{F}$。事件族的封闭性保证取反和合并事件后仍能谈它们的概率。概率 P 为事件赋予介于零与一之间的数。

$(\Omega, \mathcal{F}, P)$：$\mathcal{F}$ 对补与可列并封闭；$P(\Omega) = 1$；互斥 A_i 满足 $P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$。任意事件有 $1_{\bigcup_i A_i} \le \sum_i 1_{A_i}$，取期望得并集界。

并集界可以重复计数。即使多个事件会同时发生，右侧仍然是上界，因此不需要独立性。

条件概率与 Bayes 公式

知道 B 已经发生后，只在 B 所包含的结果中讨论 A，于是把交集的原概率除以 $P(B)$。若 B_i 是互不相交且覆盖样本空间的事件，可以按每种情形分别计算 A 发生的概率，再相加。

$P(A|B) = \frac{P(A \cap B)}{P(B)}$，$P(B) > 0$。对正概率划分 (B_i)，

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i),$$

$$P(B_j|A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}.$$

Bayes 公式的分子是某一种情形与观测同时出现的概率，分母是这项观测在所有情形下出现的概率。它用已知观测重新分配各情形的概率。

独立性

两个事件独立，意味着知道其中一个发生不会改变另一个的概率。对多个事件，需要任意一组选出的事件都满足交集概率等于各自概率之积。仅检查每一对不足以推出相互独立。

相互独立 $\Leftrightarrow P\left(\bigcap_{i\in J} A_i\right) = \prod_{i\in J} P(A_i)$ 对每个非空子集 J 成立。离散随机变量独立 $\Leftrightarrow p_{X,Y}(x,y) = p_X(x)p_Y(y)$ 对所有 x, y 成立。

对离散随机变量，逐个检查联合取值的概率即可。任意取值集合的联合概率都可由这些单点概率求和。

随机变量的分布与期望

离散随机变量是把试验结果映射为数值的函数。它的分布只记录每个数值出现的概率。期望是按这些概率对数值求加权平均，对 $g(X)$ 的期望则改为对变换后的数值求平均。

$p_X(x) = P(X = x)$，$\mathbb{E}[g(X)] = \sum_x g(x)p_X(x)$。

$$\mathbb{E}[aX + bY] = \sum_{x,y} (ax + by)p_{X,Y}(x,y) = a\mathbb{E}X + b\mathbb{E}Y.$$

独立时 $\mathbb{E}[XY] = \sum_{x,y} xyp_X(x)p_Y(y) = \mathbb{E}X\mathbb{E}Y$。指示变量 $\mathbb{E}1_A = P(A)$；非负整数 $X = \sum_{k\ge1} 1_{X\ge k}$，故 $\mathbb{E}X = \sum_{k\ge1} P(X \ge k)$。

期望的线性性来自对联合分布拆开求和，因此不要求独立。独立性在乘积的期望中才用于把联合概率写成两个边缘概率的乘积。

方差与协方差

方差以均值为中心，度量偏离量的平方平均。协方差把两个变量各自减去均值再相乘，度量它们偏离的方向是否倾向一致。计算和的方差时，必须处理平方展开产生的交叉项。

$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2] = \mathbb{E}[X^2] - (\mathbb{E}X)^2$，

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y, \quad \mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}X + b^2\,\mathrm{Var}Y + 2ab\,\mathrm{Cov}(X,Y).$$

展开中心化变量的平方即得；独立 $\Rightarrow \mathrm{Cov}(X,Y) = 0$。

具体地，令 $X_0 = X - \mathbb{E}X$、$Y_0 = Y - \mathbb{E}Y$，对 $(aX_0 + bY_0)^2$ 逐项取期望。两条交叉项合并为 $2ab\mathbb{E}[X_0Y_0]$。协方差为零可以消去这一项，但一般不能反过来推出独立。

全期望公式

先把 Y 固定为一个值，按照该条件下 X 的分布求平均。条件期望 $\mathbb{E}[X|Y]$ 随 Y 改变，本身也是随机变量。再按 Y 的分布平均，就能恢复原来的期望。

$\mathbb{E}[X|Y = y] = \sum_x xp_{X|Y}(x|y)$。求和消去条件：

$$\mathbb{E}[\mathbb{E}[X|Y]]=\sum_y p_Y(y)\sum_x xp_{X|Y}(x|y)=\sum_x x\sum_y p_{X,Y}(x,y)=\mathbb{E}X.$$

等式的关键是条件概率乘以条件事件的概率给出联合概率。将 y 求和后，只剩 X 的边缘分布。

概率生成函数

把非负整数型随机变量的概率依次放在幂级数的系数里。读出 z^k 的系数就得到取值 k 的概率。对独立变量之和，取值概率要对所有可能拆分求和，而这恰好是乘两个幂级数时的系数运算。

$X\in\mathbb{N}_0$，$G_X(z)=\sum_{k\ge0}p_X(k)z^k=\mathbb{E}[z^X]$。逐项求导（从左趋近 1）：

$$G_X'(1)=\mathbb{E}X,\quad G_X''(1)=\mathbb{E}[X(X-1)],\quad \mathrm{Var}X=G_X''(1)+G_X'(1)-G_X'(1)^2.$$

$$X,Y\text{ 独立}\Rightarrow G_{X+Y}(z)=\mathbb{E}\left[z^Xz^Y\right]=G_X(z)G_Y(z).$$

N 独立于 iid X_i，$S=\sum_{i=1}^N X_i$：$G_S(z)=\mathbb{E}[G_X(z)^N]=G_N(G_X(z))$。

随机项数情形先固定 $N=n$。给定项数后，和的生成函数是单项生成函数的 n 次幂；再对 N 取期望，得到函数复合。

Bernoulli 与二项分布

一次成功概率为 p 的试验，用 $B=1$ 记成功、$B=0$ 记失败。重复 n 次相互独立的相同试验，成功总数就是二项分布。

$G_B(z)=1-p+pz$，$\mathbb{E}B=p$，$\mathrm{Var}B=p(1-p)$。$X=\sum_{i=1}^n B_i$（iid），$G_X(z)=(1-p+pz)^n$，

$$[z^k]G_X=\binom{n}{k}p^k(1-p)^{n-k},\quad \mathbb{E}X=np,\quad \mathrm{Var}X=np(1-p).$$

恰有 k 次成功时，先选这 k 个位置，再乘每个位置组合的概率。均值和方差也可从独立指示变量逐项相加得到。

几何分布

重复进行相互独立的 Bernoulli 试验，记 X 为第一次成功所在的试验编号。若 $X=k$，事件顺序已固定为前 $k-1$ 次失败、第 k 次成功。这里计数包含成功的那一次。

$P(X=k)=(1-p)^{k-1}p$，$k\ge1$。几何级数给 $G_X(z)=p\frac{z}{1-(1-p)z}$，求导得 $\mathbb{E}X=\frac{1}{p}$、$\mathrm{Var}X=\frac{1-p}{p^2}$。

因为 $P(X\ge k)=(1-p)^{k-1}$，均值也可直接由尾和公式与几何级数求得。

Poisson 分布

Poisson 分布给非负整数计数赋概率，参数 λ 等于平均计数。它可由大量独立、单次成功率很小的试验近似得到，只要成功总次数的均值保持在 λ 附近。

$p_X(k)=e^{-\lambda}\frac{\lambda^k}{k!}$，$G_X(z)=e^{\lambda(z-1)}$，$\mathbb{E}X=\mathrm{Var}X=\lambda$。独立 Poisson 和的参数相加（PGF 相乘）；$\mathrm{Bin}\left(n,\frac{\lambda}{n}\right)$ 的 PGF 趋于 $e^{\lambda(z-1)}$，固定 k 的二项概率亦趋于 $p_X(k)$。

证明极限时固定 k，将二项概率分解为 $\binom{n}{k}\left(\frac{\lambda}{n}\right)^k$ 和 $\left(1-\frac{\lambda}{n}\right)^{n-k}$。两因子分别趋向 $\frac{\lambda^k}{k!}$ 与 $e^{-\lambda}$，给出每个取值的极限概率。

11 · 熵、散度与互信息

记号和零概率约定

本章用 p_x 表示 $X=x$ 的概率，用 $p_{x,y}$ 表示两个变量的联合概率。底为二的对数使信息量的单位为 bit。对零概率项采用连续延拓，条件分布只需在条件有正概率时定义。

本页 $\log=\log_2$；$0\log0=0$；$P(x)>0=Q(x)$ 时 $D(P\,\|\,Q)=\infty$。条件量只对正概率条件加权。

熵与条件熵

观察到概率为 p 的结果，记其信息量为 $-\log p$。概率越小，观察到它所获得的信息越多。熵是这个量的期望。条件熵先在每个给定的 $Y=y$ 下计算 X 的熵，再对实际会看到的 y 平均。

$$H(X)=-\sum_x p_X(x)\log p_X(x),\quad H(X|Y)=\sum_y p_Y(y)H\left(P_{X|Y=y}\right).$$

$$p_{X,Y}=p_Yp_{X|Y}\Rightarrow H(X,Y)=H(Y)+H(X|Y).$$

迭代得 $H(X_1,\dots,X_n)=\sum_i H(X_i|X_1,\dots,X_{i-1})$。若 $Y=f(X)$，则 $H(Y|X)=0$，故 $H(Y)\le H(X)$。

联合熵的推导将 $-\log p_{x,y}$ 拆为 $-\log p_y$ 与 $-\log p_{x|y}$。第一项对 x 求和后是 Y 的熵，第二项按定义是条件熵。

KL 散度

用分布 Q 分配概率时，同一结果的信息量变成 $-\log q_x$。若结果实际从 P 产生，KL 散度就是按 P 平均的额外信息量。它有方向，交换 P, Q 后一般会变。

$D(P\parallel Q)=\sum_{x:p_x>0}p_x\log\left(\frac{p_x}{q_x}\right)$。由 $\ln u\le u-1$，

$$(\ln 2)D(P\parallel Q)\ge\sum_{x:p_x>0}(p_x-q_x)\ge 0.$$

等号 $\Leftrightarrow P=Q$。设 U 是有限字母表 $\mathcal{X}$ 上的均匀分布，

$$D(P\parallel U)=\log|\mathcal{X}|-H(P)\Rightarrow 0\le H(P)\le\log|\mathcal{X}|.$$

下界等号为点质量，上界等号为均匀分布。

非负性的证明对每个 $p_x>0$ 代入 $u=\frac{q_x}{p_x}$。第二个不等号来自 Q 在这些点上的概率和最多为一。等号要求比值都为一，并且 Q 在其余点没有剩余概率。

互信息

互信息比较两个分布：实际的联合分布，以及在保持各自分布不变的情况下假定两者独立所得到的乘积分布。它也可看作观察 Y 之后，对 X 的平均不确定性减少了多少。

$$I(X;Y)=\sum_{x,y}p_{x,y}\log\left(\frac{p_{x,y}}{p_xp_y}\right)=D\big(P_{X,Y}\parallel P_XP_Y\big).$$

$$I(X;Y)=H(X)+H(Y)-H(X,Y)=H(X)-H(X|Y)=H(Y)-H(Y|X).$$

展开对数并按边缘求和即得。$I(X;Y)\ge 0$，等号 $\Leftrightarrow X,Y$ 独立。因此 $H(X|Y)\le H(X)$，且 $H(X_1,\dots,X_n)\le\sum_i H(X_i)$。

将对数中的商拆开。含 $p_{x,y}$ 的项给负联合熵，含 p_x 或 p_y 的项对另一变量求和后给两个边缘熵。

条件互信息

给定 $Z=z$ 后仍可问 X 与 Y 是否相关。先计算此条件下的互信息，再对 z 平均，就得到条件互信息。链式法则把同时获得两项观测的信息量拆为先看一项、再补看另一项。

$$I(X;Y|Z)=\sum_z p_zD\big(P_{X,Y|z}\parallel P_{X|z}P_{Y|z}\big)=H(X|Z)-H(X|Y,Z)\ge 0.$$

$$I(X;Y,Z)=\underbrace{H(X)-H(X|Y)}_{I(X;Y)}+\underbrace{H(X|Y)-H(X|Y,Z)}_{I(X;Z|Y)}.$$

条件互信息为零意味着每个正概率条件下，两个变量的条件联合分布都是条件边缘分布的乘积。

数据处理不等式

假定 Z 只由 Y 来生成，给定 Y 后，Z 的生成过程不再依赖 X。要比较 Y 和 Z 各自携带多少关于 X 的信息，可以把 (Y,Z) 视为同一个联合观测，并按两种顺序展开其互信息。

$X\to Y\to Z$ 表示 $P_{X,Z|Y}=P_{X|Y}P_{Z|Y}$。因此

$$I(X;Z)\le I(X;Z)+I(X;Y|Z)=I(X;Y,Z)=I(X;Y)+0.$$

直观：后续处理只能保留或损失关于输入的信息。

先看 Y 时，补看 Z 不增加信息。先看 Z 时，补看 Y 所增的信息非负。两种顺序总量相同，便得到所需不等式。

散度的链式法则

两个联合分布之间的差异分成两部分：第一变量本身的分布差异，以及固定第一变量后第二变量的条件分布差异。后者按真实分布 P_X 加权。

$D\big(P_{Y|X}\parallel Q_{Y|X}\mid P_X\big)=\sum_x p_xD\big(P_{Y|X=x}\parallel Q_{Y|X=x}\big)$。

$$\log\left(\frac{p_{x,y}}{q_{x,y}}\right)=\log\left(\frac{p_x}{q_x}\right)+\log\left(\frac{p_{y|x}}{q_{y|x}}\right)$$

$$\Rightarrow D\big(P_{X,Y}\parallel Q_{X,Y}\big)=D(P_X\parallel Q_X)+D\big(P_{Y|X}\parallel Q_{Y|X}|P_X\big).$$

对同一随机处理 W，$p_{x,y}=p_xW(y|x), q_{x,y}=q_xW(y|x)$，条件散度为零。改按 Y 先展开，得输出散度 $D(P_Y\parallel Q_Y)\le D(P\parallel Q)$。

若一种输入变换对两边使用同一个随机规则，按输入先展开时，条件散度为零；按输出先展开时，联合散度至少等于输出散度。由此得到随机处理不会增大散度。

熵的凹性

将若干分布按权重混合，等同于先随机选一个分布编号，再按选中的分布采样。若只看到样本而不知道编号，样本的熵至少与已知编号时的平均熵一样大。

$P=\sum_j w_j P_j$，令 $P_{X|J=j}=P_j$、$P(J=j)=w_j$，则 $H(P)-\sum_j w_j H(P_j)=I(X;J)\ge 0$。

12 • 集中不等式与 Sanov

由均值或方差控制尾概率

如果非负变量达到阈值 a，它对期望的贡献至少是 a 乘该事件的概率。这就是 Markov 不等式。若要控制一般变量偏离均值的大小，可以先把偏离量平方，得到非负变量，再使用同一论证。

$X\ge 0, a>0$,

$$a1_{X\ge a}\le X\Rightarrow P(X\ge a)\le \mathbb{E}\frac{X}{a}.$$

$$P(|X-\mathbb{E}X|\ge t)=P((X-\mathbb{E}X)^2\ge t^2)\le \mathrm{Var}\frac{X}{t^2}\quad (t>0).$$

iid 有限方差时 $P\left(\left|\bar{X}_n-\mu\right|\ge t\right)\le \frac{\sigma^2}{nt^2}\to 0$。

对独立同分布的样本，平均值的方差是单个样本方差除以样本数。Chebyshev 因此给出样本平均值依概率趋向总体均值。

Chernoff 指数界

直接使用原变量的均值往往只能得到较粗的界。将变量变换为 e^{tX} 后，大值获得更大的权重。$t>0$ 时指数变换保序，所以可以将原上尾事件改写成指数变量的上尾事件，再用 Markov 不等式。

$M_X(t)=\mathbb{E}e^{tX}$。指数 Markov 与独立性给

$$P(X\ge a)\le \inf_{t>0} e^{-ta}M_X(t),\quad M_{\sum_i X_i}(t)=\prod_i M_{X_i}(t).$$

上尾 $t>0$；下尾 $P(X\le a)$ 取 $t<0$。

要点是对独立变量的和，指数变成乘积，取期望后可以逐项相乘。参数 t 要选到右侧尽可能小；对下尾使用负参数是因为这时小的 X 才对应大的指数变量。

Bernoulli 样本和的指数界

设 S 为 n 次独立试验中的成功总数，单次成功概率为 p。我们要控制观察到的成功比例达到 $q>p$ 的概率。先对一次试验计算指数矩，再取 n 次方。为了选最好的参数，对上界取自然对数求极小。

$S\sim \mathrm{Bin}(n,p)$，$0<p<q<1$，令 $g(t)=\ln(1-p+pe^t)-tq$。

$$P(S\ge nq)\le \inf_{t>0} e^{ng(t)},\quad g'(t)=p\frac{e^t}{1-p+pe^t}-q.$$

$$g'(t_*)=0\Rightarrow e^{t_*}=\frac{q(1-p)}{p(1-q)},\quad g(t_*)=-D_e(\mathrm{Bern}(q)\,\|\,\mathrm{Bern}(p)).$$

因此 $P(S\ge nq)\le e^{-nD_e}$；$q<p$ 时 $t_*<0$，得下尾。$D_e=(\ln 2)D$，以自然对数计算。

求导所得到的 t_* 为正，且二阶导数非负，故给出上尾情形的最小值。代回后的两个对数项分别对应成功、失败的频率，这正是两个 Bernoulli 分布之间的自然对数散度。

Stirling 公式与二项系数

组合计数中的阶乘比，常需要判断它随样本量增长的主要指数项。把 $n!$、$k!$ 和 $(n-k)!$ 分别取对数后，主要的 $n\log n$ 项抵消，留下由比例 $\frac{k}{n}$ 决定的熵项。

$n!\sim\sqrt{2\pi n}\left(\frac{n}{e}\right)^n$。从阶乘比得

$$\ln\binom{n}{k}=nh_e\left(\frac{k}{n}\right)+O(\ln(n+1)),\quad h_e(u)=-u\ln u-(1-u)\ln(1-u).$$

经验分布与类型类

一串样本中每个符号出现的比例组成经验分布。经验分布相同的序列构成一个类型类。一个类型类可以有很多不同排列，但在同一个独立抽样模型下，它们的概率相同，因为概率只取决于各符号出现了多少次。

$\Omega=[L]$，$s_i=\big|\{j:x_j=i\}\big|$，$q_i=\frac{s_i}{n}$，$T_q=\{x^n:\mathrm{type}(x^n)=q\}$。

$$|T_q|=\frac{n!}{\prod_i s_{i!}},\quad \text{类型数}\le K=(n+1)^{L-1}.$$

对 $x^n\in T_q$，

$$Q^n(x^n)=e^{-nH_e(Q)},\quad P^n(x^n)=e^{-n(H_e(Q)+D_e(Q\|P))}.$$

$$Q^n\big(T_q\big)\le 1\Rightarrow |T_q|\le e^{nH_e(Q)}.$$

下界：在 Q 下，q 是概率最大的类型之一。若计数向量 $r\neq s$ 且 $Q^n\big(T_{\frac{r}{n}}\big)>0$，选 $r_i>s_i,r_j<s_j$，从 i 移一项至 j，类型概率乘以

$$\frac{r_iq_j}{(r_j+1)q_i}=\frac{r_i}{s_i}\cdot\frac{s_j}{r_j+1}\ge 1.$$

迭代至 s，故 $Q^n(T_q)\ge\frac{1}{K}$，即 $|T_q|\ge\frac{e^{nH_e(Q)}}{K}$。

下界证明里的移动逐步缩小计数向量与目标向量的距离。每次从出现过多的符号拿走一个位置，补给出现过少的符号，且在目标分布下类型概率不下降。这样才能推出目标类型至少占所有类型的平均概率。

Sanov 界

现在把问题从某一种经验分布扩大为一组可能的经验分布。先估计每个类型类的概率，再把这些互不相交的概率加起来。样本量增加时，可达类型的个数至多按多项式增长，而不同类型的概率差别可能是指数级。

对非空可达类型集 $\mathcal{A}$，$d=\min_{q\in\mathcal{A}}D_e(Q\,\|\,P)$，

$$K^{-1}e^{-nd}\le P^n(\mathrm{type}\in\mathcal{A})\le Ke^{-nd}.$$

证明：$P^n\big(T_q\big)=Q^n\big(T_q\big)e^{-nD_e(Q\|P)}$，支持不相容时该概率为零。下界选最小散度类型，上界至多 K 项相加。直观：偏离真实分布的指数代价由 KL 散度决定。

因此只看指数增长率时，贡献最大的是与真实分布散度最小的可达类型。下界只保留这一项，上界用最大的那项控制每一项。

13 • 离散 Fourier 变换

特征标与正交性

我们想将群上的函数拆成一些对平移行为很简单的函数。特征标就是从加法群到单位圆乘法群的同态，满足 $\chi(x+y)=\chi(x)\chi(y)$。因此把输入平移 y，只会把整个函数乘上固定的复数 $\chi(y)$。

$G=\prod_j\mathbb{Z}_{n_j}$（加法），$N=|G|$。$\chi:G\to S^1$ 为群同态，

$$\chi_a(x)=\prod_j\exp\left(2\pi i a_j\frac{x_j}{n_j}\right),\quad \chi_a\chi_b=\chi_{a+b},\quad \bar{\chi_a}=\chi_{-a},\quad \chi_a(x)=\chi_x(a).$$

$S=\sum_x\chi(x)$，$\chi(y)S=\sum_x\chi(x+y)=S$。非平凡 χ 取 $\chi(y)\neq 1$，得 $S=0$。因此

$$\frac{1}{N}\sum_x\chi_a(x)\bar{\chi_b}(x)=1_{a=b}.$$

上式给出 N 个正交向量，达到 $\mathbb{C}^G$ 的维数；任何额外特征标也须与这组基正交，故不存在。

不同特征标的内积化为它们的商在全群上的求和。对非平凡特征标，平移不改求和结果，却将它乘以某个不等于一的数，所以这项和只能是零。

系数与反演

既然特征标是一组正交基，就可以通过内积取出函数在每个方向上的系数。本文按讲义在正变换中除以群的大小 N。反演时直接把各系数乘回相应基函数再相加。

$$\hat{f}(a)=\frac{1}{N}\sum_{x\in G}f(x)\bar{\chi_a}(x),\quad f(x)=\sum_{a\in G}\hat{f}(a)\chi_a(x).$$

$$\sum_a \hat{f}(a)\chi_a(x) = \sum_y f(y)\underbrace{\left(\frac{1}{N}\sum_a \chi_a(x-y)\right)}_{1_{x=y}} = f(x).$$

证明反演公式时，将系数的定义代回，把两个有限和交换。正交性使 $x \neq y$ 的项全部消失，只留下 $f(x)$。

内积和平方和

在原空间，用所有位置上的乘积平均定义内积，其中第二个因子取复共轭。转换到 Fourier 系数之后，内积变成同频率系数乘积之和。

按平均内积，

$$\frac{1}{N}\sum_x f(x)\bar{g(x)} = \sum_a \hat{f}(a)\bar{\hat{g}(a)}.$$

展开 f, g 后正交性消去 $a \neq b$；取 $g = f$ 得 $\mathbb{E}|f|^2 = \sum_a \left|\hat{f}(a)\right|^2$。

把两边的函数都展开为特征标的线性组合。相乘会产生同频和异频项，正交性恰好消去异频项。对同一函数取内积得到 Parseval 等式。

卷积与乘积

卷积先固定总和 x，再对所有分解 $x = y + z$ 的乘积 $f(y)g(z)$ 求和。对群上的独立随机变量，和的分布也由这种运算给出。

$(f * g)(x) = \sum_y f(y)g(x-y)$。

$$\widehat{f * g}(a) = \frac{1}{N}\sum_{y,z} f(y)g(z)\chi_a(\bar{y+z}) = N\hat{f}(a)\hat{g}(a).$$

逐点乘积满足 $\widehat{fg}(c) = \sum_{a+b=c}\hat{f}(a)\hat{g}(b)$，因为 $\chi_a\chi_b = \chi_{a+b}$。直观：平移对特征标只乘一个相位，故卷积变为逐频率相乘。

求卷积的 Fourier 系数时，用 $z = x - y$ 把求和指标改成相互独立的 y, z。由于特征标在加法下分解成乘积，双重和随之分成两个单和。

布尔立方体的 Fourier 展开

输入是长度为 n 的比特串，按逐坐标模二相加构成一个群。频率可以用坐标子集表示，特征标的值由这组坐标中一的个数是奇数还是偶数决定。

$x \in \{0,1\}^n$ 均匀，$S \subseteq [n]$，

$$\chi_S(x) = (-1)^{\sum_{i\in S} x_i},\quad \chi_S\chi_T = \chi_{S\triangle T},\quad \mathbb{E}\chi_S\chi_T = 1_{S=T}.$$

$$f = \sum_S \hat{f}(S)\chi_S,\quad \hat{f}(S) = \mathbb{E}[f\chi_S],\quad \mathbb{E}f = \hat{f}(\varnothing),\quad \mathrm{Var}f = \sum_{S\neq\varnothing}\hat{f}(S)^2.$$

最后一式对实值 f，由 Parseval 减去均值平方。

均匀取输入时，非空子集的奇偶性取两个值的概率相同，因此相应特征标的均值为零。两个特征标相乘时，出现两次的坐标抵消，留下对称差。

固定部分坐标与随机限制

选定一组坐标 A，再将这些坐标固定为一组比特值 α，其余坐标仍可自由变化。固定后，原来依赖于 A 中坐标的特征标会把这些坐标的作用变成一个符号因子。

$A \subseteq [n]$，$f_{A\leftarrow\alpha}$ 将 A 中坐标固定为 α。由逐项代入，

$$\chi_{T(x_{A\leftarrow\alpha})} = \chi_{T\cap A}(\alpha)\chi_{T\setminus A}(x).$$

对均匀 α 取期望，$T \cap A \neq \varnothing$ 的项消失。故

$$\mathbb{E}_\alpha\left[\widehat{f_{A\leftarrow\alpha}}(S)\right] = 1_{S\cap A=\varnothing}\hat{f}(S).$$

各坐标独立地以概率 $1-p$ 固定，$P(S\cap A = \varnothing) = p^{|S|}$，因此

$$\mathbb{E}_{A,\alpha}\left[\widehat{f_{A\leftarrow\alpha}}(S)\right] = p^{|S|}\hat{f}(S).$$

直观：阶数越高，所依赖的坐标越容易有一个被固定。

若固定值是独立均匀的，凡是涉及至少一个被固定坐标的原 Fourier 项，其符号因子平均为零。接着对坐标集合本身平均，一个频率项被保留的概率正是它涉及的全部坐标都未被固定的概率。

14 · 压缩与信道编码

不同的熵量

最小熵只考察最可能发生的结果。碰撞熵考察两次独立抽样得到同一结果的概率，即 $\sum_x p_x^2$。

Shannon 熵按真实概率平均每个结果的信息量，最大熵量则只看有正概率的结果有多少。

$H_{\min} = -\log \max_x p_x \le H_2 = -\log \sum_x p_x^2 \le H = \mathbb{E}[-\log p_X] \le H_{\max} = \log|\operatorname{supp} P|$。依次用 $\sum p_x^2 \le \max p_x$、Jensen、$D(P \,\|\, U) \ge 0$。

中间的不等号可理解为先对随机量 p_X 求平均再取负对数，结果不会超过先取负对数再平均。

无损编码与最短平均码长

给每个源符号分配一个有限比特串。单射保证单个符号可恢复；唯一可译要求连接任意多个码字后仍能唯一拆分；前缀码要求任何码字都不是另一条的开头，因而可以从左到右即时拆分。将比特串视为二叉树上的路径，前缀码的每个码字占据一整棵互不重叠的子树。

前缀码 $\Rightarrow$ 唯一可译 $\Rightarrow$ 单射。二元前缀码长 ℓ_x 满足 Kraft：$K = \sum_x 2^{-\ell_x} \le 1$。深度 $L = \max \ell_x$，码字占用互斥的 $2^{L-\ell_x}$ 片叶；反向按长度递增选空闲结点，选择深度 ℓ 时可用数为 $2^{\ell}\left(1 - \sum_{\text{已选}} 2^{-\ell_x}\right) \ge 1$。

$$q_x = \frac{2^{-\ell_x}}{K} \Rightarrow \mathbb{E}\ell_X = H(P) + D(P \,\|\, Q) - \log K \ge H(P).$$

取 $\ell_x = \lceil -\log p_x \rceil$，则 $K \le 1$ 且 $H \le \mathbb{E}\ell_X < H + 1$。Huffman：交换使最小两权落在最深兄弟叶；收缩该对后，总代价减少其权重之和，归纳得最优。

Huffman 的归纳要比较同一棵树中不同位置的权重。交换两枚叶子时，较轻权重去较深处不会增加加权总深度。可把最小的两个权重放到一对最深的兄弟叶上，然后合并成父节点。合并所损失的那一层恰好贡献这两个权重的和，所以若合并后的树还能改善，原树也能改善。

允许小概率错误的块压缩

一次编码 n 个独立同分布的源符号。样本序列的负对数概率等于各符号的信息量之和，所以由大数定律，每符号负对数概率接近 H 的序列总概率趋于一。这些序列组成典型集。

iid 下 $-\frac{1}{n}\log P(X^n) \to H$（依概率）。典型集 T 上 $2^{-n(H+\varepsilon)} \le P(x^n) \le 2^{-n(H-\varepsilon)}$，且 $P(T) \to 1$。故 $|T| \le 2^{n(H+\varepsilon)}$，编码其索引即可。反向：至多 2^{nR} 个序列可正确恢复，$P_{\text{成功}} \le P(T^c) + 2^{n(R-H+\varepsilon)}$。最优渐近速率为 H。

可达性只需给典型序列不同索引，将其他序列计入失败；下界则从解码端出发，每个索引最多正确恢复一个序列。若速率低于熵，能正确恢复的典型序列即使全部选在典型集内，其总概率也趋于零。

未知分布的压缩与解码端旁信息

分布未知时，仍可从观察到的序列计算各符号出现的次数。将经验分布的描述与序列在该类型类中的编号一起发送，解码者便能恢复原序列。若解码端另有一串相关观测 Y^n，则它只需在与这串观测相容的典型候选里寻找 X^n。

先编码类型，再编码类型类内序号，长度 $\le O(\log n) + nH(Q)$，$Q \to P$。解码端已知 Y^n：随机分箱，对给定 y^n，候选数 $\le 2^{n(H(X|Y)+\varepsilon)}$，同箱概率 2^{-nR}，并界得 $R > H(X|Y)$。分别编码 X, Y 时，按竞争序列只错 X、只错 Y、两者都错三类，依次得到

$$R_X > H(X|Y), \quad R_Y > H(Y|X), \quad R_X + R_Y > H(X,Y).$$

随机分箱先把每条可能的序列独立均匀地分到一个编号，发送时只传编号。真实序列典型时，失败意味着某条竞争序列与它号码相同。用候选数乘以单条候选撞号概率即可给出上界。

信道与容量

信道是给定输入符号后输出符号的条件分布。无记忆指给定整串输入后，每个位置由同一规则单独产生。单次传输所携带的信息量依赖输入分布，取能获得的最大互信息就是信道容量。

无记忆信道 $W^n(y^n|x^n) = \prod_i W(y_i|x_i)$，$C = \max_{P_X} I(X;Y)$。BSC 翻转率 p：$I(X;Y) = H(Y) - h_2(p) \le 1 - h_2(p)$，均匀输入取等。BEC 擦除率 ε：$H(X|Y) = \varepsilon H(X)$，故 $C = 1 - \varepsilon$。重复码奇数次多数判决，错误率 $P\left(\mathrm{Bin}(L,p) > \frac{L}{2}\right)$，$p < \frac{1}{2}$ 时随 L 指数衰减，速率 $\frac{1}{L}$。

二元对称信道传错时输出仍然是一个比特，接收者不知道哪一位翻转。二元擦除信道则会明确标出没有收到的位；未擦除时输入已知，擦除发生时保留原先的不确定性。

用随机线性码证明可达速率

把 k 位消息当作二元向量，用同一个生成矩阵得到长度 n 的码字。两个不同消息之差是非零向量，乘随机矩阵后在所有长度 n 的向量中均匀。因此即使码字彼此相关，仍可逐个估计竞争码字造成错误的概率。

选均匀二元 $k\times n$ 矩阵 G。任意 $u\neq 0$，uG 均匀。BSC 最近典型噪声解码，给定真实噪声 $z\in T$，错误需 $uG\in z+T$ 对某个 $u\neq 0$：

$$P_e\le P(Z\notin T)+2^k\frac{|T|}{2^n}\to 0\quad 若\quad \frac{k}{n}<1-h_2(p).$$

BEC 给定 r 个未擦除坐标，错误需某个 $u\neq 0$ 的投影为零，并界为 2^{k-r}，故 $\frac{k}{n}<1-\varepsilon$ 可达。

接收者尝试寻找唯一一个与接收串只相差典型噪声的码字。若真实噪声典型，错误只会发生在至少还有一条竞争码字符合这一条件时。二元擦除信道中则只须比较没被擦掉的坐标。

容量上界

设消息等概率，因而消息熵为 $\log M$。从接收串得到估计后，要描述真实消息，可以先说明估计是否正确。若正确，无需再提供消息信息；若错误，最多还剩 $M-1$ 种候选。

均匀消息 U，M 个值，经 n 次信道使用得到估计 $\hat U$，$E=1_{\hat U\neq U}$，$P_e=P(E=1)$。

$$H\big(U|\hat U\big)=H\big(E|\hat U\big)+H\big(U|E,\hat U\big)\le h_2(P_e)+P_e\log(M-1).$$

$$I\big(U;\hat U\big)\le I(X^n;Y^n)\le\sum_i I(X_i;Y_i)\le nC.$$

第二步用 $H(Y^n)\le\sum_i H(Y_i)$ 与 $H(Y^n|X^n)=\sum_i H(Y_i|X_i)$。故 $(1-P_e)\log M\le nC+h_2(P_e)$。

因此若错误概率趋于零，估计中必须包含几乎全部消息信息。另一方面，每次信道使用最多贡献 C，而解码也不能增加信息，所以每次使用可传的消息位数不能超过 C。

最小距离与 Reed–Solomon 码

两个码字的 Hamming 距离就是不同坐标的数目。若错误数量小于最小距离，结果不会变成另一个合法码字。若半径 t 内的球互不相交，接收串至多属于一个这样的球，便可纠正至多 t 个错误。

$d_{\min}=\min_{c\neq c'}d_H(c,c')$。检 $d_{\min}-1$ 错，纠 $t=\left\lfloor\frac{d_{\min}-1}{2}\right\rfloor$ 错（半径 t 球不交）。随机二元码 M 词、长 L，每对距离 $\sim\mathrm{Bin}\big(L,\frac12\big)$，故 $P(d_{\min}\le\delta L)\le\binom{M}{2}2^{-L(1-h_2(\delta))}$（$\delta<\frac12$）。RS 在 L 个不同点评价 $\deg f<k$，$d_{\min}=L-k+1$（非零差多项式至多 $k-1$ 根，且可恰取 $k-1$ 根）。收词 y_i，求 $Q(a_i)=y_iE(a_i)$，$\deg E\le t,\deg Q<k+t,E\neq 0$。错误定位多项式给出解；$Q-fE$ 在至少 $L-t\ge k+t$ 点为零，故 $Q=fE$，由 $\frac{Q}{E}$ 恢复。

Reed–Solomon 码把消息写成低次多项式的系数，码字为它在不同点上的值。两个消息的差是非零低次多项式，能重合的位置数由根数控制。纠错时把未知错误位置变成一个多项式的根。正确位置满足接收值等于原多项式值，错误位置则由定位多项式为零消去。

15 · 马尔可夫链

马尔可夫性质与转移矩阵

用 X_t 表示系统在时刻 t 的状态。给定当前状态之后，过去的所有状态对下一步的条件分布没有额外影响。若这个转移规则不随时刻变，就可用一个矩阵记录它。本文把分布写成行向量，矩阵的第 x 行给出从 x 出发的一步分布。

$P(X_{t+1}=y|X_0,...,X_t)=P(X_{t+1}=y|X_t)$；时间齐次时右式为 $P_{X_t,y}$。$P\mathbf{1}=\mathbf{1}$，行分布 $\mu_t=\mu_0P^t$；由全概率 $P^{s+t}(x,y)=\sum_z P^s(x,z)P^t(z,y)$。

多步转移先按中间状态拆开，乘上到达中间状态与从那里继续转移的概率，再求和。这正是矩阵乘法。

稳态分布的存在与不可约性

稳态分布是经过一步后保持不变的概率分布。直接追踪第 t 步的分布可能不断振荡。存在性证明改看前 T 个时刻的分布平均，平移一步时只有首尾两项不同。有限维概率单纯形紧，因此这些平均至少有一个收敛子列。

$\pi P=\pi$。有限时，$\nu_T=T^{-1}\sum_{t=0}^{T-1}\mu_0P^t$ 有收敛子列；由

$$\nu_TP-\nu_T=\frac{\mu_0P^T-\mu_0}{T}\to 0$$

得极限为稳态。**不可约**：$\forall x,y\exists t:P^t(x,y)>0$。稳态全正：$\pi_x>0$ 且 $P^t(x,y)>0\Rightarrow\pi_y\ge\pi_xP^t(x,y)>0$。

不可约性指任意两状态之间都存在一条概率为正的有限路径。只要稳态对某个起点有正概率，就可沿这样的路径向任何终点送去正概率，因此所有状态的稳态概率都大于零。

稳态为什么唯一

两种稳态的差是总质量为零的有符号向量。若差不为零，则至少有一处为正、一处为负。找一个能把正的位置送到负的位置的步数，再用一半概率做此多步转移、一半概率停在原地。

设 $\delta=\pi-\pi'$ 非零，则有 $\delta_x>0>\delta_y$。取 $P^t(x,y)>0$，$Q=\frac{I+P^t}{2}$。在 y 坐标正负项严格相消，故

$$\|\delta\|_1=\|\delta Q\|_1<\sum_{i,j}|\delta_i|Q_{i,j}=\|\delta\|_1,$$

矛盾。

该混合转移既把一部分正质量送到负的位置，也保留该位置上原有的一部分负质量。于是在此坐标上，绝对值的三角不等式严格，从而总绝对质量变小。但两种稳态都保持原样，差也必须原样不动，矛盾。

首次到达与首次返回

到达集合的时间允许等于零，起点已经在目标集合内时无需等待。返回同一状态的时间从一步开始计数。要算期望到达时间，可以先走一步，再按这一步可能落到的状态平均剩余时间。

记 $\tau_A=\min\{t\ge 0:X_t\in A\}$，$\tau_x^+=\min\{t\ge 1:X_t=x\}$。令 $h(x)=\mathbb{E}_x\tau_A$。链从集合外出发时，第一步贡献一个时间单位，随后从 y 出发的平均等待时间为 $h(y)$。因此

$$h|_A=0,\quad h(x)=1+\sum_y P(x,y)h(y)\quad (x\notin A).$$

若研究的是先到 A 还是先到另一个不相交集合 B，令 $u(x)=P_{x(\tau_A<\tau_B)}$。在 A 上答案为一，在 B 上为零。在其余状态第一步只改变接下来计算概率的起点，因此 $u(x)=\sum_y P(x,y)u(y)$，没有等待时间递推中的常数项。

返回时间与稳态的关系。 对有限不可约链，从任何状态出发都能在某个共同的有限步数内以至少某个正概率到达 x。把时间分成等长的块，未到达的概率随块数按几何速度下降，所以 $\mathbb{E}_x\tau_x^+$ 有限。记一次返回周期内访问 y 的期望次数为

$$v_y=\mathbb{E}_x\sum_{0\le t<\tau_x^+}1_{X_t=y}.$$

由于 $t<\tau_x^+$ 由截至时刻 t 的历史决定，对第 $t+1$ 步按当前状态取条件期望：

$$(vP)_y=\mathbb{E}_x\sum_{0\le t<\tau_x^+}P(X_t,y)=\mathbb{E}_x\sum_{1\le t\le\tau_x^+}1_{X_t=y}.$$

最后一项与 v_y 的唯一区别是去掉起点、加入终点，两点都是 x，因此 $vP=v$。总访问次数等于周期长度，且 $v_x=1$。归一化得到稳态，由唯一性知

$$\pi_y=\frac{v_y}{\mathbb{E}_x\tau_x^+},\quad \mathbb{E}_x\tau_x^+=\frac{1}{\pi_x}.$$

周期与分布的收敛

周期是所有可实现的正返回长度的最大公因数。即使所有状态都能互相到达，某些时刻也可能因周期限制无法到达某些状态。非周期性消除了这一障碍。

记 $d(x)=\gcd\{t\ge 1:P^t(x,x)>0\}$。若 x 到 y 和 y 到 x 分别有长度 a,b 的路径，那么在 y 插入任一长度为 r 的返回路径，得到从 x 返回的两种长度 $a+b$ 与 $a+r+b$。所以 $d(x)$ 整除 r，交换 x,y 即知周期相同。

若周期为一，可从某个状态的返回长度中选出有限多个，其最大公因数已经是一。它们的非负整数和包含所有足够大的整数。再在这些返回路径前后接上从起点到该状态、从该状态到终点的路径，就能让任意一对状态之间的路径达到所有足够大的长度。状态数有限，因此存在同一个 m，使 P^m 每一项都为正。

令 $b_y=\min_x P^m(x,y)$，$\alpha=\sum_y b_y>0$。每一行都含有同一份质量向量 b。若 $\alpha<1$，从每行减去 b 后除以 $1-\alpha$，得到转移矩阵 Q。两个概率分布相减时，总质量为零，这份相同的向量便抵消：

$$(\mu-\nu)P^m=(1-\alpha)(\mu-\nu)Q.$$

转移矩阵不会增大绝对值之和，因为 $\sum_y\left|\sum_x d_xQ_{x,y}\right|\le\sum_x|d_x|$。取 $\nu=\pi$ 并重复计算：

$$\left\|\mu P^{km}-\pi\right\|_1\le(1-\alpha)^k\|\mu-\pi\|_1\to 0.$$

若 $\alpha=1$，各行相同，一步这样的块转移后就没有差异。块之间剩余的步数也不会增大距离。将 P 换成 $\frac{I+P}{2}$ 称为懒化，它保持稳态，且每个状态都有一步返回，所以周期为一。

细致平衡与抽样用的转移过程

细致平衡要求在目标分布下，从 x 流向 y 的概率质量逐对等于反向流量。将所有流入量相加即可得到稳态。为了用马尔可夫链抽样，可先提出一个候选状态，再按接受率决定是否移动。

$\pi_x P(x,y)=\pi_y P(y,x)\Rightarrow \sum_x \pi_x P(x,y)=\pi_y$。无向连通图游走 $\pi_x=\frac{\deg(x)}{2|E|}$。给定目标 $\pi>0$、提议 Q，接受率

$$a(x,y)=\min\left(1,\frac{\pi_y Q(y,x)}{\pi_x Q(x,y)}\right)\quad (Q(x,y)>0).$$

拒绝则留原地；两向流量均为 $\min\big(\pi_x Q(x,y),\pi_y Q(y,x)\big)$。

如果提议本身在两个方向上不对称，就要同时校正目标概率和提议概率。接受率选到使两向实际流量都等于两种潜在流量中较小的一种。拒绝时留在原位置，保证每行概率和仍为一。

总变差、耦合与混合时间

总变差距离衡量两种分布在任何一个事件上最多能差多少。把两种分布写成坐标质量，正差部分的和等于负差部分绝对值的和。因此找到所有正差坐标，就能取到最大的事件差。

$d_{\mathrm{TV}}(\mu,\nu)=\frac12\sum_x|\mu_x-\nu_x|=\max_A|\mu(A)-\nu(A)|$（取 $A=\{x:\mu_x\ge\nu_x\}$）。

$$|P(X\in A)-P(Y\in A)|\le P(X\neq Y).$$

耦合先共享 $\min(\mu_x,\nu_x)$ 的质量，剩余支持不交，故 $\min P(X\neq Y)=d_{\mathrm{TV}}(\mu,\nu)$。

$$d(t)=\max_x d_{\mathrm{TV}}(P^t(x,\cdot),\pi),\quad t_{\mathrm{mix}}(\varepsilon)=\min\{t:d(t)\le\varepsilon\}.$$

一条链从 x 出发，另一条从 π 出发，相遇后同步，$d(t)\le\max_x P_x(T>t)$。

耦合是在同一个试验中同时生成两个变量，同时保留各自原来的分布。若它们取值相同，对任意事件的两个指示变量也相同；只有取值不同时才可能产生差。这给出了控制总变差的方法。混合时间取的是最难初始状态达到给定误差所需的时间。

16 • 图论基础与谱

顶点、边、游走与连通性

简单无向图的每条边连接两个不同顶点。顶点的度是与它相接的边数。游走是沿相邻顶点依次移动得到的序列，可以重复经过顶点。若两个点之间存在游走，删掉中间重复的回路后也能得到一条路径。

$G=(V,E)$，$n=|V|,m=|E|$，$\sum_v \deg v=2m$（每边贡献两端）。路径为无重复顶点的游走；可达性给出连通分支。任意连通图删圈边直到无圈，得到生成树。

连通分支把图分成若干相互不能到达的部分。在连通图中删掉一个圈上的边，原来经该边相连的两个端点仍能沿圈的剩余部分到达，所以连通性保持。

树的等价刻画

树既连通又没有圈。连通保证两点间至少有一路径；无圈保证这条路径唯一。为了数边，取最长路径的一端。若它还有另一条边，另一端要么在原路径上而形成圈，要么不在路径上而使路径可延长。

连通无圈 $\Leftrightarrow$ 任意两点间唯一路径 $\Leftrightarrow$ 连通且 $m=n-1$。两条不同路径的分叉至重合形成圈；无圈连通图的最长路径端点为叶，删叶归纳得边数。森林 $m=n-c$，c 为分支数。

因此端点是叶，删去它只少一个顶点和一条边，并留下更小的树。反过来，一个有 $n-1$ 条边的连通图若有圈，就能删边得到边数更少的生成树，与刚得到的边数公式矛盾。

邻接矩阵与有向无环图

邻接矩阵的第 u 行第 v 列标记是否能从 u 一步走到 v。矩阵相乘时枚举倒数第二步所在顶点，恰好把长度固定的游走分成不重不漏的若干类。

$A_{u,v}=1_{(u,v)\in E}$，$\big(A^k\big)_{u,v}$ 为长度 k 的游走数：

$$\big(A^{k+1}\big)_{u,v}=\sum_w \big(A^k\big)_{u,w}A_{w,v}.$$

DAG 有入度零点（否则沿入边回溯形成有向圈），逐次删除得到拓扑序。源点出发的路径数按拓扑序递推，或取 $I+A+\dots+A^{n-1}$，因 $A^n=0$。

有向无环图可将顶点排成所有边都由前指向后的顺序。依此求从起点到每个顶点的路径数时，所有前驱的值都已算好。没法再前进时自然停止，因此不需要无限求和。

二分图与 Euler 路迹

二分图要求把顶点分成两个集合，每条边都跨越两侧。闭合的游走每走一步切换一次侧，因而任何圈都是偶数长。反过来若按到根的距离奇偶分组后有同侧边，这条边与根树中的路径会组成奇圈。

二分 $\Leftrightarrow$ 无奇圈。逐分支按距根距离的奇偶分两侧，若同侧有边，将两条树路径截至最近公共祖先，再加此边，得奇圈。Euler 回路要求非孤立部分连通、各点偶度；每次入点对应出点，充分性由先走闭迹再拼接剩余闭迹。恰两奇点时加一条连接两点的边，取回路后删去此边得到开放迹。

Euler 路迹要求每条边恰好使用一次，顶点可以重复。若想闭合，每次进点须配一次出点。充分性证明先走出一条闭迹，再从闭迹上还有未用边的顶点启动另一条闭迹，将其插入原闭迹。所有边用完时得到所求。

图拉普拉斯矩阵与矩阵树定理

任选每条边的方向，是为了给它的两个端点一正一负。关联矩阵的一列只在这两个位置非零，外积在两端对角线上各贡献一，在两个交叉位置各贡献负一。这样 (B B^T) 正好是度数矩阵减邻接矩阵。

任意给边定向，关联矩阵 B 每列一 $+1$ 一 -1。$L = D - A = BB^T$，

$$x^T L x = \left\|B^T x\right\|^2 = \sum_{\{u,v\}\in E} (x_u - x_v)^2 \geq 0.$$

故 $\ker L$ 恰为每个连通分支上的常值向量，$\dim \ker L = c$。删 B 的第 i 行得 B_i，Cauchy–Binet 给

$$\det L_{\hat{\imath},\hat{\imath}} = \det\left(B_i B_i^T\right) = \sum_{S\subseteq E,|S|=n-1} \det\left(B_{i,S}\right)^2.$$

有圈则对应列线性相关，行列式为零；否则 $n-1$ 边的森林为树，按非根叶逐行展开，行列式为 ± 1。故余子式计数生成树。

删去一个顶点行后，Cauchy–Binet 将行列式按选出的边集展开。证明的关键是判断每一项的行列式：圈给出列间的相消关系；若选出的边是一棵树，可依次删去根以外的叶，其所在行只有一个非零项。最后每棵树贡献一。

平面图和染色

平面图是在平面上画出且不同边只在共有端点相交的图。固定一种画法后，面包括外部无界区域。生成树画在平面上只留下一个面，每补回一条边会在原来的一个面内把该面分成两部分。

连通平面嵌入 $n - m + f = 2$：生成树有一面，每加一条余边，m, f 各增 1。简单图 $n \geq 3$ 有 $3f \leq 2m$，故 $m \leq 3n - 6$；无三角形时 $4f \leq 2m$，得 $m \leq 2n - 4$（均在连通情形）。每个色类独立，故 $\chi(G)\alpha(G) \geq n$。

用每个面的边界长度求和时，每条边从两侧各算一次，桥的两侧即使属于同一个面也要算两次。染色则要求相邻顶点颜色不同，因此同一颜色的顶点集合是独立集。

图的特征值

拉普拉斯二次型是所有边两端取值之差的平方和，因此为零时每条边两端等值。第二小特征值是在排除全局常数方向后，单位长度向量能达到的最小二次型。连通时，这个最小值为正。

$0 = \lambda_1 \leq \lambda_2 \leq \ldots$ 为 L 的谱，$\lambda_2 = \min_{x\perp \mathbf{1}, x\neq 0} \frac{x^T L x}{x^T x}$。$d$ 正则时 $L = dI - A$。A 的谱 $\mu_1 \geq \ldots \geq \mu_n$：

$$\bar{d} \leq \mu_1 = \max_{x\neq 0} \frac{x^T A x}{x^T x} \leq \Delta.$$

下界取 $\mathbf{1}$；上界用 $2x_u x_v \leq x_u^2 + x_v^2$。连通图二分 $\Leftrightarrow \mu_n = -\mu_1$：二分时一侧变号使 A 的特征值反号；反向取最低特征向量 x，$|x|^T A |x| \geq -x^T A x = \mu_1 \|x\|^2$，故取等。$|x|$ 为首特征向量、由连通性知各坐标正，故每边 $x_u x_v < 0$。

邻接矩阵在二分图里只连接异侧顶点，给一侧的坐标变号就会让特征值取反。逆向证明需要连通性：对最小特征向量逐坐标取绝对值后得到最大特征向量；若它有零坐标，特征方程会迫使其邻点也为零，继而全为零。

17 · 概率方法

期望法与第一矩

为证明存在某个满足条件的对象，可以先随机选择对象，再用随机变量数其中有多少坏事件。如果总期望小于一，就不可能每个对象都至少有一个坏事件。同样，某个对象的成绩至少达到总体平均。

$X = \sum_i 1_{A_i}$，$\mathbb{E}X = \sum_i P(A_i)$。$\mathbb{E}X < 1 \Rightarrow P(X = 0) > 0$；$\mathbb{E}Y \geq a \Rightarrow \exists \omega : Y(\omega) \geq a$（有限样本空间）。

Ramsey：

$$\binom{n}{k}2^{1-\binom{k}{2}}<1\Rightarrow R(k,k)>n$$

（每个 k 点集成为单色团的概率为 $2^{1-\binom{k}{2}}$）。

Ramsey 数 $R(k,k)$ 指最小的顶点数，使完全图的每种双色边染色都含一份同色的 k 点团。若随机染色时同色团数的期望小于一，便说明当前顶点数还不够。

随机排列与集合对不等式

把各集合涉及的元素放到同一个均匀随机排列中。对每个集合对，考虑前一个集合的全部元素都出现在后一个集合的全部元素之前。与此事件有关的只是在这两个集合的并上的相对顺序。

Bollobás 集合对 $|A_i|=k,|B_i|=l$，$A_i\cap B_i=\varnothing$，$i\neq j\Rightarrow A_i\cap B_j\neq\varnothing$。事件 E_i 为所有 A_i 元素排在 B_i 前，$P(E_i)=\frac{1}{\binom{k+l}{k}}$。若 $x\in A_i\cap B_j,y\in A_j\cap B_i$，则 $E_i\cap E_j\Rightarrow x<y<x$。故 $\sum_i P(E_i)\le 1$。

两对的交叉相交条件提供两个见证元素。一对成立会要求第一个见证排在第二个之前，另一对成立会要求相反顺序。因此这些事件互斥，可直接将概率相加得到上界。

两两相交的等大小集合族

Erdős–Ko–Rado 定理给出一个限制：如果任意两个入选的 k 元子集都必须相交，最多能选多少个。证明用随机圆排列把一般的子集问题转成圆周上的等长区间问题。

$n\ge 2k$，两两相交的 k 元集族 $\mathcal{F}$ 满足 $|\mathcal{F}|\le\binom{n-1}{k-1}$。随机循环排列中至多 k 个长度 k 的区间属于 $\mathcal{F}$：固定一个入选区间 $[0,k-1]$，其他候选起点为 $\pm1,\dots,\pm(k-1)$；起点 $-j$ 与 $k-j$ 的区间不交，每对至多选一。每个成员成为区间的概率为 $\frac{n}{\binom{n}{k}}$，取期望即得。所有包含某个固定元素的集合取等。

固定一个已入选的区间，其余入选区间必须与它相交。把可能的起点配对，使每对对应两个不相交的区间，所以每对最多取一个。再从每个固定成员成为连续区间的概率出发，对入选区间数求期望。

先随机生成，再删除坏结构

直接随机生成一个完全合格的对象往往难以做到。可以同时控制候选对象的大小与坏结构数，随后删除少量元素排除所有坏结构。若删一次至多损失一个元素，又不会制造新坏结构，最终保留量至少是原大小减去原坏结构数。

X 为候选体量，Y 为坏结构数，每次删除一项打破一个坏结构，若删除不产生新坏结构，则某次结果保留至少 $\mathbb{E}(X-Y)$。高围长高色数：固定 $g\ge 3,k\ge 1$，取 $p=n^{-1+\frac{1}{2g}}$、$r=\lceil\frac{n}{2k}\rceil$，记 C 为长度 $\le g$ 的圈数，

$$\mathbb{E}C\le\sum_{\ell=3}^{g}n^{\ell}\frac{p^{\ell}}{2\ell}=O(\sqrt{n}),\quad P\Big(C>\frac{n}{4}\Big)=o(1),$$

$$P(\alpha(G)\ge r)\le\binom{n}{r}(1-p)^{\binom{r}{2}}\le 2^{n}e^{-p\frac{r(r-1)}{2}}=o(1).$$

同时满足 $C\le\frac{n}{4},\alpha(G)<r$ 的图删掉每个短圈的一点，得 $|V'|\ge 3\frac{n}{4}$、围长 $>g$、$\chi(G')\ge\frac{|V'|}{\alpha(G')}>k$。

高围长要求没有短圈，高染色数可通过限制最大独立集得到。随机生成后，分别估计短圈太多和独立集太大的概率，再用并集界说明能同时避开。删除短圈上的点不会使独立集变大。

修补与期望的进一步结论

在矩阵问题中，目标是保留尽可能多的 1，同时让每个指定大小的子矩阵至少有一个零。对原有每个全一子矩阵改掉一个一即可，改动后不会出现新的全一子矩阵。选择 p 时平衡元素数与坏结构数的增长阶。

二元 $n\times n$ 矩阵，各项以概率 $p=cn^{-\frac{2}{k+1}}$ 取 1。令 X 为 1 的数目，Y 为全 1 的 $k\times k$ 子矩阵数，

$$\mathbb{E}(X-Y)=n^2p-\binom{n}{k}^2p^{k^2}\ge\left(c-c^{k^2}\right)n^{2-\frac{2}{k+1}}\quad(0<c<1).$$

每个坏子矩阵改一个 1 得所需下界。随机竞赛图中固定次序的 $n-1$ 条边方向全对的概率为 $2^{-(n-1)}$；故存在至少 $n!2^{-(n-1)}$ 条 Hamilton 路的竞赛图。

竞赛图是对完全图的每一条边选择一个方向。固定顶点排列成为有向 Hamilton 路，需要排列中相邻的边都指向下一点。对所有排列的指示变量求和即得路数的期望。

无和集合的构造

无和指集合中不存在 $a+b=c$ 且三个数都属于该集合。模素数下，可先找到一个已知无和的区间，再乘同一个随机非零系数，保留被送进该区间的原整数。

有限 $B\subseteq\mathbb{Z}\setminus\{0\}$。取 $p=3k+2$ 素数且 $p>\max_{b\in B}|b|$，$I=[k+1,2k+1]$ 在模 p 下无和（$I+I$ 落在补集）。对均匀 $x\in\mathbb{F}_p^\times$，$A_x=\{b\in B:xb\in I\ \text{模}\ p\}$ 无和，$\mathbb{E}|A_x|=|B|\frac{k+1}{3k+1}\geq\frac{|B|}{3}$。
由于素数大于原整数的绝对值，每个原元素模素数都非零，乘均匀非零系数后得到均匀非零余数。所以每个元素被保留的概率相同。原整数若存在加法关系，模素数后仍存在，会与目标区间无和矛盾。

第二矩与出现概率

知道一个非负随机变量的平均值很大，仍可能只是少数极大的取值拉高了平均。第二矩控制这种情况。将变量乘以事件 $X>0$ 的指示变量，再用 Cauchy–Schwarz，将期望拆成平方平均与事件概率的乘积。
$X\geq0$，$0<\mathbb{E}X^2<\infty$。

$$(\mathbb{E}X)^2=\left(\mathbb{E}[X1_{X>0}]\right)^2\leq\mathbb{E}X^2\cdot P(X>0).$$

$$\mathrm{Var}X=o\big((\mathbb{E}X)^2\big)\Rightarrow P(X>0)\to1.$$

对计数变量，$\mathbb{E}X^2=\sum_{i,j}P\big(A_i\cap A_j\big)$，按重叠类型拆分。
当计数由许多指示变量组成，平方会数到所有有序的事件对。两件事重叠时不能直接把联合概率拆开，因此计算按两件事共有的顶点或边分类。

随机图、单调性质与阈值

在 $G(n,p)$ 中每条边独立决定是否出现。$G(n,m)$ 则固定边数，再均匀挑边集。单调递增性质一旦成立，加边也仍然成立。阈值描述加边概率从远低于某个尺度到远高于它时，性质发生概率如何从接近零变到接近一。
$G(n,p)$ 各边独立，$G(n,m)$ 均匀取 m 边。单调递增性质的阈值 p_* 满足 $\frac{p}{p_*}\to0,\infty$ 时概率分别趋于 $0,1$；尖锐阈值在固定 $0<\varepsilon<1$ 的 $(1\pm\varepsilon)p_*$ 两侧即达 $0,1$。取 p_* 为性质概率达到 $\frac12$ 的点（性质非空且非恒真）。k 份独立 $G(n,p)$ 的并服从 $G\big(n,1-(1-p)^k\big)$，全不满足性质的概率为 $\big(1-P_p\big)^k$。若 $\frac{p}{p_*}\to\infty$，可选 $k\to\infty$ 使 $1-(1-p_*)^k\leq p$，故 $1-P_p\leq2^{-k}$；若 $\frac{p}{p_*}\to0$，选 $k\to\infty$ 使 $1-(1-p)^k\leq p_*$，则 $1-\big(1-P_p\big)^k\leq\frac12$，故 $P_p\leq1-2^{-\frac1k}\to0$。
证明中把几个独立生成的图合并。任何一个子图满足性质，合并后便满足，所以失败概率至多是所有子图都失败的概率。上界与下界分别在阈值参数和较小参数上应用同一个论证。

连通性阈值：孤立点给出下界

连通图不能有孤立点，因此只需证明低于阈值时几乎总有一个孤立点。一个点孤立要求它的 $n-1$ 条潜在邻边全不存在。两个点同时孤立时，它们之间的边只计一次，因此共有 $2n-3$ 条边需要不存在。
连通性临界尺度 $p_*=\ln\frac{n}{n}$。低于窗口取 $p=\frac{\ln n-a}{n}$、$a\to\infty$ 且 $a=o(\ln n)$。孤立点数 X：

$$\mathbb{E}X=n(1-p)^{n-1}\sim e^a,\quad\mathrm{Var}\frac{X}{(\mathbb{E}X)^2}\leq\frac{1}{\mathbb{E}X}+\frac{p}{1-p}\to0.$$

协方差来自 $P(i,j\ \text{均孤立})=(1-p)^{2n-3}$。
这两个概率分别给出一阶矩和协方差。归一化方差趋于零，说明孤立点数集中在其正的、不断增长的均值附近，于是图以高概率不连通。

连通性阈值：无边割给出上界

高于阈值时，不连通必然存在一个大小最多为 $\frac{n}{2}$ 的非空分支。选定一个大小为 s 的候选顶点集，需要所有横跨它和补集的 $s(n-s)$ 条边都缺失。对所有这样的集合求并集上界即可。
$p=\frac{\ln n+a}{n}$，由单调性可设 $a\to\infty,a=o(\ln n)$。不连通必有 $1\leq s\leq\frac{n}{2}$ 的无边割；并界每项 $\binom{n}{s}(1-p)^{s(n-s)}$。若 $s\leq\frac{n}{(\ln n)^2}$，每项 $\leq\left(\frac{e^{1-a+o(1)}}{s}\right)^s$；若 $s>\frac{n}{(\ln n)^2}$，每项 $\leq\left(e(\ln n)^2n^{-\frac12}\right)^s$。两段求和均为 $o(1)$。
小集合的补集大小接近 n，可用较精确的指数估计抵消候选数量。大集合用较粗的 $n-s\geq\frac{n}{2}$ 即可，因为每个集合缺失这么多边的概率已极小。

Lovász 局部引理

有很多坏事件时，直接相加概率可能大于一。若每个事件只与少数其他事件相互依赖，可以逐步排除坏事件，同时控制这一条件下其余坏事件的概率。依赖图的要求是，一个事件与所有非邻居合在一起所生成的信息独立，仅两两独立不够。
设 $N(i)$ 是事件 A_i 在依赖图中的邻居集合。对任意指标集 S，令 $B_S=\bigcap_{j\in S}A_j^c$。若存在 $0\leq x_i<1$，使每个事件都满足

$$P(A_i) \le x_i \prod_{j \in N(i)} (1 - x_j),$$

则所有坏事件都不发生的概率至少为 $\prod_{i(1-x_i)} > 0$。

证明。 按 $|S|$ 归纳证明 $P(A_i | B_S) \le x_i$，其中 $i \notin S$。S 为空时，是假设直接给出的界。为了保证条件概率有定义，可同时归纳证明 $P(B_S) > 0$：逐个排除 S 中的事件，每一步所条件化的集合比 S 小，因此排除下一事件的条件概率至少是 $1 - x_j > 0$。

把 S 拆成邻居 $T = S \cap N(i)$ 与剩余非邻居 $U = S \setminus T$。若 T 为空，由独立性直接得到 $P(A_i | B_S) = P(A_i) \le x_i$。否则

$$P(A_i | B_S) = \frac{P(A_i \cap B_T | B_U)}{P(B_T | B_U)} \le \frac{P(A_i)}{P(B_T | B_U)}.$$

分子去掉了对 B_T 的要求，并使用了 A_i 与 B_U 独立。将 T 的元素逐个加入，在计算排除某个 j 的条件概率时，条件中只有 U 与此前加入的邻居，因而条件集合大小小于 $|S|$。归纳给出

$$P(B_T | B_U) \ge \prod_{j \in T} (1 - x_j).$$

代回后，假设中的邻居因子与分母抵消，余下的因子都不超过一，故上界为 x_i。最后对全体事件依次取条件，得到 $P\left(\bigcap_i A_i^c\right) \ge \prod_{i(1-x_i)}$。

对称形式。 若每个事件的概率至多 p，各事件至多有 d 个邻居，且 $ep(d+1) \le 1$，则可以让所有坏事件同时不发生。对 $d \ge 1$，取 $x_i = \frac{1}{d+1}$，利用 $\left(1 - \frac{1}{d+1}\right)^d \ge e^{-1}$ 验证上面的条件。若 $d = 0$，各事件相互独立，直接将补事件的概率相乘即可。