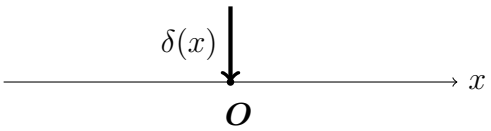
目录

1 高斯分布的几个应用场景

高斯过程的应用非常广泛，而高斯分布是高斯过程的根本。在这一节中，我们将从扩散模型、最大熵以及中心极限定理三个角度阐释高斯分布的普适性。

1.1 扩散模型 (Diffusion)

宏观运动与微观运动之间的微妙结合：宏观上，体现为逐渐向外的弥散行为，其具有方向，速度，角度以及范围。我们可以从微观和宏观两个角度来解读这个行为，宏观上总结其统计规律抑或是从微小粒子的角度上进行认知，再推及宏观进行认识。



这里，我们采取后一种策略，对扩散问题作如此建模：一些微粒从原点注入，在一维空间中开始向两侧扩散，会呈现怎么样的分布。这项工作是爱因斯坦在 1905 年首先完成的，其直接影响就是导致了阿佛加德罗常数的测定。

首先定义分布 $f(x,t)$，其与所在位置，测量时间相关；直观上，随着时间的推移，扩散进一步发生，体现在数字特征上，就是方差的增大。为得到 $f(x,t)$ 的分布，还需引入另一个量，粒子扩散速度的分布 $\rho(y,\tau)$，在 τ 的时间范围内扩散距离等于 y 的微粒所占的比例；可以对两个定义的分布 $f(x,t),\rho(y,\tau)$ 列出如下的等式：

$$\begin{aligned} f(x,t+\tau) &= \int_{\mathbb{R}} \rho(y,\tau) f(x-y,t)dy \\ \int_{\mathbb{R}} \rho(y,\tau)dy &= 1 \\ \int_{\mathbb{R}} y\rho(y,\tau)dy &= 0 \\ \int_{\mathbb{R}} y^2\rho(y,\tau)dy &= D(\tau) \end{aligned} \tag{1}$$

对 $f(x-y,t)$ 进行泰勒展开，以使用后面的三个等式从而消除 $\rho(y,\tau)$ 的存在:

$$f(x-y,t) = f(x,t) + (-y)\frac{\partial}{\partial x}f(x,t) + \frac{y^2}{2}\frac{\partial^2}{\partial x^2}f(x,t) + ... \tag{2}$$

将其代回原来的积分式，这一步逻辑上并不严谨，因为泰勒展开对原函数的贴近仅在局部奏效，积分式的积分区间却是整个实数轴，但这个方法得到的结果与实验结果高度吻合：

$$
\begin{aligned}
\int_{\mathbb{R}}[f(x,t)+(-y)\frac{\partial}{\partial x}f(x,t)+\frac{y^2}{2}\frac{\partial^2}{\partial x^2}f(x,t)]\rho(y,\tau)dy &= f(x,t+\tau)\\
f(x,t)+\frac{D(\tau)}{2}\frac{\partial^2}{\partial x^2}f(x,t) &= f(x,t+\tau)\\
f(x,t+\tau)-f(x,t) &= \frac{D(\tau)}{2}\frac{\partial^2}{\partial x^2}f(x,t)\\
\frac{f(x,t+\tau)-f(x,t)}{\tau} &= \frac{D(\tau)}{2\tau}\frac{\partial^2}{\partial x^2}f(x,t)
\end{aligned}
\tag{3}
$$

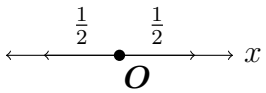
使扩散的时间间隔 τ 趋近于零，就可以将等式转为微分形式，我们对方差的定义依赖于 τ，当扩散的时长逐渐趋于 0 时，扩散的行为也近似没有发生，方差 $D(\tau)$ 也趋于 0，故可以做如下假设：

$$
\lim_{\tau\to 0}\frac{D(\tau)}{\tau}=D \implies \frac{\partial f}{\partial t}=\frac{D}{2}\frac{\partial^2 f}{\partial x^2} \tag{4}
$$

我们将上面的二阶微分方程称为扩散方程 (Diffusion Equation)。若初始的微粒分布是 $\delta(x)$，则扩散方程的解的形式如下：

$$
f(x,0)=\delta(x) \implies f(x,t)=\frac{1}{\sqrt{2\pi Dt}}\exp(-\frac{x^2}{2Dt}) \tag{5}
$$

得到的分布就是高斯分布。以上的推理都是在连续情况下进行的。从离散角度，我们可以使用随机游走模型 (Random Walk) 解释扩散模型, $P(m,n)$ 描述 m 位置于 n 时刻的微粒数目，假设微粒每次只能移动一个单位，且向左向右游走的概率均为 $\frac{1}{2}, P_{left}=P_{right}=\frac{1}{2}$



则我们可以写出如下的递归方程：

$$
\begin{aligned}
P(m,n+1) &= \frac{1}{2}P(m+1,n)+\frac{1}{2}P(m-1,n)\\
P(m,n+1)-P(m,n) &= \frac{1}{2}[P(m+1,n)-P(m,n)]-\frac{1}{2}[P(m,n)-P(m-1,n)]
\end{aligned}
\tag{6}
$$

二式可以理解为时间的一阶差分对应空间的二阶差分, 对应连续情况下的 $\frac{\partial f}{\partial t}=C\frac{\partial^2 f}{\partial x^2}$，可以解得离散情况的高斯分布。

1.2 最大熵 (Maximum Entropy)

对分布 $f(x)$，满足 $f(x)\geq 0, \int_{\mathbb{R}} f(x)dx=1$。为分布 f 定义熵, 对分布的随机性进行度量:

$$
H(f)=-\int_{\mathbb{R}} f(x)\log f(x)dx，\ 求\ \max_f H(f)
$$

不失一般性地，我们规定 f 的一二阶矩,$E[X]=m, E[X^2]=\sigma^2$, 以在一个相同条件进行比较，从而得到符合最大熵的概率分布。

在有限区间 $|x|\leq B$ 下取得最大熵的分布显然为均匀分布; 而在无限区间下的最大熵分布，我们可以通过泛函的手段解出是高斯分布，步骤如下：

假设我们已经找到了最优分布 $f_0(x)$, 令 $g(t)=H(f_0+t*h)$，h 为任意的函数，因为 f_0 已经为最优分布, $\forall t\in\mathbb{R}$, $g(0)\geq g(t) \implies \frac{dg(t)}{dt}|_{t=0}=0$

$$g(t)=H(f_0+t\cdot h)=-\int_{\mathbb{R}}(f_0+t\cdot h)\log(f_0+t\cdot h)dx \tag{7}$$

结合已有约束（一阶矩，二阶矩以及分布的总和是 1 的三个约束），使用拉格朗日乘数法对 $g(t)$ 进行优化：

$$\begin{aligned}L(t,\lambda_1,\lambda_2,\lambda_3)=&\int_{\mathbb{R}}(f_0+t\cdot h)\log(f_0+t\cdot h)\,dx\\&+\lambda_1\left[\int_{\mathbb{R}}x(f_0+t\cdot h)\,dx-m\right]\\&+\lambda_2\left[\int_{\mathbb{R}}x^2(f_0+t\cdot h)\,dx-m\right]\frac{dg(t)}{dt}|_{t=0}=0 \implies \frac{\partial L(t)}{\partial t}|_{t=0}=0\\&+\lambda_3\left[\int_{\mathbb{R}}(f_0+t\cdot h)\,dx-1\right]\\&\frac{\partial}{\partial t}L(t,\lambda_1,\lambda_2,\lambda_3)=0\end{aligned} \tag{8}$$

$$\begin{aligned}\frac{\partial}{\partial t}L&=\int_{\mathbb{R}}(h+h\log(f_0+t\cdot h))dx+\lambda_1[\int_{\mathbb{R}}x\cdot hdx-m]+\lambda_2[\int_{\mathbb{R}}x^2\cdot hdx]+\lambda_3\int_{\mathbb{R}}hdx\\&=\int_{\mathbb{R}}h(x)[\log(f_0+t\cdot h)+\lambda_1x+\lambda_2x^2+\lambda_3']dx,\ (\lambda_3'=\lambda_3+1)\\&=0,\ \forall h(x)\end{aligned} \tag{9}$$

代入 $t=0$，得到 $\int_{\mathbb{R}}h(x)[\log(f_0)+\lambda_1x+\lambda_2x^2+\lambda_3']dx=0,\ \forall h(x)$，所以

$$\log(f_0(x))=-\lambda_2x^2-\lambda_1x-\lambda_3' \implies f_0(x)=C\exp(-(\lambda_2x^2+\lambda_1x+\lambda_3')) \tag{10}$$

所以，在无限区间上具有最大熵的概率分布是高斯分布。

1.3 中心极限定理 (Central Limit Theorem)

假设 X_k 独立同分布,$E[X_k]=0, E[X_k^2]=1$，那么 $\frac{1}{\sqrt{n}}\sum_{k=1}^{n}X_k\xrightarrow{n\to\infty}\mathcal{N}(0,1)$。这项定理可以使用特征函数 $\phi_X(\omega)=E[\exp(j\omega X)]=\int_{\mathbb{R}}\exp(j\omega X)f_X(x)dx$ 推得，这个工具的本质就是傅里叶反变换，一个在电子信息领域非常常用的数学工具。

首先，证明一个引理：独立随机变量的和的分布为两者特征函数的卷积:

引理 1.1. X_1, X_2 独立, $Y = X_1 + X_2, Y \sim f_{X_1} * f_{X_2}$，步骤如下：

$$\begin{aligned} \phi_Y(\omega) = E[\exp(j\omega(X_1+X_2))] = E[\exp(j\omega X_1)]E[\exp(j\omega X_2)] = \phi_{X_1}\cdot\phi_{X_2} \\ \phi_Y = \phi_{X_1}\cdot\phi_{X_2} \overset{F.T.}{\Longleftrightarrow} f_Y = f_{X_1} * f_{X_2}(convolution) \end{aligned} \tag{11}$$

接下来，可以将两个随机变量的加和推广到多个，
$\lim\limits_{n\to\infty} \frac{\sum_{k=1}^n X_k}{\sqrt{n}}, E[X_k] = 0, var(X_k) = 1, k = 1\dots n$

$$\begin{aligned} \phi_{\frac{\sum_{k=1}^n X_k}{\sqrt{n}}}(\omega) &= E\left[\exp\left(j\omega\frac{\sum_{k=1}^n X_k}{\sqrt{n}}\right)\right] \\ &= E\left[\prod_{k=1}^n \exp\left(j\frac{\omega}{\sqrt{n}X_k}\right)\right] \\ &= \prod_{k=1}^n E\left[\exp\left(j\frac{\omega}{\sqrt{n}X_k}\right)\right] \\ &= \prod_{k=1}^n \phi_{X_k}\left(\frac{\omega}{\sqrt{n}}\right) \\ &= \left[\phi_{X_1}\left(\frac{\omega}{\sqrt{n}}\right)\right]^n \\ \phi_X\left(\frac{\omega}{\sqrt{n}}\right) = E\left[\exp\left(j\frac{\omega}{\sqrt{n}}X\right)\right] &= E\left[1 + j\frac{\omega}{\sqrt{n}}X + (j\frac{\omega}{\sqrt{n}}X)^2 + O(\frac{1}{n})\right] \end{aligned} \tag{12}$$

结合 X_1 的一二阶矩，我们可以得到：

$$\lim_{n\to\infty} \phi_{\frac{\sum_{k=1}^n X_k}{\sqrt{n}}}(\omega) = \lim_{n\to\infty}\left[1 - \frac{\omega^2}{2}\cdot\frac{1}{n} + O\left(\frac{1}{n}\right)\right] = \exp(-\frac{\omega^2}{2}) \tag{13}$$

而这个函数的傅里叶变换就是高斯分布。
（高斯函数的傅里叶变换对:$\exp\left(-\frac{t^2}{2\sigma^2}\right) \Leftrightarrow \exp\left(-\frac{\sigma^2\omega^2}{2}\right)$）

随机游走的综合效应可以用中心极限定理中多个随机变量的和来表征:

$$S_n = X_1 + \cdots + X_n, \quad X_k \overset{\text{i.i.d.}}{\sim} \mathrm{B}\begin{bmatrix} \Delta x & -\Delta x \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \quad E[X_k] = 0, \quad D(X_k) = (\Delta x)^2$$

当游走的时间间隔趋于 0，而步数趋于无穷，且两者乘积保持不变时 $n = \frac{t}{\Delta t} \to \infty$，就可以得到连续情况下的随机游走:

$$\begin{aligned} \tilde{X_k} = \frac{X_k}{\sqrt{(\Delta x)^2}}, \quad X_k = \Delta x\tilde{X_k}, \quad S_n = \Delta x\sum_{k=1}^n \tilde{X_k} \\ \frac{S_n}{\Delta x\sqrt{n}} = \frac{\sum_{k=1}^n \tilde{X_k}}{\sqrt{n}} \xrightarrow{n\to\infty} \mathcal{N}(0,1) \\ \Delta x\sqrt{n} = \Delta x\sqrt{\frac{t}{\Delta t}} = \sqrt{t}\cdot\frac{\Delta x}{\sqrt{\Delta t}} = \sqrt{t}\cdot\sqrt{\frac{(\Delta x)^2}{\Delta t}} = \sqrt{t}\cdot\sqrt{D} \end{aligned} \tag{14}$$

在扩散模型的离散情况中，我们定义了 $\lim\limits_{\tau\to 0}\frac{D(\tau)}{\tau} = D$, 即随着时间的缩短，扩散的距离逐渐趋于零，并假设了两者比例的极限为 D。上式的 $(\Delta x)^2$ 恰好对应单步转移的方差

$D(X_k)$，即连续情况下的 $D(\tau)$；所以，原式对应 $\sqrt{D}$。将所得结果代回上面的概率分布推导，可以得到：

$$\frac{S_n}{\Delta x\sqrt{n}} = \frac{S_n}{\sqrt{D\cdot t}} \xrightarrow{n\to\infty} \mathcal{N}(0,1), \quad D(S_n) = D\cdot t \tag{15}$$

所以，使用中心极限定理推得的扩散模型概率分布与爱因斯坦的结果完全吻合，均为高斯分布；不过这个阐述是从单个微粒的游走切入，而爱因斯坦的工作是从整体的统计性质入手。

2 高斯过程的性质

定义 2.1. *$X(t)$ 是高斯过程:*
$\forall n\in\mathbb{N}, \forall t_1,\ldots,t_n$, $X=(X(t_1),\ldots,X(t_n))^T$, $X\sim\mathcal{N}(\mu,\Sigma)$, *其中均值* $\mu=E[X]$，*协方差矩阵* $\Sigma=[(X-\mu)(X-\mu)^T]$。

- n=1 时,$X\sim\mathcal{N}(\mu,\sigma^2)$, $f_X(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- n=2 时,$X\sim\mathcal{N}(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$,
  $f_{X_1,X_2}(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right)\right]$
- 推广到更高维度，我们需使用矢量、矩阵的形式来表征其概率密度函数:

$$\begin{aligned}&\forall X\in\mathbb{N},\ X\sim\mathcal{N}(\mu,\Sigma),\ \mu\in\mathbb{R}^n,\ \Sigma\in\mathbb{R}^{n\times n},\\ &f_X(x)=\frac{1}{(2\pi)^{\frac{n}{2}}(\det\Sigma)^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right]\end{aligned} \tag{16}$$

首先，我们验证这个函数确实是一个概率密度，检验其是否具有非负性全空间积分为 1 的性质：

- 非负性：协方差矩阵 Σ 是正定阵，其行列式必然为正，故该分布函数一定非负；
- 积分为 1：$\int_{\mathbb{R}^n} f_X(x)dx=\frac{1}{(2\pi)^{\frac{n}{2}}(\det\Sigma)^{\frac{1}{2}}}\int_{\mathbb{R}^n}\exp\left[-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right]dx=1$

下面对积分为 1 进行验证：首先，指数上的二次型需消去交叉项，对协方差矩阵的逆进行对角化；
已知协方差阵 $\Sigma=\Sigma^T$, Σ 正定, 可以对其进行特征值分解:

$$\begin{aligned}\Sigma&=U^T\Lambda U,\quad \Lambda=diag(\lambda_1,\cdots,\lambda_n),\quad U^{-1}=U^T\\ \Sigma^{-1}&=\left(U^T\Lambda U\right)^{-1}=U^{-1}\Lambda^{-1}(U^T)^{-1}=U^T\cdot diag(\lambda_1^{-1},\cdots,\lambda_n^{-1})\cdot U\\ &=U^T\cdot diag(\lambda_1^{-\frac{1}{2}},\cdots,\lambda_n^{-\frac{1}{2}})\cdot diag(\lambda_1^{-\frac{1}{2}},\cdots,\lambda_n^{-\frac{1}{2}})\cdot U=U^T\Lambda^{-\frac{1}{2}}\Lambda^{-\frac{1}{2}}U\\ &=B^T\cdot B(B=\Lambda^{-\frac{1}{2}}U)\end{aligned} \tag{17}$$

代回原积分式：

$$\begin{aligned}
&\frac{1}{(2\pi)^{\frac{n}{2}}(\det\Sigma)^{\frac{1}{2}}}\int_{\mathbb{R}^n}\exp\left[-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right]dx\\
&=\frac{1}{(2\pi)^{\frac{n}{2}}(\det\Sigma)^{\frac{1}{2}}}\int_{\mathbb{R}^n}\exp\left[-\frac{1}{2}(X-\mu)^TB^TB(X-\mu)\right]dx\\
&=\frac{1}{(2\pi)^{\frac{n}{2}}}\int_{\mathbb{R}^n}\exp\left[-\frac{1}{2}Y^T\cdot Y\right]dY(Y=B(X-\mu),dY=(\det\Sigma)^{-\frac{1}{2}}dx)\\
&=\frac{1}{(2\pi)^{\frac{n}{2}}}\left[\int_{\mathbb{R}}-\frac{1}{2}y_k^2dy_k\right]^n\\
&=1
\end{aligned}\tag{18}$$

故，该函数是一个概率密度函数。然而，通过概率密度函数来检验是否为 n 维高斯分布没有可行性；在此，我们引入 n 维特征函数 (Characteristic Function) 来处理高斯分布：

定义 2.2. *n 维特征函数, 随机变量 $X\in\mathbb{R}^n$，特征函数 $\phi_X(\omega)=E[\exp(j\omega^TX)]$*

n 维高斯 $X\sim\mathcal{N}(\mu,\,\Sigma)$ 的特征函数 $\phi_X(\omega)$：

$$\begin{aligned}
\phi_X(\omega)&=\int_{\mathbb{R}}\exp(j\omega^Tx)f_X(x)dx\\
&=\frac{1}{(2\pi)^{\frac{n}{2}}(\det\Sigma)^{\frac{1}{2}}}\int_{\mathbb{R}^n}\exp(j\omega^Tx)\exp\left[-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right]dx
\end{aligned}\tag{19}$$

这里采用一个并不严谨但是非常直观的方法，找到高维表达式在一维的对应，然后将算式的一维结果对应回高维：

$$j\omega^TX+\left[-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right]\Leftrightarrow-\frac{1}{2\sigma^2}(x-\mu)^2+j\omega x$$

$$-\frac{1}{2}(X-\mu-j\Sigma\omega)^T\Sigma^{-1}(X-\mu-j\Sigma\omega)+j\omega^T\mu-\frac{1}{2}\omega^T\Sigma\omega\Leftrightarrow-\frac{1}{2\sigma^2}(x-\mu-j\sigma^2\omega)^2+j\omega\mu-\frac{1}{2}\sigma^2\omega^2$$

将配方得到的结果代回积分式：

$$\begin{aligned}
&\phi_X(\omega)\\
&=\frac{1}{(2\pi)^{\frac{n}{2}}(\det\Sigma)^{\frac{1}{2}}}\int_{\mathbb{R}^n}\exp\left[-\frac{1}{2}(X-\mu-j\Sigma\omega)^T\Sigma^{-1}(X-\mu-j\Sigma\omega)+j\omega^T\mu-\frac{1}{2}\omega^T\Sigma\omega\right]dx\\
&=\exp(j\omega^T\mu-\frac{1}{2}\omega^T\Sigma\omega)\cdot\frac{1}{(2\pi)^{\frac{n}{2}}(\det\Sigma)^{\frac{1}{2}}}\int_{\mathbb{R}^n}\exp\left[-\frac{1}{2}(X-\mu-j\Sigma\omega)^T\Sigma^{-1}(X-\mu-j\Sigma\omega)\right]dx\\
&=\exp(j\omega^T\mu-\frac{1}{2}\omega^T\Sigma\omega)\cdot1
\end{aligned}\tag{20}$$

积分内剩下的内容恰好对应一个高斯密度函数，故积分为 1，而高斯分布的随机变量 X 的特征函数为 $\phi_X(\omega)=\exp\left(j\omega^T-\frac{1}{2}\omega^T\Sigma\omega\right)$, 比概率密度的表示更简洁扼要，不涉及协方差阵的逆。

2.1 高斯分布的线性（Linearity Property）

定义 2.3. 高斯分布具有线性:$X \in \mathbb{R}^n$, $X \sim \mathcal{N}(\mu, \Sigma)$, $A \in \mathbb{R}^{m\times n}$，则

$$Y = AX \in \mathbb{R}^m \implies Y \sim \mathcal{N}(A\mu, A\Sigma A^T)$$

证明如下：

$$\begin{aligned}\phi_Y(\omega) &= E[\exp(j\omega^T Y)] = E[\exp(j\omega^T AX)] \\ &= E[\exp(j(A^T\omega)^T X)] = \phi_X(A^T\omega) = \exp(j(A^T\omega)^T\mu - \frac{1}{2}(A^T\omega)^T\Sigma(A^T\omega)) \\ &= \exp(j\omega^T(A\mu) - \frac{1}{2}\omega^T(A\Sigma A^T)\omega) \implies Y \sim \mathcal{N}(A\mu, A\Sigma A^T)\end{aligned} \tag{21}$$

只要特征函数符合高斯分布的形态，那么该随机变量就是高斯分布。

推论 2.4. 联合高斯分布的任意边缘分布都是高斯分布：$X = (X_1, \cdots, X_n)^T \sim \mathcal{N}(\mu, \Sigma)$, $\tilde{X} = (X_{n_1}, \cdots, X_{n_k})^T$, $\{n_1, \cdots, n_k\} \subset \{1, \cdots, n\}$

可以构造一个 $k \times n$ 矩阵，左乘在 X 上，可以得到目标边缘分布的随机变量矢量，由定理可知，其是高斯分布。联合高斯必然导致边缘高斯，然而边缘高斯无法反推得到联合高斯。这里给出一个高斯分布的判据：

$$X \in \mathbb{R}^n, X \sim \mathcal{N} \iff \forall \alpha \in \mathbb{R}^n, \alpha^T X \sim \mathcal{N}$$

证明:$\Rightarrow$ 显然，这里只讨论 $\Leftarrow$ 的情况：

$$\begin{aligned}\phi_X(\omega) &= E[\exp(j\omega^T X)](\omega \in \mathbb{R}^n \implies \omega^T X \sim \mathcal{N}) \\ &= \phi_{\omega^T X}(1) = exp(j \cdot 1 \cdot \mu_{\omega^T X} - \frac{1}{2}1 \cdot \sigma^2_{\omega^T X} \cdot 1) = \exp(j\mu_{\omega^T X} - \frac{1}{2}\sigma^2_{\omega^T X})\end{aligned} \tag{22}$$

$$\begin{cases}\mu_{\omega^T X} = E[\omega^T X] = \omega^T \mu_X \\ \sigma^2_{\omega^T X} = E[\omega^T X - \omega^T \mu_X]^2 = E[\omega^T (X - \mu_X)]^2 = \omega^T \Sigma_X \omega\end{cases} \Rightarrow \phi_X(\omega) = \exp(j\omega^T\mu_X - \frac{1}{2}\omega^T\Sigma_X\omega)$$

所以,$X \sim \mathcal{N}(\mu_X, \Sigma_X)$

若随机变量符合联合高斯分布，则任意两个随机变量不相关等价于所有随机变量相互独立；一般情况下，独立可以推导出不相关，但不相关无法得到独立，但在联合高斯的前提下，两者是等价的，这一点可以通过协方差矩阵进行验证：

$$\begin{aligned}X = (X_1, \cdots, X_n)^T \sim \mathcal{N}, \quad Uncorrelated : E[X_i X_j] = E[X_i]E[X_j] \\ \Sigma_{ij} = E[X_i - E[X_i]]E[X_j - E[X_j]] = E[X_i X_j] - E[X_i]E[X_j] = 0 (i \neq j)\end{aligned} \tag{23}$$

这说明，非对角线的元均为 0，协方差矩阵 Σ 为对角阵 $diag(\sigma_1^2, \cdots, \sigma_n^2)$，对应的概率

密度为：

$$\begin{aligned} f_{X_1,\cdots,X_n}(x_1,\cdots,x_n) &= \frac{1}{(2\pi)^{\frac{n}{2}}(\det\Sigma)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\sum_{k=1}^{n}\frac{(X_k-\mu_k)^2}{\sigma_k^2}\right) \\ &= \prod_{k=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(X_k-\mu_k)^2}{2\sigma_k^2}\right) \\ &= \prod_{k=1}^{n} f_{X_k}(x_k) \end{aligned} \tag{24}$$

所以，若一组随机变量符合联合高斯分布，则两两不相关等价于两两相互独立。在主成分分析 (Principal Component Analysis) 中，我们通过线性变换去掉了各个随机变量间的相关性；若进一步要求这一组随机变量满足联合高斯分布，则对其使用 PCA，就等同于使用独立分量分析 (Identical Component Analysis)；这也是这套分析方法的起源，其中比较常见的应用就是盲源分离 (Blind Source Separation)。

而另一个非常时兴的高斯分布线性的应用就是 AIGC 画图使用的 Diffusion 算法中的一个技术点: $X_0, X_1, X_2, \cdots, X_N \in \mathbb{R}^n, X_k = \sqrt{1-\alpha_k}X_{k-1} + \sqrt{\alpha_k}\epsilon_k, \epsilon_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$ 该随机变量 X_N 是一系列独立同分布的高斯随机变量 ϵ_k 线性组合而成，其分布显然是高斯分布，具体参数的求解在原始论文中使用了 Reparametric Trick, 定义参量 $\beta_k = 1-\alpha_k$：

$$\begin{aligned} X_k &= \sqrt{\beta_k}X_{k-1} + \sqrt{1-\beta_k}\epsilon_k = \sqrt{\beta_k}(\sqrt{\beta_{k-1}}X_{k-2} + \sqrt{1-\beta_{k-1}}\epsilon_{k-1}) + \sqrt{1-\beta_k}\epsilon_k \\ &= \sqrt{\beta_k\beta_{k-1}}X_{k-2} + \sqrt{\beta_k(1-\beta_{k-1})}\epsilon_{k-1} + \sqrt{1-\beta_k}\epsilon_k \end{aligned} \tag{25}$$

注意到 $\epsilon_{k-1}, \epsilon_k$ 是相互独立的高斯变量，故而两者的和就可以写作

$$\mathcal{N}(0, \sqrt{\beta_k(1-\beta_{k-1})}I) + \mathcal{N}(0, \sqrt{1-\beta_k}I) \Rightarrow \mathcal{N}(0, \sqrt{1-\beta_k\beta_{k-1}}I)$$

而原式推导可以写作 $X_k = \sqrt{\beta_k\beta_{k-1}}X_{k-2} + \sqrt{1-\beta_k\beta_{k-1}}\epsilon_0$。以此类推，可以得到：

$$\begin{cases} X_N = \sqrt{\tilde{\beta}}X_0 + \sqrt{1-\tilde{\beta}}\epsilon_0 \\ \tilde{\beta} = \prod_{k=0}^{N}\beta_k \end{cases}$$

2.2 Cochran 定理以及条件分布

高斯分布线性的另一个应用就是 Cochran 定理：

定理 2.5. *(Cochran 定理)* $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$，样本均值 $\bar{X} = \frac{1}{n}\sum_{k=1}^{n}X_k$，样本方差 $\bar{S} = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$，样本均值 $\bar{X}$ 与样本方差 $\bar{S}$ 相互独立。

通过样本均值 $\bar{X}$ 可以获得对随机变量均值 $\hat{\mu}$ 的估计，而通过样本方差 $\bar{S}$ 可以获得

对随机变量方差 $\hat{\sigma^2}$ 的估计。构造一个酉矩阵 (unitary matrix)A，如下：

$$A = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}_{n\times n}$$

令 Y 等于 X 左乘 A,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_n \end{bmatrix} = A \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_n \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_n \end{bmatrix}$$

由于 A 是酉矩阵，不会对随机变量的能量性质造成影响，故 $\sum_{i=1}^{n} Y_i^2 = \sum_{j=1}^{n} X_j^2$; 同时由于酉矩阵的各行向量彼此正交且 $X_1, \cdots, X_n$ 相互独立，所以 Y 的各分量之间相互独立；$Y_1 = \sqrt{n}\bar{X}, \quad \Sigma_Y = A\Sigma_X A^T = A(\sigma^2 I)A^T = \sigma^2(AIA^T) = \sigma^2 I$

$$\begin{aligned} (n-1)\bar{S} &= \sum_{k=1}^{n}(X_k - \bar{X})^2 = \sum_{k=1}^{n}(X_k^2 - 2\cdot\bar{X}X_k + \bar{X}^2) \\ &= \sum_{k=1}^{n} X_k^2 - 2\cdot\bar{X}\cdot n\bar{X} + n\bar{X}^2 = \sum_{k=1}^{n} X_k^2 - n\bar{X}^2 \\ (n-1)\bar{S} &= \sum_{k=1}^{n} X_k^2 - n\bar{X}^2 = \sum_{k=1}^{n} Y_k^2 - Y_1^2 = \sum_{k=2}^{n} Y_k^2 \end{aligned} \tag{26}$$

由于 $Y_2, \cdots, Y_n$ 与 $Y_1 = \sqrt{n}\bar{X}$ 独立，所以样本均值 $\bar{X}$ 与样本方差 $\bar{S}$ 相互独立。同时我们注意到，$(n-1)\bar{S}$ 有 n-1 个自由度 $(Y_2, \cdots, Y_n)$，所以应该除以 $(n-1)$ 进行标准化。到这里，对高斯分布的线性性质已经进行了较为深入的讨论。

下面，我们对两个高斯随机变量的条件分布进行探讨：

$$(X_1, X_2) \sim \mathcal{N}\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\right), \quad X_1, \mu_1 \in \mathbb{R}^m,\ X_2, \mu_2 \in \mathbb{R}^n \quad X_2|X_1 \sim ?$$

$$\begin{aligned} f_{X_2|X_1} &= \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_1}(x_1)} \\ &= C\exp\left(-\frac{1}{2}\begin{pmatrix}X_1^T - \mu_1^T & X_2^T - \mu_2^T\end{pmatrix}\begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}^{-1}\begin{pmatrix}X_1 - \mu_1 \\ X_2 - \mu_2\end{pmatrix} + \frac{1}{2}(X_1^T - \mu_1^T)\Sigma_{11}^{-1}(X_1 - \mu_1)\right) \end{aligned} \tag{27}$$

这里的关键是对协方差矩阵求逆，首先我们对该矩阵进行对角化:

$$\begin{aligned}
&\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix} = \begin{bmatrix} I & 0\\ \Sigma_{21}\Sigma_{11}^{-1} & I\end{bmatrix}\begin{bmatrix}\Sigma_{11} & 0\\ 0 & \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\end{bmatrix}\begin{bmatrix} I & \Sigma_{11}^{-1}\Sigma_{12}\\ 0 & I\end{bmatrix}\\
&\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}^{-1} = \begin{bmatrix} I & -\Sigma_{11}^{-1}\Sigma_{12}\\ 0 & I\end{bmatrix}\begin{bmatrix}\Sigma_{11}^{-1} & 0\\ 0 & (\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\end{bmatrix}\begin{bmatrix} I & 0\\ -\Sigma_{21}\Sigma_{11}^{-1} & I\end{bmatrix}
\end{aligned} \tag{28}$$

代回原式指数部分的二次型进行化简，可以得到：

$$\begin{aligned}
&\begin{pmatrix}X_1^T-\mu_1^T & X_2^T-\mu_2^T\end{pmatrix}\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}^{-1}\begin{pmatrix}X_1-\mu_1\\ X_2-\mu_2\end{pmatrix}\\
&=\begin{pmatrix}X_1^T-\mu_1^T & X_2^T-\mu_2^T-(X_1^T-\mu_1^T)\Sigma_{11}^{-1}\Sigma_{12}\end{pmatrix}\begin{bmatrix}\Sigma_{11}^{-1} & 0\\ 0 & (\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\end{bmatrix}\\
&\begin{pmatrix}X_1-\mu_1\\ X_2-\mu_2-\Sigma_{21}\Sigma_{11}^{-1}(X_1-\mu_1)\end{pmatrix}\\
&=(X_1^T-\mu_1^T)\Sigma_{11}^{-1}(X_1-\mu_1)\\
&+[X_2^T-\mu_2^T-(X_1^T-\mu_1^T)\Sigma_{11}^{-1}\Sigma_{12}](\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}(X_2-\mu_2-\Sigma_{21}\Sigma_{11}^{-1}(X_1-\mu_1))
\end{aligned} \tag{29}$$

故，高斯变量条件分布的密度函数可以写作：

$$\begin{aligned}
&f_{X_2|X_1}=C\exp[-\frac{1}{2}([X_2^T-\mu_2^T-(X_1^T-\mu_1^T)\Sigma_{11}^{-1}\Sigma_{12}](\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}(X_2-\mu_2-\Sigma_{21}\Sigma_{11}^{-1}(X_1-\mu_1)))]\\
&X_2|X_1\sim\mathcal{N}(\mu_2+\Sigma_{21}\Sigma_{11}^{-1}(X_1-\mu_1),\,\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})
\end{aligned} \tag{30}$$

矩阵 $\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ 是正定的，这一点可以通过柯西不等式证明得出。当两个随机变量均是一维高斯随机变量时，$X_1,X_2\in\mathbb{R}^1$

$$E[X_2|X_1]=\mu_2+\frac{\sigma_{21}}{\sigma_{11}}(X_1-\mu_1),\ D(X_2|X_1)=\sigma_{22}-\frac{\sigma_{21}\sigma_{12}}{\sigma_{11}}=\sigma_{22}-\frac{\sigma_{12}^2}{\sigma_{11}}$$

高斯随机变量的条件期望是均方意义下的最优估计；条件后的均值 $E[X_2|X_1]$ 更精确，是随机变量本身的均值加上投影系数的线性修正项 (X_1 对 X_2 的影响，先验知识)，这在高斯的前提下不仅是最简单的也是最优的；而条件后的方差 $D(X_2|X_1)$ 小于原有方差，因为先验知识减小了随机的不确定度，而其不确定性体现在方差中。

若两者不相关,$E[X_1X_2]=E[X_1]E[X_2],\sigma_{12}=0$，则 $E[X_2|X_1]=E[X_2],\ D(X_2|X_1)=D(X_2)$,$X_1$ 的先验知识对 X_2 不构成影响，也无法减小其随机性。

例 2.6. $X_1,X_2\overset{i.i.d.}{\sim}\mathcal{N}(0,1),\ E[X_1+X_2|X_1-X_2]=?,\quad E[(X_1+X_2)^2|X_1-X_2]=?$

令 $X_1'=X_1+X_2,\ X_2'=X_1-X_2,\ \begin{bmatrix}X_1'\\ X_2'\end{bmatrix}=\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}X_1\\ X_2\end{bmatrix}$,

$$(X_1',X_2')\sim\mathcal{N}(A\mu,A\Sigma A^T)=\mathcal{N}(\begin{bmatrix}0\\ 0\end{bmatrix},\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix})$$

$$
\begin{aligned}
X_1'|X_2' &\sim \mathcal{N}(E[X_1'] + \Sigma_{12}\Sigma_{22}^{-1}(X_2' - E[X_2']), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \\
E[X_1'|X_2'] &= E[X_1'] + \Sigma_{12}\Sigma_{22}^{-1}(X_2' - E[X_2']) = 0 \\
E[(X_1')^2|X_2'] &= D(X_1'|X_2') + (E[X_1'|X_2'])^2 = [\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}] + [E[X_1'] + \Sigma_{12}\Sigma_{22}^{-1}(X_2' - E[X_2'])]^2 \\
&= 2 + 0 = 2(trivial)
\end{aligned} \tag{31}
$$

$X_1'|X_2'$ 本身也是一个随机变量，随机性体现在 X_2' 上，故可以定义一个新的随机变量 $Y_{X_1'|X_2'} = X_1 + X_2|X_1 - X_2 \sim \mathcal{N}(0, 2)$。将前面的一二阶矩一般化，讨论更高阶矩的情况 $(E[(X_1')^n|X_2'] = E[Y^n_{X_1'|X_2'}])$，并由此简要介绍一些高斯函数的非线性应用。

$$
Y \sim \mathcal{N}(0, \sigma^2) \Rightarrow E[Y^n] = \begin{cases} 0, n & odd \\ ?, n & even \end{cases}
$$

$$
\begin{aligned}
E[Y^{2k}] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}} y^{2k}\exp\left(-\frac{y^2}{2\sigma^2}\right)dy = \frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}} y^{2k} d\left(-\sigma^2\exp\left(-\frac{y^2}{2\sigma^2}\right)\right) \\
&= \frac{1}{\sqrt{2\pi}\sigma}\left[\int_{\mathbb{R}} d\left(-\sigma^2 y^{2k-1}\exp\left(-\frac{y^2}{2\sigma^2}\right)\right) - (-\sigma^2)\cdot(2k-1)\int_{\mathbb{R}} y^{2k-2}\exp\left(-\frac{y^2}{2\sigma^2}\right)dy\right] \\
&= \sigma^2\cdot(2k-1)\cdot\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}} y^{2k-2}\exp\left(-\frac{y^2}{2\sigma^2}\right)dy = \sigma^2\cdot(2k-1)\cdot E[Y^{2k-2}] \\
&= \sigma^{2k}\cdot(2k-1)!!
\end{aligned} \tag{32}
$$

$$
\Longrightarrow\ Y \sim \mathcal{N}(0, \sigma^2) \Rightarrow E[Y^n] = \begin{cases} 0, n & odd \\ \sigma^{2k}\cdot(2k-1)!!, n & even \end{cases}
$$

定理 2.7. $Y \sim \mathcal{N}(\mu, \sigma^2), E[\cos Y] = \frac{1}{2}\exp(-\frac{\sigma^2}{2})\cos(\mu)$

展开为复指数函数，再使用特征函数进行处理，即可得证：

$$
\begin{aligned}
E[\cos Y] &= E[\frac{1}{2}(\exp(jY) + \exp(-jY))] = \frac{1}{2}[\phi_Y(1) + \phi_Y(-1)] \\
&= \frac{1}{2}[\exp(j\cdot 1\cdot\mu - \frac{\sigma^2\cdot 1}{2}) + \exp(j\cdot(-1)\cdot\mu - \frac{\sigma^2\cdot 1}{2})] \\
&= \frac{1}{2}\exp(-\frac{\sigma^2}{2})\cos(\mu)
\end{aligned} \tag{33}
$$

出于篇幅考量，这里再介绍一个常用于处理高斯分布非线性应用的数学定理，Price 定理：

定理 2.8. *(Price Theorem)* $g(X_1, X_2)$ 是个一般的非线性函数，$(X_1, X_2) \sim \mathcal{N}, \mathcal{N}, \sigma^{\in}_{\infty}, \sigma^{\in}_{\in}, \rho$，则有

$$
\frac{\partial E[g(X_1, X_2)]}{\partial\rho} = \sigma_1\sigma_2 E[\frac{\partial^2 g(X_1, X_2)}{\partial X_1\partial X_2}]
$$

3 总结

这篇报告主要来自我对张颢老师随机过程课程中高斯过程的笔记整理。首先从扩散模型，最大熵以及中心极限定理的角度介绍了高斯分布的普遍存在性，穿插着随机游走的两种不同解释；之后对多维高斯分布的线性性质进行了较为细致的介绍以及推导，最后稍微提及了部分非线性的应用。