[AI2613 随机过程][第十四讲] 马尔可夫过程、郎之万蒙特卡洛与扩散模型

张驰豪

最后更新：*2025 年 6 月 13 日*

目录

1 马尔可夫半群理论 **2**

1.1 马尔可夫转移核与半群，算子解释 2

1.2 马尔可夫半群的生成元 3

1.3 柯尔莫哥洛夫前向与后向方程（Kolmogorov Forward / Backward Equations） 5

1.4 使用 KBE/KFE 研究扩散过程的分布演化 7

2 朗之万动力学的收敛性（Convergence of Langevin Dynamics） **9**

3 去噪扩散概率模型（Denoising Diffusion Probabilistic Model） **10**

3.1 OU 过程的逆过程的演化方程证明（1-维情况） 10

3.2 分数匹配（Score Matching） 12

4 高维推广 **13**

1 马尔可夫半群理论

本课程中反复出现的一个主题便是如何描述一个随机过程。我们通常有两种方式：其一是描述随机变量在每一步的局部游走（例如，使用随机微分方程），其二是定义转移“矩阵”。我们这一节将为伊藤过程定义出对应的转移矩阵，实际上由于我们的状态空间通常是无限的，我们称之为转移算子。实际上，我们要做的事情和之前在连续时间离散空间的马尔可夫链的研究中要做的事情十分类似。当然，一旦涉及无穷，许多在有限的场合自然而然的性质（比如求和号的交换）就不一定成立。和之前离散空间的做法类似，我们今天不太纠结于特别严格的论述一些类似操作的合法性。

对于马尔可夫半群理论比较严格的阐述，可以参考专著 *Brownian Motion, Martingales, and Stochastic Calculus* by Jean-François Le Gall。

1.1 马尔可夫转移核与半群，算子解释

我们忆及，有限马尔可夫链的转移矩阵 $P = (P_{ij})_{i,j\in[n]}$ 定义如下：

$$P(i,j) = \mathbb{P}\left[X_{t+1} = j \mid X_t = i\right].$$

并且我们有

$$P^t(i,j) = \mathbb{P}\left[X_{s+t} = j | X_s = i\right],$$

其中 $P^t(i,j)$ 表示 X_s 在恰好 t 步内从 i 游走到 j 的概率。依据全概率公式（或矩阵乘法的定义），我们有

$$P^{s+t}(i,j) = (P^s \cdot P^t)(i,j) = \sum_{k\in[n]} P^s(i,k)P^t(k,j) \tag{1}$$

若 π_t 是 X_t 的概率分布，则有 $\pi_t^\top P = \pi_{t+1}^\top$。

对于一个连续的时齐马尔可夫过程 $(X_t)_{t\geq 0}$ （我们会考虑的通常是满足随机微分方程 $\mathrm{d}X_t = \mu(X_t)dt + \sigma(X_t)\,\mathrm{d}B_t$）的时齐伊藤过程），我们可以定义它的马尔可夫半群（Markov semigroup）为一族函数 $\mathcal{P} = (P_t)_{t\geq 0}$ 满足

$$P_t(x,A) = \mathbb{P}\left[X_t \in A \mid X_0 = x\right].$$

换句话说，对于每一个固定的 x，$P_t(x,\cdot)$ 是一个概率测度，用来表示从 x 出发，走了 t 时间之后位置的分布。这里每一个 P_t 都被称为一个马尔可夫转移核（Markov transition kernel）。根据全概率公式，我们显然有下面的的查普曼-柯尔莫哥洛夫（Chapman-Kolmogorov）方程

命题 1 (查普曼-柯尔莫哥洛夫方程)**.** *对于任何* $s,t \geq 0$, $P_{s+t}(x,A) = \int P_s(x,\mathrm{d}y)P_t(y,A)$。

作为算子的 P 让我们回到 $P \in [0,1]^{n\times n}$ 是矩阵的场合。刻画 P 的另一种方式是将其视为一个算子，也就是考察它右乘一个函数的效果。令 $f\colon [n] \to \mathbb{R}$ 为一任意函数。我们可以等价地视其为向量 $f = (f(1), \ldots, f(n))^\top \in \mathbb{R}^n$。若我们将 f 右乘于 P，根据定义，我们有

$$[Pf](i) = \sum_{j\in[n]} P(i,j)f(j) = \mathbf{E}\left[f(X_{t+1}) \mid X_t = i\right]. \tag{2}$$

其结果 Pf 是定义在 $[n]$ 上的另一函数，其在输入 i 处的值为 $\mathbf{E}\left[f(X_{t+1}) \mid X_t = i\right]$。此值有一个直观的诠释：$[Pf](i)$ 是当随机游走 P 当前位于 i 时，其下一步位置 j 上所有函数值 $f(j)$ 的平均。

同理，我们有

$$[P^t f](i) = \mathbf{E}\left[f(X_{s+t}) \mid X_s = i\right]. \tag{3}$$

我们也可以从同样的角度来理解我们刚才在连续场合定义的马尔可夫半群 $(P_t)_{t\geq 0}$。对于一个函数 $f\colon \mathbb{R} \to \mathbb{R}$，我们定义

$$P_t f(x) := \int_{\mathbb{R}} f(y) P_t(x, \mathrm{d}y) = \mathbf{E}\left[f(X_t) \mid X_0 = x\right].$$

从这个角度来说，查普曼-柯尔莫哥洛夫方程（命题 1）可以被视为一个算子方程：

$P_s P_t f := P_s(P_t f)$

$$P_{s+t} = P_s P_t.$$

我们考虑的马尔可夫半群和测试函数 f 通常需要满足一些技术性的条件。比如，我们会要求在 $t \to 0$ 时，$\|P_t f - f\|_\infty \to 0$。换句话说 P_t 会收敛到单位算子 I。这类似于离散时间马尔可夫链的转移矩阵 P^t 在 $t = 0$ 时是单位矩阵 I。

1.2 马尔可夫半群的生成元

我们在讲连续时间离散状态的马尔可夫链的时候，我们引入过一个叫做转移速率矩阵的概念，被用来类比于离散时间场合的转移矩阵 $P - I$。我们现在重新定义它，并给予一个更酷名字，叫做*生成元*（*generator*）。

在这门课里，我们不去仔细讨论生成元的定义域。我们可以不妨假设所有关心的 f 都是在无穷远处趋于零的连续函数，或者定义在一个紧集上的连续函数。

定义 2 (马尔可夫半群的生成元)**.** *假设 $\mathcal{P} = (P_t)_{t\geq 0}$ 是一个马尔可夫半群。算子 $\mathcal{L}$ 是 $\mathcal{P}$ 的生成元，当且仅当*

$$\mathcal{L}f = \lim_{t\downarrow 0} \frac{P_t f - f}{t}$$

对于那些 $f \in D(\mathcal{L}) := \{f \mid \text{上述极限存在}\}$ 成立。

我们有 $\mathcal{L} = \frac{\mathrm{d}P_t}{\mathrm{d}t}\Big|_{t=0}$，因此 $\mathcal{L}$ 可以看成 P_t 的变化率。和离散状态类似，P_t 亦可由 $\mathcal{L}$ 描述。

命题 3. 对任意 $t \geq 0$，$P_t = e^{t\mathcal{L}}$ 成立。

注意，算子 A 的指数形式为

$$e^A := \sum_{k\geq 0} \frac{A^k}{k!}$$

证明. 根据定义，我们有

$$\begin{aligned}
\mathcal{L}(P_t f) &= \lim_{s\downarrow 0} \frac{P_s(P_t f) - P_t f}{s} \\
&\overset{\text{(Chapman-Kolmogorov)}}{=} \lim_{s\downarrow 0} \frac{P_{t+s} f - P_t f}{s} \\
&= \left(\lim_{s\downarrow 0} \frac{P_{t+s} - P_t}{s}\right) f = \left(\frac{\mathrm{d}}{\mathrm{d}t} P_t\right) f
\end{aligned}$$

微分方程 $\mathcal{L}P_t = \frac{\mathrm{d}}{\mathrm{d}t}P_t$ 的解是 $P_t = e^{t\mathcal{L}}$。 □

$\mathcal{L}$ 的一个有用性质是，我们可以将 $\mathcal{L}$ 与 P_t 交换次序。

命题 4. 对任意 $t \geq 0$，$\mathcal{L}$ 与 P_t 是可交换的。

证明. 根据定义，我们有

$$\begin{aligned}
\mathcal{L}(P_t f) &= \lim_{s\downarrow 0} \frac{P_s P_t f - P_t f}{s} \\
&\overset{\text{(Chapman-Kolmogorov)}}{=} \lim_{s\downarrow 0} \frac{P_t P_s f - P_t f}{s} \\
&= \lim_{s\downarrow 0} P_t \frac{P_s f - f}{s} \\
&= P_t \lim_{s\downarrow 0} \frac{P_s f - f}{s} = P_t \mathcal{L} f.
\end{aligned}$$

以上第三和第四个等式成立是因为对任意 $s \geq 0$，P_s 是一个连续线性算子。 □

我们现在可以计算时齐伊藤过程的生成元了。

引理 5. 假设 $(X_t)_{t\geq 0}$ 满足

$$\mathrm{d}X_t = \mu(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t$$

则它的生成元为

$$\mathcal{L}f = f'\mu + \frac{1}{2} f''\sigma^2$$

证明. 根据生成元的定义，我们有

$$\begin{aligned}\mathcal{L}f(x) &= \lim_{h\downarrow 0}\frac{\mathbf{E}\left[f(X_h)\mid X=x\right]-f(x)}{h}\\ &= \lim_{h\downarrow 0}\frac{\mathbf{E}\left[f(X_h)-f(X_0)\mid X_0=x\right]}{h}\\ &= \mathbf{E}\left[\left.\frac{\mathrm{d}f(X_t)}{\mathrm{d}t}\right|_{t=0}\middle| X_0=x\right]\end{aligned}$$

我们对 $\mathrm{d}f(X_t)$ 应用伊藤引理，

$$\begin{aligned}\mathcal{L}f(x) &= \frac{1}{\mathrm{d}t}\mathbf{E}\left[\left(f'(x)\mu(x)+\frac{1}{2}f''(x)(\sigma(x))^2\right)\mathrm{d}t+f'(x)\sigma(x)\,\mathrm{d}B_t\right]\\ &= f'(x)\mu(x)+\frac{1}{2}f''(x)(\sigma(x))^2+\frac{f'(x)\sigma(x)}{\mathrm{d}t}\mathbf{E}\left[\mathrm{d}B_t\right].\end{aligned}$$

由于 $\{B_t\}_{t\geq 0}$ 是一个标准布朗运动，$\mathbf{E}\left[\mathrm{d}B_t\right]=0$。这表明 $\mathcal{L}f(x)=f'(x)\mu(x)+\frac{1}{2}f''(x)(\sigma(x))^2$。 □

示例 6 (标准布朗运动的生成元)**.** 标准布朗运动满足 $\mathrm{d}X_t=\mathrm{d}B_t$。因此直接应用引理 5，我们得到

$$\mathcal{L}f(x)=\tfrac{1}{2}\cdot f''(x)$$

1.3 柯尔莫哥洛夫前向与后向方程（Kolmogorov Forward / Backward Equations）

我们在离散马尔可夫链的场合已经接触过科尔莫哥洛夫前向和后向方程，它同样存在于我们现在连续的场合。我们在前面命题 3的证明中实际上已经给出来了

(科尔莫哥洛夫后向方程（KBE）) $$\frac{\mathrm{d}}{\mathrm{d}t}P_t=\mathcal{L}P_t.$$

我们知道 $\mathcal{L}$ 和 P_t 是可交换的（命题 4），因此我们也有

(科尔莫哥洛夫前向方程（KFE）) $$\frac{\mathrm{d}}{\mathrm{d}t}P_t=P_t\mathcal{L}.$$

科尔莫哥洛夫前向方程又被称为福克-普朗克方程（Fokker-Planck equation）。如果我们和有限状态离散时间，即 $P\in\mathbb{R}^{n\times n}$ 为矩阵的场合类比，KBE 和 KFE 分别对应于如下两个平凡的等式：

(KBE) $$P^t-P^{t-1}=(P-I)P^{t-1},$$

(KFE) $$P^t-P^{t-1}=P^{t-1}(P-I).$$

伴随算子（Adjoint Operator） 我们接下来要解释 KFE 和 KBE 为什么分别叫做*前向方程*和*后向方程*。我个人认为，它实际上来源于方程的某种特殊形式。我们先来考虑 KFE，或者 Fokker-Planck 方程。为了解释“前向”的意义，我们先引入 P^*，即 P 的伴随算子。在函数空间 $L^2(\mathbb{R})$ 上，我们可以定义内积 $\langle f, g\rangle = \int_{\mathbb{R}} f(x)g(x)\,\mathrm{d}x$。于是，一个马尔可夫转移核 P 的伴随算子 P^* 定义为满足

$L^2(\mathbb{R})$ 为所有满足 $\int_{\mathbb{R}} f(x)^2\,\mathrm{d}x < \infty$ 的函数的集合。

P^* 的存在性由里斯表示定理保证。

$$\langle Pf, g\rangle = \langle f, P^* g\rangle$$

的算子。

有了这些定义，我们来考虑马尔可夫过程 $(X_t)_{t\geq 0}$。我们用 π_t 表示 X_t 的概率密度函数。于是，对于任何一个测试函数 $f \in L^2(\mathbb{R})$，我们有

$$\mathbf{E}\left[f(X_t)\right] = \int_{\mathbb{R}} f(x)\pi_t(x)\,\mathrm{d}x = \langle f, \pi_t\rangle.$$

我们使用另外一种方式来计算 $\mathbf{E}\left[f(X_t)\right]$：

$$\mathbf{E}\left[f(X_t)\right] = \mathbf{E}\left[\mathbf{E}\left[f(X_t)|X_0\right]\right] = \int_{\mathbb{R}} \mathbf{E}\left[f(X_t)|X_0 = x\right]\pi_0(x)\,\mathrm{d}x = \int_{\mathbb{R}} P_t f(x)\pi_0(x)\,\mathrm{d}x = \langle P_t f, \pi_0\rangle.$$

于是，我们得到

$$\langle f, \pi_t\rangle = \langle P_t f, \pi_0\rangle = \langle f, P_t^*\pi_0\rangle.$$

由于上式对于所有的 $f \in L^2(\mathbb{R})$ 成立，我们有 $\pi_t = P_t^*\pi_0$。可以看到，P_t^* 扮演了有限状态场合 $(P^t)^\top$ 的角色，用来给出从初始分布 π_0 到 t 时刻分布 π_t 的演化。

我们把 KFE 两边均取伴随，便得到 $\frac{\mathrm{d}}{\mathrm{d}t}P_t^* = \mathcal{L}^* P_t^*$。再把他们应用到 π_0 上，便得到了

更一般的，对于任何 $s \geq 0$ 以及固定的 x_s，我们可以用 KFE 得到

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}\left[X_{s+t} = x \mid X_s = x_s\right] = \mathcal{L}^*\,\mathbb{P}\left[X_{s+t} = x \mid X_s = x_s\right].$$

$$\frac{\mathrm{d}}{\mathrm{d}t}P_t^*\pi_0 = \mathcal{L}^* P_t^*\pi_0 \iff \frac{\mathrm{d}}{\mathrm{d}t}\pi_t = \mathcal{L}^*\pi_t.$$

这也许是 KFE 一种更常见的形式，它描述了 $(X_t)_{t\geq 0}$ 在随着时间演化时，其概率分布的变化规律。

我们接着来解释 KBE 称作后向方程的原因。对于任何一个可测集 A，我们取一个测试函数 $f(x) = \mathbb{1}\left[x \in A\right]$，并把它代入 KBE 中便能得到

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}P_t f(x) = P_t\mathcal{L}f(x) &\iff \forall s \geq 0,\ \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\left[\mathbb{1}\left[X_{s+t} \in A\right] \mid X_s = x\right] = \mathcal{L}\,\mathbf{E}\left[\mathbb{1}\left[X_{s+t} \in A\right] \mid X_s = x\right] \\ &\iff \forall s \geq 0,\ \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}\left[X_{s+t} \in A \mid X_s = x\right] = \mathcal{L}\,\mathbb{P}\left[X_{s+t} \in A \mid X_s = x\right].\end{aligned}$$

换句话说，这个方程定义了从 $X_s = x$ 出发，经过时间 t 后，X_{s+t} 落在集合 A 的概率 $\mathbb{P}\left[X_{s+t} \in A \mid X_s = x\right]$ 随时间的演化规律。由于它描述了从初始状态 X_s 的变化对于后续状态 X_{s+t} 到达某个固定集合的规律，因此被称为后向方程。

1.4 使用 KBE/KFE 研究扩散过程的分布演化

考虑时齐的伊藤过程：

$$\mathrm{d}X_t = \mu(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t.$$

我们将研究 X_t 的密度函数 $\pi_t(x)$ 的演化。已知 π_t 满足柯尔莫哥洛夫前向方程 $\frac{\mathrm{d}}{\mathrm{d}t}\pi_t = \mathcal{L}^*\pi_t$。因此我们需要找出算子 $\mathcal{L}^*$。我们在前面已经计算过这个过程的生成元

$$\mathcal{L}f(x) = \mu(x)\cdot\frac{\mathrm{d}}{\mathrm{d}x}f(x) + \frac{1}{2}\sigma^2(x)\cdot\frac{\mathrm{d}^2}{\mathrm{d}x^2}f(x).$$

我们使用定义来计算 $\mathcal{L}^*$。固定一个二次可微的函数 f。一方面，

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\left[f(X_t)\right] = \frac{\mathrm{d}}{\mathrm{d}t}\langle\pi_t, f\rangle = \langle\frac{\mathrm{d}}{\mathrm{d}t}\pi_t, f\rangle = \langle\mathcal{L}^*\pi_t, f\rangle,$$

其中最后一个等式由 KFE 得到。另一方面，

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\left[f(X_t)\right] &= \lim_{s\to 0}\mathbf{E}\left[\frac{f(X_{t+s})-f(X_t)}{s}\right] \\
&= \lim_{s\to 0}\mathbf{E}\left[\mathbf{E}\left[\frac{f(X_{t+s})-f(X_t)}{s}\,\middle|\,X_t\right]\right] \\
\text{(伊藤公式)}\qquad &= \mathbf{E}\left[f'(X_t)\mu(X_t) + \frac{1}{2}f''(X_t)\sigma(X_t)\right],
\end{aligned}$$

利用分部积分公式，我们得到

$$\begin{aligned}
\mathbf{E}\left[f'(X_t)\mu(X_t)\right] &= \int_{\mathbb{R}} f'(x)\mu(x)\pi_t(x)\,\mathrm{d}x \\
&= \mu(x)\pi_t(x)f(x)|_{-\infty}^{\infty} - \int_{\mathbb{R}} f(x)\frac{\partial}{\partial x}\left(\mu(x)\pi_t(x)\right)\mathrm{d}x \\
&= -\langle f, \frac{\partial}{\partial x}(\mu\cdot\pi_t)\rangle.
\end{aligned}$$

其中最后一个不等式成立，是因为 $\pi_t(x)$ 在无穷远处为零。类似地，两次使用分部积分公式，我们得到

$$\begin{aligned}
\mathbf{E}\left[f''(X_t)\sigma^2(X_t)\right] &= \int_{\mathbb{R}} f''(x)\sigma^2(x)\pi_t(x)\,\mathrm{d}x \\
&= -\int_{\mathbb{R}} f'(x)\frac{\partial}{\partial x}\left(\sigma^2(x)\pi_t(x)\right)\mathrm{d}x \\
&= \int_{\mathbb{R}} f(x)\frac{\partial^2}{\partial x^2}\left(\sigma^2(x)\pi_t(x)\right)\mathrm{d}x \\
&= \langle f, \frac{\partial^2}{\partial x^2}\left(\sigma^2\pi_t\right)\rangle.
\end{aligned}$$

实际上，我们可以利用伴随算子的一些法则，如 $(A+B)^* = A^* + B^*$，$(AB)^* = B^*A^*$，$\left(\frac{\partial}{\partial x}\right)^* = -\frac{\partial}{\partial x}$，直接从 $\mathcal{L}$ 得到 $\mathcal{L}^*$。从这个角度出发，我们有

$$\begin{aligned}
\mathcal{L}^*f &= \left(\mu\cdot\frac{\partial}{\partial x}f + \tfrac{1}{2}\sigma^2\cdot\frac{\partial^2}{\partial x^2}f\right)^* \\
&= \left(\mu\cdot\frac{\partial}{\partial x}f\right)^* + \left(\tfrac{1}{2}\sigma^2\cdot\frac{\partial^2}{\partial x^2}f\right)^* \\
&= -\frac{\partial}{\partial x}\mu^*f + \tfrac{1}{2}\frac{\partial^2}{\partial x^2}(\sigma^2)^*f \\
&= -\frac{\partial}{\partial x}(\mu f) + \tfrac{1}{2}\frac{\partial^2}{\partial x^2}\left(\sigma^2 f\right).
\end{aligned}$$

上面的计算中用到 $\mu(x)$ 和 $\sigma^2(x)$ 都是实值函数，可以看成乘上一个实数的算子，因此它们的伴随算子就是它们本身。这个证明和我们使用定义证明没有本质区别，因为伴随算子的这些法则同样需要使用分部积分验证。

因此 $\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{E}\left[f(X_t)\right] = \langle -\frac{\partial}{\partial x}\left(\mu\pi_t\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\sigma^2\pi_t\right), f\rangle$。故而

$$\mathcal{L}^*(g) = -\frac{\partial}{\partial x}(\mu g) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\sigma^2 g\right).$$

因此，我们关心的这个伊藤过程的概率分布 π_t 的演化方程为

$$\frac{\partial}{\partial t}\pi_t(x) = -\frac{\partial}{\partial x}\left(\mu(x)\pi_t(x)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\sigma^2(x)\pi_t(x)\right). \tag{4}$$

作为一个应用，我们假设 $\pi(x)$ 是某伊藤过程的平稳分布，将其代入 (4)，我们得到

$$\frac{\partial}{\partial x}(\mu(x)\pi(x)) = \frac{1}{2}\frac{\partial}{\partial x}\left(\sigma^2(x)\pi(x)\right). \tag{5}$$

现在让我们看几个熟悉的扩散例子。

示例 7 (朗之万动力学)**.** 其随机微分方程为

$$\mathrm{d}X_t = -f'(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t.$$

将上述方程中的 μ 和 σ 代入 (5) 可得

$$\frac{\partial}{\partial x}\left(-f'(x)\pi(x)\right) = \frac{\partial^2}{\partial x^2}\pi(x).$$

消去一个微分算子:

$$-f'(x)\pi(x) = \pi'(x),$$

其解为

$$\pi(x) \propto e^{-f(x)}. \tag{6}$$

$\pi(x) \propto e^{-f(x)}$ 的意思是存在一个（与 x 无关的）常数 C，使得 $\pi(x) = C \cdot e^{-f(x)}$。在这里 $C = \left(\int_{\mathbb{R}} e^{-f(x)}\,\mathrm{d}x\right)^{-1}$，使得 $\pi(x)$ 是一个合法的概率密度函数。

Ornstein-Uhlenbeck 过程是朗之万动力学的一个特例，其平稳分布是高斯分布。

示例 8 (Ornstein-Uhlenbeck 过程)**.** 其随机微分方程为

$$\mathrm{d}X_t = -X_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t.$$

将 $f'(x) = x$（等价于 $f(x) = \frac{x^2}{2}$）代入 6 中的 $f(x)$，我们得到

$$\pi(x) \propto e^{-\frac{x^2}{2}}.$$

2 朗之万动力学的收敛性（Convergence of Langevin Dynamics）

在本节中，我们分析朗之万动力学的收敛速率。令 $f:\mathbb{R}^d\to\mathbb{R}$ 为一个 μ-强凸（strongly convex）函数。大家在凸优化课分析过梯度下降算法，以及它的连续版本梯度流。回顾在分析梯度流时，我们计算了 $\frac{\mathrm{d}\|X_t-Y_t\|^2}{\mathrm{d}t}$，其中 X_t 和 Y_t 分别来自常微分方程 $\mathrm{d}X_t=-\nabla f(X_t)\,\mathrm{d}t$ 和 $\mathrm{d}Y_t=-\nabla f(Y_t)\,\mathrm{d}t$。我们对与郎之万动力学的分析是十分类似的，只需要一个额外耦合的技巧来处理随机项。

我们说一个可微函数 f 是 μ-强凸的，如果其满足对于任何 x,y，

$$f(y)\ge f(x)+\langle\nabla f(x),y-x\rangle+\frac{\mu}{2}\|x-y\|_2^2.$$

除非另有说明，本节的 $\|\cdot\|$ 指的是欧几里得范数。

对于朗之万扩散，令 $\{X_t\}$ 和 $\{Y_t\}$ 为由以下方程生成的两个过程

$$\begin{cases}\mathrm{d}X_t=-\nabla f(X_t)\,\mathrm{d}t+\mathrm{d}B_t\\ \mathrm{d}Y_t=-\nabla f(Y_t)\,\mathrm{d}t+\mathrm{d}B_t\end{cases}. \tag{7}$$

我们让 Y_0 从平稳分布 $\pi\propto e^{-f}$ 中抽取，而 X_0 的分布可以是任意的。我们使用来计算衡量 X_t 和 Y_t 分布之间的 W_2 距离的演化。

对于两个分布 μ 和 ν，令 ζ 为 μ 和 ν 之间的所有耦合的集合。那么 Wasserstein 距离 $W_2(\mu,\nu):=\inf_{\omega\in\zeta}\mathrm{E}_{(X,Y)\sim\omega}\left[\|X-Y\|^2\right]^{\frac{1}{2}}$。

我们用相同的布朗运动来耦合 X_t 和 Y_t。从 (7)式我们有

$$\frac{\mathrm{d}}{\mathrm{d}t}(X_t-Y_t)=\nabla f(Y_t)-\nabla f(X_t).$$

因此，

$$\begin{aligned}\frac{\mathrm{d}\mathbf{E}\left[\|X_t-Y_t\|^2\right]}{\mathrm{d}t}&=2\mathbf{E}\left[\left\langle\frac{\mathrm{d}\,(X_t-Y_t)}{\mathrm{d}t},X_t-Y_t\right\rangle\right]\\&=2\mathbf{E}\left[\langle\nabla f(Y_t)-\nabla f(X_t),X_t-Y_t\rangle\right]\\&\overset{(\heartsuit)}{\le}-2\mu\cdot\mathbf{E}\left[\|X_t-Y_t\|^2\right],\end{aligned} \tag{8}$$

其中 (♡) 源于 f 的强凸性保证了

$$\begin{cases}f(y)-f(x)\ge\langle\nabla f(x),y-x\rangle+\frac{\mu}{2}\|x-y\|_2^2\\ f(x)-f(y)\ge\langle\nabla f(y),x-y\rangle+\frac{\mu}{2}\|x-y\|_2^2\end{cases}.$$

于是 (8) 给出了朗之万动力学的收敛上界

$$\mathbf{E}\left[\|X_t-Y_t\|\right]\le e^{-2\mu t}\cdot\mathbf{E}\left[\|X_0-Y_0\|^2\right].$$

真正在计算机中实现这个算法进行采样时，需要对时间进行离散化（正如同梯度下降算法般）。我们将在作业中分析离散化后的朗之万算法。

3 去噪扩散概率模型（Denoising Diffusion Probabilistic Model）

值得注意的是，上述分析在很大程度上依赖于 π 的强对数凹性（或等价地，f 的强凸性），这在实践中的大多数分布中并不成立。最近一种流行的方法，用于从一般分布 π 中采样，即所谓的*去噪扩散概率模型*（*DDPM*）。

DDPM 的思想是对初始样本 $X_0 \sim \pi$ 添加高斯噪声，直到其分布足够接近高斯分布。然后，**它从高斯分布中抽取一个样本，并执行一个逆向过程**，以生成目标分布的样本。

实际上，就是反过来走马尔可夫链。用离散时间的马尔可夫链的语言来说的话，设 $P(x,y)$ 为一步从 x 到 y 的概率。于是 $\mathbb{P}[X_{t+1}=y \mid X_t=x]=P(x,y)$。那么反过来的链一步从 y 到 x 的转移概率 $\mathbb{P}[X_t=x \mid X_{t+1}=y]=P(x,y)\cdot\frac{\mathbb{P}[X_t=x]}{\mathbb{P}[X_{t+1}=y]}$。注意到，这不再是一个时齐的马尔可夫链，而且每一步转移的概率与正向过程的轨迹有关。

具体来说，在前向过程中，我们从 $X_0 \sim \pi$ 开始运行一个 Ornstein-Uhlenbeck 过程 $(X_t)_{t\geq 0}$。回顾 OU 过程遵循随机微分方程 $\mathrm{d}X_t = -X_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$。因此我们有

$$X_t = e^{-t}X_0 + \tfrac{1}{2}e^{-t}\int_0^t \sqrt{2}\cdot e^s\,\mathrm{d}B_s \overset{D}{=} e^{-t}X_0 + \sqrt{1-e^{-2t}}\cdot\xi. \tag{9}$$

其中 $\xi \sim \mathcal{N}(0,1)$ 是一个标准高斯随机变量。因此，当 T 足够大时，X_T 接近一个高斯随机变量。

设 q_t 为 X_t 的分布。在逆向过程中，我们首先从一个高斯分布中采样 $\overleftarrow{X}_0$，然后根据 OU 过程的逆向方程计算 $(\overleftarrow{X}_t)_{t\in[0,T]}$：

$$\mathrm{d}\overleftarrow{X}_t = \left(\overleftarrow{X}_t + 2\nabla\log q_{T-t}\left(\overleftarrow{X}_t\right)\right)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t. \tag{10}$$

若 $\overleftarrow{X}_0 \sim \mathcal{N}(0,1)$，则 $\overleftarrow{X}_T$ 的分布在 T 足够大时将接近 π。接下来，我们将证明方程式 (10) 确实是 OU 过程的逆过程。

3.1 OU 过程的逆过程的演化方程证明（1-维情况）

我们的目标是研究对于任何 $s \leq t$ 以及 x_s，给出 $\frac{\partial}{\partial t}p(x_t|x_s)$ 的演化方程，把它与 KFE 进行比较，从而得出逆过程演化的方程。我们只研究一维的场合。首先根据条件概率的定义，我们有

我们这里稍微滥用一下记号，用 $p(x_s)$ 表示 $\mathbb{P}\left[\overleftarrow{X}_s = x_s\right]$，用 $p(x_s,x_t)$ 表示 $\mathbb{P}\left[\overleftarrow{X}_s = x_s, \overleftarrow{X}_t = x_t\right]$，用 $p(x_t|x_s)$ 表示 $\mathbb{P}\left[\overleftarrow{X}_t = x_t | \overleftarrow{X}_s = x_s\right]$。虽然这极不严格，数学上有歧义（我们今天的计算中不会出现有歧义的情况），但确实很方便很好用。

$$\begin{aligned}\frac{\partial}{\partial t}p(x_t|x_s) &= p(x_s)^{-1}\cdot\frac{\partial}{\partial t}\left(p(x_s|x_t)\cdot p(x_t)\right)\\ &= p(x_s)^{-1}\cdot\left(\frac{\partial}{\partial t}p(x_s|x_t)\cdot p(x_t) + p(x_s|x_t)\cdot\frac{\partial}{\partial t}p(x_t)\right).\end{aligned} \tag{11}$$

上述两个彩色标出来的偏导数我们可以通过对原过程的 KBE 和 KFE 来计算。KBE 可以给我们提供

注意这里的符号，我们是对 t 求偏导，但同时我们是对原过程使用 KBE，因此要取两次负号。

$$\frac{\partial}{\partial t}p(x_s|x_t) = -f'(x_t)\cdot\frac{\partial}{\partial x_t}p(x_s|x_t) + \frac{\partial^2}{\partial x_t^2}p(x_s|x_t),$$

而 KFE 告诉我们

$$\frac{\partial}{\partial t}p(x_t) = -\frac{\partial}{\partial x_t}\left(f'(x_t)\cdot p(x_t)\right) - \frac{\partial^2}{\partial x_t^2}p(x_t).$$

我们将上述两个偏导数代入 (11)，得到

$$\begin{aligned} p(x_s)\cdot\frac{\partial}{\partial t}p(x_t|x_s) = &\left(-f'(x_t)\cdot\frac{\partial}{\partial x_t}p\left(x_s|x_t\right) + \frac{\partial^2}{\partial x_t^2}p\left(x_s|x_t\right)\right)\cdot p(x_t) \\ &- p(x_s|x_t)\cdot\left(\frac{\partial}{\partial x_t}\left(f'(x_t)\cdot p(x_t)\right) + \frac{\partial^2}{\partial x_t^2}p(x_t)\right). \end{aligned} \tag{12}$$

我们分别对上式两个彩色的项进行展开，可以得到：

$$\frac{\partial}{\partial x_t}p\left(x_s|x_t\right) = \frac{\partial}{\partial x_t}\left(\frac{p(x_s,x_t)}{p(x_t)}\right) = \frac{\frac{\partial}{\partial x_t}p(x_s,x_t)}{p(x_t)} - \frac{p(x_s,x_t)\cdot\frac{\partial}{\partial x_t}p(x_t)}{p(x_t)^2}, \tag{13}$$

以及

$$\frac{\partial}{\partial x_t}\left(f'(x_t)\cdot p(x_t)\right) = f''(x_t)\cdot p(x_t) + f'(x_t)\cdot\frac{\partial}{\partial x_t}p(x_t). \tag{14}$$

把 (13) 和 (14) 代入 (12)，并作一定化简，可以得到

$$\begin{aligned} &p(x_s)\cdot\frac{\partial}{\partial t}p(x_t|x_s) = \\ &\quad p(x_t)\cdot\frac{\partial^2}{\partial x_t^2}p(x_s|x_t) - f'(x_t)\cdot\frac{\partial}{\partial x_t}p(x_s,x_t) - p(x_s|x_t)\cdot\frac{\partial^2}{\partial x_t^2}p(x_t) - f''(x_t)\cdot p(x_s,x_t). \end{aligned} \tag{15}$$

上式绿色的两项的合并为

$$\quad + \quad = \frac{\partial}{\partial x_t}\left(f'(x_t)\cdot p(x_s,x_t)\right). \tag{16}$$

而对于蓝色的两项，其形式出现在两个乘积的二次导的展开之中：

$$\frac{\partial^2}{\partial x_t^2}p(x_s,x_t) = \frac{\partial^2}{\partial x_t^2}\left(p(x_t)\cdot p(x_s|x_t)\right) \tag{17}$$

$$= p(x_t)\cdot\frac{\partial^2}{\partial x_t^2}p(x_s|x_t) + \frac{\partial^2}{\partial x_t^2}p(x_t)\cdot p(x_s|x_t) + 2\frac{\partial}{\partial x_t}p(x_t)\cdot\frac{\partial}{\partial x_t}p(x_s|x_t). \tag{18}$$

我们将 (16) 和 (18) 代入 (15)，并消去第一个蓝色项，可以得到

$$p(x_s)\cdot\frac{\partial}{\partial t}p(x_t|x_s) = \frac{\partial^2}{\partial x_t^2}p(x_s,x_t) - 2\left(\frac{\partial^2}{\partial x_t^2}p(x_t)\cdot p(x_s|x_t) + \frac{\partial}{\partial x_t}p(x_t)\cdot\frac{\partial}{\partial x_t}p(x_s|x_t)\right) - \frac{\partial}{\partial x_t}f'(x_t)\cdot p(x_s,x_t). \tag{19}$$

注意到上式黄色部分可以写成

$$\quad = \frac{\partial}{\partial x_t}\left(\frac{\partial}{\partial x_t}p(x_t)\cdot p(x_s|x_t)\right). \tag{20}$$

我们把 (20) 代入 (19)并进行整理，可以得到

$$\begin{aligned}
p(x_s)\cdot\frac{\partial}{\partial t}p(x_t|x_s) &= \frac{\partial^2}{\partial x_t^2}p(x_s,x_t)-\frac{\partial}{\partial x_t}\left(2\cdot\frac{\partial}{\partial x_t}p(x_t)\cdot p(x_s|x_t)+f'(x_t)\cdot p(x_s,x_t)\right)\\
&= \frac{\partial^2}{\partial x_t^2}p(x_s,x_t)-\frac{\partial}{\partial x_t}\left(p(x_s,x_t)\cdot\left(2\cdot\frac{\frac{\partial}{\partial x_t}p(x_t)}{p(x_t)}+f'(x_t)\right)\right)\\
&= \frac{\partial^2}{\partial x_t^2}p(x_s,x_t)-\frac{\partial}{\partial x_t}\left(p(x_s,x_t)\cdot\left(2\cdot\frac{\partial}{\partial x_t}\log p(x_t)+f'(x_t)\right)\right).
\end{aligned}$$

两边同除 $p(x_s)$，我们得到

$$\frac{\partial}{\partial t}p(x_t|x_s)=-\frac{\partial}{\partial x_t}\left(p(x_t|x_s)\cdot\left(2\cdot\frac{\partial}{\partial x_t}\log p(x_t)+f'(x_t)\right)\right)+\frac{\partial^2}{\partial x_t^2}p(x_t|x_s). \tag{21}$$

和 KFE 的系数进行比较，这便足以给出逆过程的演化方程了。

3.2 分数匹配（Score Matching）

有了逆过程的演化方程，我们只需要模拟它即可。我们从一个高斯变量 $\overleftarrow{X}_0$ 出发，不断的离散化执行 (22)。

$$\mathrm{d}\overleftarrow{X}_t=\left(\overleftarrow{X}_t+2\nabla\log q_{T-t}\left(\overleftarrow{X}_t\right)\right)\mathrm{d}t+\sqrt{2}\,\mathrm{d}B_t. \tag{22}$$

为了执行逆过程，最重要的问题是对于每一个 t，估计 $\nabla\log q_t(x)$ 的值。它又被称为分数函数（score function）。一般来说，这个函数的计算是困难的。但在实际中，我们有一些来自分布 q_t 的样本（我们有一些来自 q_0 的样本，比如 q_0 是所有白色的猫的分布，一些白色的猫的图片便是 q_0 的样本。通过从这些样本出发模拟 OU 过程即可得到 q_t 的样本），因此，我们期待使用这些样本来训练一个神经网络来估计 $\nabla\log q_t(x)$。

我们这个所谓计算困难是基于如下查询模型：用户可以对于任何一个 $x\in\mathbb{R}$，查询 $f(x)$ 或者 $\nabla f(x)$ 的值。可以证明，对于一般的分布 $p(x)\propto e^{f(x)}$，没有办法使用关于维数的多项式次查询来估计 $\nabla\log q_t(x)$ 的值。

下面这一个略感神奇的公式让我们方便的训练这样一个神经网络。

定理 9 (Tweedie 公式)**.** *设 $Y=X+\sqrt{t}\cdot\xi$，其中 $\xi\sim\mathcal{N}(0,\mathrm{Id})$，是一个 d-维标准正态分布。那么*

$$\mathbf{E}\left[X\mid Y=y\right]=y+t\nabla\log p_Y(y),$$

其中 $p_Y(y)$ 是 Y 的概率密度函数。

证明. 我们会在最后一次作业中让大家证明这个定理，并在作业截止后更新此处的证明。 □

我们可根据 (9) 获得一堆 (X_i,Y_i)，其中 X_i 是（一定比例放缩后）q_0 的样本而 Y_i 是由 X_i 平移某个高斯分布之后分布为 q_t 的样本。我们现在的目标是寻找一个函数 F_θ，使得 $F_\theta(y)\approx\mathbf{E}\left[X\mid Y=y\right]$，这样使用 Tweedie 公式我们便能估计 $\nabla\log P_Y(y)$ 的值了。这是一个回归模型，我们可以通过已知的样本来最小化均方误差 $\mathbf{E}\left[\|X-F_\theta(Y)\|^2\right]$。容易验证，最优的参数 θ^* 满足 $F_{\theta^*}(y)=\mathbf{E}\left[X\mid Y=y\right]$。

4 高维推广

我们之前大部分计算都限制在一维的场合。现在我们来推广到高维的情况。【TODO】