

生成式模型基础

在概率分布空间中做计算

陆一平 北京大学北京国际数学研究中心

本讲的问题。如果计算的未知量是一整个概率分布，我们应当怎样表示它、比较它、优化它，并最终从它产生样本？一条贯穿本讲的路径是：把分布表示成粒子，把分布的改变写成演化方程，再把方程中的速度或 score 变成可以从数据学习的函数。

三个可执行的动作 移动位置对应输运，加入随机扰动对应扩散，改变权重或复制与删除对应生灭。Fokker-Planck 方程及其带反应项的扩展把这三个动作写在同一个式子里。扩散模型与 Flow Matching 的训练目标，则告诉我们如何从样本确定方程中的未知系数。

证明怎么读。每个关键结论先解释为什么要做这一步，再指出证明中真正起作用的变换，最后在明确的假设下完成计算。贯穿全文的一维高斯例子用于核对符号、常数和方向；它也展示了一个容易混淆的事实：同一时刻的分布相同，整条随机轨迹仍然可以不同。第 8 节集中证明三种粒子操作与 PDE 的严格对应，并给出有限粒子的精确方程与粒子数误差。

预备知识：条件期望、方差、线性代数、微积分及常微分方程的基本概念。Brownian motion、Itô 积分与 SDE 将在正文中从头介绍。正文给出所用光滑设定下的证明，并在相应位置注明密度、边界、可积性与解的唯一性条件。整理日期：2026 年 10 月 11 日。

目录与符号

符号与约定	3
1 计算对象从一个点变成一个分布	4
2 用散度与 ELBO 确定计算目标	5
3 最优运输：把分布差异写成移动成本	9
4 Wasserstein 梯度流：目标怎样决定动力学	15
5 Brownian motion：平方根尺度如何改变微积分	20
6 SDE 与 Fokker–Planck：从轨道到概率分布	25
7 生灭与权重：带归一化反应的粒子系统	27
8 从粒子到 PDE：三种操作的严格对应	33
9 扩散模型：从加噪路径学习反向动力学	43
10 Flow Matching：先设计分布路径，再回归边缘速度	51
11 把几何、粒子操作与学习目标连起来	56
A 时间离散：误差从哪里来？	61
B 带解答的练习	62
参考文献与原课程资料	64

符号与约定

符号	含义
$\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$	概率测度； $\mathcal{P}_2(\mathbb{R}^d)$ 还要求二阶矩有限。
ρ, π	相对于 Lebesgue 测度的概率密度；只在密度存在时使用这些写法。
$\mu(\varphi), \mathbb{E}_\rho[\varphi]$	$\int \varphi d\mu$ 与 $\int \varphi(x)\rho(x) dx$ 。
$T_\# \mu$	映射 T 把 μ 推前得到的测度，即 $T(X)$ 的分布。
μ^N	粒子经验测度 $N^{-1} \sum_i \delta_{X^i}$ ，或带权版本 $\sum_i w_i \delta_{X^i}$ 。
$\mathcal{F}(\rho), \xi_\rho$	分布上的泛函及其第一变分 $\delta \mathcal{F} / \delta \rho$ 。
v_t, b_t	连续性方程中的速度与 SDE 中的漂移；二者在存在扩散时可以不同。
W_t, Σ_t	标准 Brownian motion 与 SDE 的扩散矩阵。
$a_t = \Sigma_t \Sigma_t^\top$	瞬时协方差矩阵；各向同性情形写成 $a_t = g_t^2 I_d$ 。
$s_t = \nabla \log \rho_t$	分布的 score；只有光滑正密度时才按此逐点定义。
$r_t, \bar{r}_t$	生长率及其分布平均 $\bar{r}_t = \int r_t \rho_t$ 。
$\mathcal{L}_t, \mathcal{L}_t^*$	作用于测试函数的生成元及作用于密度的伴随算子。
$p_t, s = T - t$	扩散模型正向加噪的密度与反向采样的递增时钟。
α_t, σ_t	线性加噪转移的信号系数与条件标准差； σ_t 与矩阵 Σ_t 区分。

约定。 未特别说明时， ∇ 、 $\nabla^\top$ 、 Δ 均对空间变量作用。 $\|A\|_F$ 表示矩阵 Frobenius 范数。空间为 $\mathbb{R}^d$ 时，使用分部积分的密度与系数满足相应的可积性和无穷远衰减条件；使用紧支撑测试函数的弱方程则不需要先对密度求导。在有界区域上，相应公式采用零通量边界；本讲的高斯例子均在整个 $\mathbb{R}$ 上。

与第一讲的联系。 第一讲讨论如何从有限数据估计函数。本讲中要估计的函数将变成速度场或 score；学到函数以后，还要把它积分成一个分布。因此，函数估计误差和数值积分误差将在最后一节重新相遇。

1 计算对象从一个点变成一个分布

1.1 先问：我们能对一个分布做什么？

在普通优化中，未知量是 $x \in \mathbb{R}^d$ ，算法反复更新这个点。在采样、Bayesian 推断和生成式建模中，未知量往往是一整个概率分布 μ 。我们希望计算 $\int \varphi d\mu$ ，希望把 μ 调整得更接近目标分布，最后还希望产生服从目标分布的新样本。

这三件事可以用同一种表示连接起来。取位置 $X^1, \dots, X^N$ 和非负权重 $w_1, \dots, w_N$, $\sum_i w_i = 1$ ，令

$$\mu^N = \sum_{i=1}^N w_i \delta_{X^i}, \quad \mu^N(\varphi) = \sum_{i=1}^N w_i \varphi(X^i). \quad (1.1)$$

于是，分布上的积分变成有限求和；分布的更新变成位置或权重的更新。这里 δ_x 表示全部质量集中在 x 的概率测度，称为 Dirac 测度。

证明直觉 网格方法把空间划成很多小格，存储每一格中的质量。粒子方法把存储位置交给样本本身：哪里有质量，哪里就放粒子。它避免了显式存储整个高维密度网格，但仍需要回答两个问题：有限个粒子对哪些观测量准确？怎样移动这些粒子才会产生正确的目标分布？

两种常见输入方式需要区分。采样问题可能给出一个未归一化密度 $\tilde{\pi}(x) = e^{-V(x)}$ ，但不知道积分 $Z = \int e^{-V(x)} dx$ 。生成式学习则给出 $X_1, \dots, X_n \sim \rho_{\text{data}}$ ，连 $\rho_{\text{data}}(x)$ 的数值也通常不知道。前一种情形可以利用 V 或 ∇V 设计动力学；后一种情形需要从数据学习动力学中的未知函数。

1.2 为什么测试函数是自然的语言？

若 $X \sim \mu$ ，则 $\mu(\varphi) = \mathbb{E}[\varphi(X)]$ 。测试函数 φ 可以是一维投影、平方距离、一个局部的平滑指示函数，或我们真正关心的观测量。研究分布的演化，就可以先研究这些平均量怎样变化。

命题 1.1 (固定观测量的粒子误差)。设 $X^1, \dots, X^N$ 独立同分布于 μ ，且 $\varphi \in L^2(\mu)$ 。对等权经验测度有

$$\mathbb{E}[\mu^N(\varphi)] = \mu(\varphi), \quad \mathbb{E}[|\mu^N(\varphi) - \mu(\varphi)|^2] = \frac{\text{Var}_\mu(\varphi)}{N}. \quad (1.2)$$

因此，对 $\delta \in (0, 1)$ ，以至少 $1 - \delta$ 的概率， $|\mu^N(\varphi) - \mu(\varphi)| \leq \sqrt{\text{Var}_\mu(\varphi)/(N\delta)}$ 。

证明。第一式由期望的线性性得到。记 $Y_i = \varphi(X^i) - \mu(\varphi)$ ；独立性使 $i \neq j$ 时 $\mathbb{E}[Y_i Y_j] = 0$ ，故

$$\mathbb{E} \left[\left(\frac{1}{N} \sum_i Y_i \right)^2 \right] = \frac{1}{N^2} \sum_i \mathbb{E}[Y_i^2] = \frac{\text{Var}_\mu(\varphi)}{N}.$$

最后对平方误差应用 Markov 不等式，即得到 Chebyshev 界。 $\square$

这个 $N^{-1/2}$ 是一个固定观测量的误差尺度。方差常数可能依赖维数；同时控制一整个函数类还需要第一讲中的复杂度工具。如果粒子经过相互作用或重采样，证明中的交叉项通常不再为零。后面写出粒子算法时，我们会把这两点单独说明。

说明 1.2 (经验测度不是一张光滑密度图)。若 π 有 Lebesgue 密度，而 μ^N 只有有限个原子，令 $A = \{X^1, \dots, X^N\}$ ，则 $\pi(A) = 0$ ，而 $\mu^N(A) = 1$ 。所以 μ^N 不绝对连续于 π ， $\text{KL}(\mu^N \parallel \pi) = +\infty$ 。同样， $\nabla \log \mu^N$ 不是一个逐点有定义的 score。要计算这些量，需要一个平滑密度模型、密度估

计，或等价的可采样训练目标。这也解释了为什么后面会用条件 score 的回归，代替对经验分布直接取对数和梯度。

1.3 一个始终可以算完的例子

取简单分布 $Z \sim \mathcal{N}(0, 1)$ ，目标为 $\pi = \mathcal{N}(2, 1/4)$ 。映射 $T(z) = 2 + z/2$ 立即产生正确样本。但在一般问题中，我们只知道目标样本，不能直接写出这样的 T 。可以先把静态映射拆成一串小变换：

$$X_t = m_t + \eta_t Z, \quad m_0 = 0, \eta_0 = 1, m_1 = 2, \eta_1 = 1/2, \quad (1.3)$$

其中 $\eta_t > 0$ 为标准差函数。直接求导并把 $Z = (X_t - m_t)/\eta_t$ 代回，得到

$$\dot{X}_t = v_t(X_t), \quad v_t(x) = \dot{m}_t + \frac{\dot{\eta}_t}{\eta_t}(x - m_t). \quad (1.4)$$

因此，产生样本的问题变成了学习一个随时间变化的速度场。第一项移动均值，第二项伸缩云团。它既是一个可直接检验的 ODE，也是后面理解 Flow Matching 的最小模型。

1.4 三种粒子操作与一个分布方程

对粒子而言，最直接的更新只有几类。我们可以按向量场移动位置，可以加入独立的随机扰动，也可以在位置不变时调整权重，或用复制与删除实现权重的变化。它们在连续极限中对应

$$\partial_t \rho_t = -\nabla \cdot (b_t \rho_t) + \frac{1}{2} \sum_{i,j=1}^d \partial_{ij} (a_{ij,t} \rho_t) + (r_t - \bar{r}_t) \rho_t, \quad \bar{r}_t = \int r_t \rho_t \, dx. \quad (1.5)$$

这里 $a_t = \Sigma_t \Sigma_t^\top$ 是随机扰动的瞬时协方差。当 $r_t = 0$ 时，式 (1.5) 是经典 Fokker-Planck 方程；最后一项是额外的归一化反应项。本节先把它作为全篇要解释的目标公式，后面将逐项推导。

乘以光滑紧支撑函数 φ 并分部积分，得到更便于计算的弱形式：

$$\frac{d}{dt} \mathbb{E}_{\rho_t}[\varphi] = \mathbb{E}_{\rho_t} \left[b_t \cdot \nabla \varphi + \frac{1}{2} \text{Tr}(a_t \nabla^2 \varphi) \right] + \text{Cov}_{\rho_t}(r_t, \varphi). \quad (1.6)$$

输运读取观测量的一阶变化，扩散读取二阶变化，生灭读取观测量与生长率的相关性。对于常数观测量 $\varphi = 1$ ，三个贡献都为零，因此总质量保持为 1。在矩条件允许时取 $\varphi(x) = x$ ，可以读出均值变化；取 $\varphi(x) = xx^\top$ ，就能读出扩散对协方差的直接贡献。

后面各部分各解决一个问题。散度与 ELBO 说明我们想把分布变成什么样；最优运输与梯度流说明如何选择运动方向；Itô 公式把随机粒子更新转成分布方程；扩散模型与 Flow Matching 则把方程中的未知系数变成可从数据估计的函数。从始至终，算法只操作粒子或网络，分布方程负责解释它究竟算出了什么。

这些内容与原课程的最优运输及粒子系统讲义相衔接[1, 2]。接下来先为“接近目标分布”建立数学含义。

2 用散度与 ELBO 确定计算目标

当未知量是一个概率分布时，首先要回答的是：怎样判断当前分布离目标还有多远？本节先把这个问题写成一个标量泛函，再说明如何把其中难算的积分改写成可以采样估计的目标。这一步确

定了想优化什么；后面引入最优运输，才会确定允许怎样改变分布。原课程的最优运输与分布空间计算材料是这一组织方式的起点 [1, 2]。

2.1 密度比与 f 散度

设 μ, ν 是同一可测空间上的概率测度，且 $\mu \ll \nu$ 。记密度比为 $r = d\mu/d\nu$ ，于是

$$\int r \, d\nu = 1, \quad \mathbb{E}_\mu h = \mathbb{E}_\nu[rh].$$

这里比较的是两个概率测度。若它们在 $\mathbb{R}^d$ 上分别有密度 ρ, π ，那么 $r(x) = \rho(x)/\pi(x)$ ，以下公式便可以写成通常的积分。

定义 2.1 (f 散度)。设 $f: [0, \infty) \rightarrow \mathbb{R} \cup \{+\infty\}$ 凸且 $f(1) = 0$ 。在 $\mu \ll \nu$ 时定义

$$D_f(\mu\|\nu) = \int f\left(\frac{d\mu}{d\nu}\right) d\nu. \quad (2.1)$$

以下有限值计算均假设相应积分存在。密度形式是 $D_f(\rho\|\pi) = \int \pi(x)f(\rho(x)/\pi(x)) \, dx$ 。

几个常见的选择如下。表中总变差采用 $\|\mu - \nu\|_{\text{TV}} = \sup_A |\mu(A) - \nu(A)|$ 的约定。

选择	$f(u)$	相应散度
相对熵	$u \log u$	$\text{KL}(\mu\ \nu) = \mathbb{E}_\mu \log r$
反向相对熵	$-\log u$	$\text{KL}(\nu\ \mu)$ ，在互相绝对连续时
Pearson 散度	$(u - 1)^2$	$\mathbb{E}_\nu(r - 1)^2$
总变差	$\frac{1}{2} u - 1 $	$\ \mu - \nu\ _{\text{TV}}$

约定 $0 \log 0 = 0$ 。对 KL，若 $\mu \not\ll \nu$ ，定义 $\text{KL}(\mu\|\nu) = +\infty$ 。一般 f 散度也可以计入奇异部分来扩展定义，本节的密度比推导只使用绝对连续的情形。

命题 2.2 (非负性来自凸性)。有 $D_f(\mu\|\nu) \geq 0$ 。若 f 严格凸且散度有限，则等号当且仅当 $\mu = \nu$ 。

证明直觉 在参考测度 ν 下，密度比 r 的平均值必须等于 1。凸函数会惩罚密度比围绕 1 的波动，因此把这些波动平均以后，不会优于常数 $r = 1$ 。

证明. Jensen 不等式给出

$$\int f(r) \, d\nu \geq f\left(\int r \, d\nu\right) = f(1) = 0.$$

严格凸时，Jensen 等号要求 r 在 ν 下几乎处处为常数；归一化使该常数只能为 1，等价于 $\mu = \nu$ 。□

密度比刻画的是相对质量，而不是空间中的移动距离。例如把两个分布同时通过一个双射重新编号，密度比在对应点的值不变，上述散度也不变。到了最优运输部分，我们会额外指定“把一单位质量从 x 移到 y 要付多少钱”，从而把空间几何纳入比较。

2.2 Fenchel 变分表达：把密度比换成可估计的期望

直接计算 D_f 需要密度比。有样本而没有密度公式时，这通常很困难。凸对偶给出了另一种描述：用一个测试函数 T 区分两个分布，再通过一个惩罚项限制 T 的规模。

将 f 在负半轴上扩展为 $+\infty$ ，定义其凸共轭

$$f^*(t) = \sup_{u \geq 0} \{tu - f(u)\}.$$

由定义立即得到 Fenchel 不等式 $tu \leq f(u) + f^*(t)$ 。若 f 在 $u > 0$ 可微，则选择 $t = f'(u)$ 可以取到等号：这正是凸函数在 u 处的支撑直线。

定理 2.3 (一个条件明确的散度变分公式)。设 f 为下半连续凸函数，在 $(0, \infty)$ 上有限且可微，并且密度比满足 $0 < a \leq r \leq b < \infty$ ， ν 几乎处处成立。则

$$D_f(\mu \parallel \nu) = \sup_T \{\mathbb{E}_\mu T - \mathbb{E}_\nu f^*(T)\}. \quad (2.2)$$

这里上确界取遍满足 $\mathbb{E}_\nu |f^*(T)| < \infty$ 的有界可测函数。最优测试函数可以取为 $T^* = f'(r)$ 。

证明直觉 对每一个可能的密度比，都用一条支撑直线从下方逼近 f 。固定测试函数只能给出下界；如果能让每个位置都选中正确的切线，就恰好恢复原来的散度。整个证明只需把这个逐点事实积分。

证明. 对任意允许的 T ，逐点应用 Fenchel 不等式：

$$rT - f^*(T) \leq f(r).$$

对 ν 积分，并使用 $\mathbb{E}_\nu[rT] = \mathbb{E}_\mu T$ ，得到式 (2.2) 的右边不超过左边。另一方面，凸可微函数的导数在紧区间 $[a, b] \subset (0, \infty)$ 上有界，所以 $T^* = f'(r)$ 有界。支撑直线取等号给出

$$f^*(f'(r)) = rf'(r) - f(r).$$

右边在 $r \in [a, b]$ 时有界，因此 T^* 确实属于允许的函数类。代入后逐点等于 $f(r)$ ，积分便得到反向不等式。□

说明 2.4 (函数类限制与适用范围)。密度比有上下界是这里保证取等号与可积性的充分条件。对于一般密度比，Fenchel 下界仍成立；要宣称上确界等于散度，还需检验最优测试函数的可积性，或补充适当的截断逼近论证。实际计算若把 T 限制在一个神经网络类中，得到的是该类给出的变分下界；有限样本优化还会引入额外误差。

以 $f(u) = u \log u$ 为例，计算一元最大值可得

$$f^*(t) = e^{t-1}, \quad T^*(x) = 1 + \log r(x).$$

于是 KL 具有 $\sup_T \{\mathbb{E}_\mu T - \mathbb{E}_\nu e^{T-1}\}$ 的表达。再把 $T = h + c$ 中的常数 c 单独优化：

$$\sup_{c \in \mathbb{R}} \left\{ \mathbb{E}_\mu h + c - e^{c-1} \mathbb{E}_\nu e^h \right\} = \mathbb{E}_\mu h - \log \mathbb{E}_\nu e^h.$$

在定理的密度比条件下， $h = \log r$ 合法且取等号，从而

$$\text{KL}(\mu \parallel \nu) = \sup_{h \text{ 有界可测}} \left\{ \mathbb{E}_\mu h - \log \mathbb{E}_\nu e^h \right\}. \quad (2.3)$$

这一表达把“计算一个密度比”变成“计算两个分布下的期望”。

2.3 ELBO: 难算的证据与后验误差之间的恒等式

设潜变量模型的联合密度为

$$p_\theta(x, z) = p_\theta(z)p_\theta(x | z), \quad p_\theta(x) = \int p_\theta(x, z) dz.$$

给定观测 x , 假设 $0 < p_\theta(x) < \infty$ 。困难在于后验 $p_\theta(z | x) = p_\theta(x, z)/p_\theta(x)$ 含有一个高维积分。我们引入一个便于采样和计算密度的概率密度 $q(z | x)$, 用它近似后验。下列恒等式是变分推断与 VAE 的基本出发点 [5]。

定义 2.5 (证据下界)。 固定 x, θ , 定义

$$\mathcal{L}(\theta, q; x) = \int q(z | x) \log \frac{p_\theta(x, z)}{q(z | x)} dz. \quad (2.4)$$

当它有限时, 称为 evidence lower bound, 简称 ELBO。模型分解后还可以写为

$$\mathcal{L}(\theta, q; x) = \mathbb{E}_q[\log p_\theta(x | Z)] - \text{KL}(q(\cdot | x) \| p_\theta(\cdot)), \quad (2.5)$$

其中假设拆开的两项均有定义且有限。

定理 2.6 (ELBO 与后验 KL 的精确分解)。 若 $q(\cdot | x) \ll p_\theta(\cdot | x)$, 且下式中的 KL 有限, 则

$$\log p_\theta(x) = \mathcal{L}(\theta, q; x) + \text{KL}(q(\cdot | x) \| p_\theta(\cdot | x)). \quad (2.6)$$

因此 $\mathcal{L}(\theta, q; x) \leq \log p_\theta(x)$, 且等号当且仅当 $q(\cdot | x) = p_\theta(\cdot | x)$ 几乎处处成立。

证明直觉 不知道归一化常数, 并不妨碍写出“近似后验除以真后验”。对这个比值取对数, 未知的证据就变成一个与 z 无关的加法常数。积分后, 剩下的两项正好是可优化的 ELBO 和非负的后验误差。

证明. 在 q 有质量的位置, Bayes 公式给出

$$\log \frac{q(z | x)}{p_\theta(z | x)} = \log \frac{q(z | x)}{p_\theta(x, z)} + \log p_\theta(x).$$

对 $q(z | x) dz$ 积分。由于 $\int q dz = 1$, 最后一项仍为 $\log p_\theta(x)$, 第一项则为 $-\mathcal{L}(\theta, q; x)$ 。这证明恒等式。KL 的非负性与取等条件给出其余结论。式 (2.5) 由 $\log p_\theta(x, z) = \log p_\theta(x | z) + \log p_\theta(z)$ 直接代入得到。□

固定模型参数 θ 时, 提高 ELBO 等价于减少后验 KL。若同时更新 θ , 证据和后验都会改变, 应使用完整的恒等式理解训练。即使神经网络参数优化得很好, 受限的 q 类也未必包含真实后验。

例 2.7 (高斯重参数化把随机性固定下来)。 设

$$q_\eta(z | x) = \mathcal{N}(m_\eta(x), \text{diag}(\sigma_{\eta,1}^2(x), \dots, \sigma_{\eta,k}^2(x))), \quad \sigma_{\eta,j}(x) > 0.$$

令 $\epsilon \sim \mathcal{N}(0, \text{Id}_k)$, 则

$$Z = m_\eta(x) + \sigma_\eta(x) \odot \epsilon \implies Z \sim q_\eta(\cdot | x),$$

其中 $\odot$ 表示逐坐标相乘。对可积函数 H_η ,

$$\mathbb{E}_{q_\eta} H_\eta(Z) = \mathbb{E}_\epsilon H_\eta(m_\eta(x) + \sigma_\eta(x) \odot \epsilon).$$

如果右侧关于 η 的导数在参数邻域内有可积的共同支配函数, 就可以把求导移入期望, 得到可由独立 ϵ 样本估计的梯度。这里固定的是噪声的分布; 参数通过一个可微映射改变样本。

若先验为 $\mathcal{N}(0, \text{Id}_k)$, KL 项还可精确算出:

$$\text{KL}(q_\eta(\cdot | x) \| \mathcal{N}(0, \text{Id}_k)) = \frac{1}{2} \sum_{j=1}^k (m_{\eta,j}^2 + \sigma_{\eta,j}^2 - 1 - \log \sigma_{\eta,j}^2). \quad (2.7)$$

证明只需把两个高斯对数密度相减, 再用 $\mathbb{E}(Z_j - m_{\eta,j})^2 = \sigma_{\eta,j}^2$ 与 $\mathbb{E}Z_j^2 = m_{\eta,j}^2 + \sigma_{\eta,j}^2$ 。ELBO 的 Monte Carlo 估计在期望意义下等于 ELBO, 单次抽样得到的数值不必仍是证据的下界。

ELBO 是关于一个概率分布的**静态目标**。它本身没有时间变量, 也没有指定粒子如何运动。要把“提高 ELBO”变成一个演化方程, 还需要选择概率空间上的几何; 第 4 节将把这个选择具体做出来。

3 最优运输: 把分布差异写成移动成本

密度比只比较相同位置上的质量。最优运输进一步允许质量在不同位置之间移动, 并且明确计算这种移动的代价。原课程先从运输计划与对偶讨论这一点, 再通过粒子轨迹连接连续性方程 [1, 2]。

3.1 从一个映射到一份运输计划

设 $c(x, y)$ 是从 x 移动到 y 的成本。Monge 问题寻找一个映射:

$$\inf_{T: T_{\#}\mu = \nu} \int c(x, T(x)) \, d\mu(x). \quad (3.1)$$

约束 $T_{\#}\mu = \nu$ 表示: 若 $X \sim \mu$, 则 $T(X) \sim \nu$ 。它要求每个出发点的质量全部送到一个终点。有时这样的映射根本不存在。例如 $\mu = \delta_0$ 与 $\nu = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$: 单个出发点通过确定性映射, 无法变成两个不同的点。

定义 3.1 (耦合与 Kantorovich 问题)。记 $\Gamma(\mu, \nu)$ 为所有边缘分别是 μ, ν 的联合概率测度。元素 $\gamma \in \Gamma(\mu, \nu)$ 称为一个耦合或运输计划。Kantorovich 运输成本是

$$\text{OT}_c(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \int c(x, y) \, d\gamma(x, y). \quad (3.2)$$

当 $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, 即它们都有有限二阶矩时, 定义

$$W_2^2(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \int \|x - y\|^2 \, d\gamma(x, y). \quad (3.3)$$

本讲的二次成本是 $\|x - y\|^2$, 不额外带 $\frac{1}{2}$ 。

任意映射 T 都产生耦合 $(\text{Id}, T)_{\#}\mu$, 所以运输计划包含了映射这一特例。同时, 独立耦合 $\mu \otimes \nu$ 总是存在。放宽成联合分布以后, 原来可能不可行的确定性搬运问题成为一个线性优化问题。

3.2 有限支持强对偶：用价格给每份计划估值

设

$$\mu = \sum_{i=1}^m \alpha_i \delta_{x_i}, \quad \nu = \sum_{j=1}^n \beta_j \delta_{y_j}, \quad c_{ij} = c(x_i, y_j).$$

设 $\alpha_i, \beta_j \geq 0$ ，且两组权重都和为 1；允许零权重，便于稍后做网格离散化。运输计划就是非负矩阵 $P = (P_{ij})$ ，其中 P_{ij} 表示从 x_i 运到 y_j 的质量。两组边缘约束是

$$\sum_j P_{ij} = \alpha_i, \quad \sum_i P_{ij} = \beta_j.$$

给出发点 and 终点分别标价 ϕ_i, ψ_j 。若每条路线都满足 $\phi_i + \psi_j \leq c_{ij}$ ，则这些价格对任何可行运输计划都会给出一个下界。

定理 3.2 (有限运输问题的强对偶)。若 $c_{ij} \in \mathbb{R}$ ，则

$$\min_{\substack{P_{ij} \geq 0 \\ \sum_j P_{ij} = \alpha_i \\ \sum_i P_{ij} = \beta_j}} \sum_{i,j} c_{ij} P_{ij} = \max_{\phi_i + \psi_j \leq c_{ij}} \left(\sum_i \alpha_i \phi_i + \sum_j \beta_j \psi_j \right). \quad (3.4)$$

证明直觉 弱对偶只需沿每条路线检查价格没有超过成本。强对偶则回答：能否找到一套价格，把最优运输的成本完全解释出来？有限维分离定理给出答案。如果声称某个低于最优值的费用可行，它会落在“所有可实现边缘和费用”的凸集合之外；把它分离出去的超平面，经过归一化就是所需的价格。

证明。先证弱对偶。对任意可行 P, ϕ, ψ ,

$$\sum_{i,j} c_{ij} P_{ij} \geq \sum_{i,j} (\phi_i + \psi_j) P_{ij} = \sum_i \alpha_i \phi_i + \sum_j \beta_j \psi_j.$$

原问题有可行点 $P_{ij} = \alpha_i \beta_j$ ，且可行集合闭、有界，因此最小值 C_* 存在。

为证明相反方向，在 $\mathbb{R}^{m+n+1}$ 中考虑闭凸锥

$$\mathcal{K} = \left\{ (r, s, z) : \begin{array}{l} P_{ij} \geq 0, \quad r_i = \sum_j P_{ij}, \quad s_j = \sum_i P_{ij}, \\ z \geq \sum_{i,j} c_{ij} P_{ij} \quad \text{对某个 } P \text{ 成立} \end{array} \right\}.$$

它的闭性也可以直接验证：若 (r, s, z) 的一列元素收敛，则相应非负矩阵的总质量 $\sum_i r_i$ 有界，故可取收敛子列，极限矩阵仍满足所有约束。

对任意 $z_0 < C_*$ ，点 (α, β, z_0) 不在 $\mathcal{K}$ 中。有限维闭凸锥的分离定理给出 (A, B, λ) ，使

$$A \cdot r + B \cdot s + \lambda z \geq 0 \quad ((r, s, z) \in \mathcal{K}), \quad A \cdot \alpha + B \cdot \beta + \lambda z_0 < 0.$$

因为 $(0, 0, z) \in \mathcal{K}$ 对所有 $z \geq 0$ 成立，所以 $\lambda \geq 0$ 。若 $\lambda = 0$ ，将一个可行运输矩阵代入第一式便得到 $A \cdot \alpha + B \cdot \beta \geq 0$ ，与第二式矛盾。故 $\lambda > 0$ 。对只在 (i, j) 放一单位质量的矩阵应用分离不等式，得到

$$A_i + B_j + \lambda c_{ij} \geq 0.$$

令 $\phi_i = -A_i/\lambda$ ， $\psi_j = -B_j/\lambda$ ，便有 $\phi_i + \psi_j \leq c_{ij}$ ，并且

$$\sum_i \alpha_i \phi_i + \sum_j \beta_j \psi_j > z_0.$$

让 $z_0 \uparrow C_*$, 证明了两个最优值相等。

最后验证右边的上确界可以取到。给定任意可行价格, 依次令

$$\psi'_j = \min_i (c_{ij} - \phi_i), \quad \phi'_i = \min_j (c_{ij} - \psi'_j).$$

每次更新都保留可行性, 并且不会降低目标值; 还满足 $\psi'_j = \min_i (c_{ij} - \phi'_i)$ 。利用加减同一常数的不变性, 将 ϕ'_1 归一化为 0。记 $M = \max_{ij} |c_{ij}|$, 则

$$|\phi'_i - \phi'_k| \leq \max_j |c_{ij} - c_{kj}| \leq 2M, \quad |\psi'_j| \leq 3M.$$

因此可以在一个固定紧盒子内选择最优序列。取收敛子列, 约束与目标值在极限中保留, 最大值遂取到。□

互补条件也很直观: 若 $P_{ij}^* > 0$, 最优价格必满足 $\phi_i^* + \psi_j^* = c_{ij}$ 。否则在该路线留下的严格空隙会使弱对偶积分严格, 违背最优值相等。因此真正使用的路线, 恰好是价格接触成本的位置。

3.3 紧支撑连续对偶: 有限网格怎样逼近原问题

有限维结论不能仅靠把求和符号改成积分就自动推广。下面给出一个足够清楚、也足够常用的连续版本 [3]。

定理 3.3 (紧空间、连续成本的对偶)。设 $K, L \subset \mathbb{R}^d$ 为紧集, μ, ν 分别是其上的概率测度, 且 $c \in C(K \times L)$ 。则

$$\inf_{\gamma \in \Gamma(\mu, \nu)} \int c \, d\gamma = \sup_{\substack{\phi \in C(K), \psi \in C(L) \\ \phi(x) + \psi(y) \leq c(x, y)}} \left\{ \int \phi \, d\mu + \int \psi \, d\nu \right\}. \quad (3.5)$$

证明直觉 把每个小格子里的质量搬到代表点, 连续成本只会改变一个一致可控的小量。有限问题已经有强对偶。剩下的工作有两件: 把有限运输计划还原为连续计划, 把格点价格延伸成处处合法的连续价格。两种还原的误差都随格子变细而消失。

证明. 弱对偶仍由积分 $\phi(x) + \psi(y) \leq c(x, y)$ 得到。取有限 Borel 分割 $K = \bigsqcup_i A_i$, $L = \bigsqcup_j B_j$, 各格直径不超过 δ , 代表点为 $x_i \in A_i, y_j \in B_j$ 。记对应质量为 α_i, β_j , 格点最优成本为 C_δ 。定义一致连续模

$$\omega(\delta) = \sup_{\substack{|x-x'| \leq \delta, |y-y'| \leq \delta \\ x, x' \in K, y, y' \in L}} |c(x, y) - c(x', y')|.$$

紧性保证 $\omega(\delta) \rightarrow 0$ 。

任意连续耦合在格子 $A_i \times B_j$ 上的质量组成一个可行矩阵, 成本变化至多 $\omega(\delta)$ 。反过来, 任意可行矩阵可以还原成

$$\gamma = \sum_{\alpha_i \beta_j > 0} P_{ij} \frac{\mu|_{A_i}}{\alpha_i} \otimes \frac{\nu|_{B_j}}{\beta_j}.$$

它具有正确边缘, 成本变化也至多 $\omega(\delta)$ 。若记连续问题的下确界为 C_* , 就得到

$$|C_\delta - C_*| \leq \omega(\delta).$$

取格点问题的最优对偶价格 ϕ_i, ψ_j ，在格上定义阶梯函数

$$\phi_0(x) = \phi_i - \omega(\delta) \quad (x \in A_i), \quad \psi_0(y) = \psi_j \quad (y \in B_j).$$

由成本误差界可知 $\phi_0(x) + \psi_0(y) \leq c(x, y)$ 。阶梯函数还不连续，但可以用成本本身将它们平滑地包起来：

$$\phi(x) = \inf_{y \in L} \{c(x, y) - \psi_0(y)\}, \quad \psi(y) = \inf_{x \in K} \{c(x, y) - \phi(x)\}.$$

由 c 的一致连续性，两函数均连续；例如

$$|\phi(x) - \phi(x')| \leq \sup_{y \in L} |c(x, y) - c(x', y)|.$$

定义还给出 $\phi \geq \phi_0, \psi \geq \psi_0$ ，且 $\phi + \psi \leq c$ 。因此连续对偶值至少为

$$C_\delta - \omega(\delta) \geq C_* - 2\omega(\delta).$$

令 $\delta \downarrow 0$ ，结合弱对偶，即得结论。 $\square$

这一定理的范围是紧支撑与连续成本。非紧空间、无界成本以及一般弱正则性下的对偶需要进一步的紧性和可积性论证，本讲不将上面的有限网格证明直接当作那些结论的证明。

3.4 一维高斯：最优耦合就是共用同一份随机性

命题 3.4 (一维高斯的二次运输距离)。设 $\mu_0 = \mathcal{N}(m_0, \sigma_0^2)$ 、 $\mu_1 = \mathcal{N}(m_1, \sigma_1^2)$ ，其中 $\sigma_0, \sigma_1 > 0$ 。则

$$W_2^2(\mu_0, \mu_1) = (m_1 - m_0)^2 + (\sigma_1 - \sigma_0)^2, \quad (3.6)$$

且最优映射为

$$T(x) = m_1 + \frac{\sigma_1}{\sigma_0}(x - m_0).$$

证明直觉 均值差是无法省去的整体平移。去掉均值后，想让两个随机变量尽可能靠近，就应让它们的涨落尽可能同步。共用同一个标准高斯恰好达到协方差的最大值。

证明. 对任意耦合 (X, Y) ，Cauchy-Schwarz 不等式给出 $\text{Cov}(X, Y) \leq \sigma_0 \sigma_1$ ，故

$$\mathbb{E}(X - Y)^2 = (m_0 - m_1)^2 + \sigma_0^2 + \sigma_1^2 - 2\text{Cov}(X, Y) \geq (m_0 - m_1)^2 + (\sigma_0 - \sigma_1)^2.$$

令 $G \sim \mathcal{N}(0, 1)$ ，取 $X = m_0 + \sigma_0 G$ ， $Y = m_1 + \sigma_1 G$ ，便达到下界。解出 $G = (X - m_0)/\sigma_0$ ，即得到所述映射。 $\square$

例如两个方差均为 σ^2 、均值相差 a 的高斯，

$$W_2 = |a|, \quad \text{KL}(\mathcal{N}(0, \sigma^2) \parallel \mathcal{N}(a, \sigma^2)) = \frac{a^2}{2\sigma^2}.$$

后一个式子由相减两个对数密度并取期望得到。当两个高斯越来越窄时，相同的空间位移仍然是 $|a|$ ，但相对质量失配会变得越来越严重。这给出了选择分布差异度量时的一个具体判断标准。

3.5 粒子轨迹与连续性方程

最优运输指定起点与终点的对应关系，生成算法还需要实际走过一条路径。给定速度场 $v_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ ，考虑

$$\dot{X}_t = v_t(X_t), \quad X_0 \sim \mu_0.$$

随机性可以完全来自初始点；给定初始点后，轨迹由常微分方程决定。记 $\Phi_t(x)$ 为从 x 出发的解，即

$$\partial_t \Phi_t(x) = v_t(\Phi_t(x)), \quad \Phi_0(x) = x.$$

粒子的分布便是 $\mu_t = (\Phi_t)_\# \mu_0$ 。

定理 3.5 (光滑流与连续性方程)。设 $v_t(x)$ 连续，关于 x 为 C^1 ， $D_x v_t(x)$ 关于 (t, x) 连续，并满足

$$\sup_{t \in [0, T], x \in \mathbb{R}^d} \|D_x v_t(x)\| < \infty, \quad \sup_{t \in [0, T]} |v_t(0)| < \infty.$$

这些全局 Lipschitz 与线性增长条件保证 $[0, T]$ 上存在唯一的全局流。对任意 $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ ，令 $\mu_t = (\Phi_t)_\# \mu_0$ 。则对每个 $\varphi \in C_c^1(\mathbb{R}^d)$ ，

$$\frac{d}{dt} \int \varphi(x) d\mu_t(x) = \int \nabla \varphi(x) \cdot v_t(x) d\mu_t(x). \quad (3.7)$$

更一般的时空测试函数满足

$$\int \varphi(T, x) d\mu_T - \int \varphi(0, x) d\mu_0 = \int_0^T \int (\partial_t \varphi + \nabla \varphi \cdot v_t) d\mu_t dt. \quad (3.8)$$

其中 φ 关于时空为 C^1 ，空间支撑在一个固定紧集内。在此弱解意义下，给定 μ_0 后解唯一。如果 μ_t 有密度 ρ_t ，则其密度满足

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0 \quad (3.9)$$

的分布形式；进一步光滑时即为经典偏微分方程。

证明直觉 不要直接追问每个位置的密度怎样求导。先挑一个光滑的观测函数 φ ，沿每条粒子轨迹用普通链式法则，再对初始点取平均。这样得到的是一个可直接验证的积分恒等式。把导数从测试函数移到质量流上，连续性方程便自然出现。

证明. 由推前定义，

$$\int \varphi d\mu_t = \int \varphi(\Phi_t(x)) d\mu_0(x).$$

链式法则给出

$$\frac{d}{dt} \varphi(\Phi_t(x)) = \nabla \varphi(\Phi_t(x)) \cdot v_t(\Phi_t(x)).$$

当 $\nabla \varphi \neq 0$ 时，其自变量落在一个固定紧集，而 v 在该紧集与 $[0, T]$ 上有界。因此右边具有与初始点无关的可积上界，可以交换求导与积分。再使用推前公式一次，就得到 (3.7)。对 $\varphi(t, \Phi_t(x))$ 使用同一论证，并在时间上积分，得到 (3.8)。若存在密度，空间分部积分给出

$$\int \nabla \varphi \cdot v_t \rho_t dx = - \int \varphi \nabla \cdot (\rho_t v_t) dx,$$

正是 (3.9) 的弱形式。

最后证唯一性。固定终点时间 $\tau \leq T$ 和 $h \in C_c^1(\mathbb{R}^d)$, 记从时刻 t 走到 τ 的流为 $\Phi_{t,\tau}$, 定义

$$\varphi(t, x) = h(\Phi_{t,\tau}(x)).$$

沿轨迹观察这个量, 它始终等于终点的 h 值, 所以 $\partial_t \varphi + v_t \cdot \nabla \varphi = 0$ 。全局流的增长控制保证这一测试函数具有统一的紧空间支撑。代入任意弱解的 (3.8), 得到

$$\int h \, d\mu_\tau = \int h(\Phi_{0,\tau}(x)) \, d\mu_0(x).$$

对所有这样的 h 成立, 唯一确定 $\mu_\tau = (\Phi_\tau)_\# \mu_0$ 。□

推论 3.6 (密度的 Jacobian 公式)。若 μ_0 有密度 ρ_0 , 则

$$\rho_t(\Phi_t(x)) \det D\Phi_t(x) = \rho_0(x) \quad (3.10)$$

几乎处处成立。在 $\rho_0(x) > 0$ 的轨迹上,

$$\log \rho_t(\Phi_t(x)) = \log \rho_0(x) - \int_0^t \nabla \cdot v_s(\Phi_s(x)) \, ds. \quad (3.11)$$

证明直觉 一个小体积单元被流拉大以后, 同一份质量会摊得更薄。Jacobian 记录体积改变, 散度记录它的瞬时相对增长率。

证明. 全局流是 C^1 微分同胚, 变量代换即得 (3.10)。矩阵 $J_t = D\Phi_t(x)$ 满足 $\dot{J}_t = Dv_t(\Phi_t(x))J_t$, $J_0 = \text{Id}$ 。Jacobi 行列式公式给出

$$\frac{d}{dt} \log \det J_t = \text{Tr}(J_t^{-1} Dv_t(\Phi_t(x))J_t) = \nabla \cdot v_t(\Phi_t(x)).$$

初始行列式为 1, 故之后始终为正。对 (3.10) 取对数并积分, 得到第二式。□

3.6 Benamou–Brenier: 运输距离是最小动能

粒子移动也给出一个分布路径的长度标准。若 (μ_t, v_t) 满足连续性方程, 定义总动能

$$\mathcal{A}(\mu, v) = \int_0^1 \int \|v_t(x)\|^2 \, d\mu_t(x) \, dt.$$

以下给出一个光滑版本的完整证明。它把“起终点的最优对应”与“在中间时刻怎样移动”连接起来 [2, 3]。

定理 3.7 (光滑可实现情形的动态运输公式)。设 $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, 并假设存在一个二次成本的最优映射 T , 使 $T_\# \mu_0 = \mu_1$ 且 $\int |T(x) - x|^2 \, d\mu_0 = W_2^2(\mu_0, \mu_1)$ 。进一步假设

$$T_t = (1-t)\text{Id} + tT \quad (0 \leq t \leq 1)$$

均为 C^1 微分同胚, 且

$$v_t^* = (T - \text{Id}) \circ T_t^{-1}$$

满足定理 3.5 的速度场条件。令 $\mathcal{A}_{\text{sm}}$ 为所有具有这些速度场正则性、起终点分别为 μ_0, μ_1 的流诱导曲线, 则

$$W_2^2(\mu_0, \mu_1) = \inf_{(\mu, v) \in \mathcal{A}_{\text{sm}}} \int_0^1 \int \|v_t\|^2 \, d\mu_t \, dt. \quad (3.12)$$

最小值由 $\mu_t = (T_t)_\# \mu_0$, $v_t = v_t^*$ 达到。

证明直觉 沿任意轨迹走过的位移，是速度在时间上的积分。在单位时间内，Cauchy-Schwarz 告诉我们动能至少是总位移的平方。要取等号，只需先找到最优起终点配对，再让每个粒子沿直线匀速运动。

证明. 先取任意允许的流 Φ_t 。令 $X_0 \sim \mu_0$, $X_t = \Phi_t(X_0)$ ，则 (X_0, X_1) 是 (μ_0, μ_1) 的一个耦合。因此

$$\begin{aligned} W_2^2(\mu_0, \mu_1) &\leq \mathbb{E} \|X_1 - X_0\|^2 \\ &= \mathbb{E} \left\| \int_0^1 v_t(X_t) dt \right\|^2 \\ &\leq \mathbb{E} \int_0^1 \|v_t(X_t)\|^2 dt \\ &= \int_0^1 \int \|v_t(x)\|^2 d\mu_t(x) dt. \end{aligned}$$

若动能无穷，下界自动成立；动能有限时，上述换序由非负性保证。

再取题设的 T_t 。其逐点导数是 $T(x) - x$ ，且

$$\partial_t T_t(x) = v_t^*(T_t(x)).$$

由定理 3.5，推前曲线是一个合法的连续性方程解。通过推前公式计算其动能：

$$\begin{aligned} \int_0^1 \int |v_t^*(y)|^2 d\mu_t(y) dt &= \int_0^1 \int |T(x) - x|^2 d\mu_0(x) dt \\ &= W_2^2(\mu_0, \mu_1). \end{aligned}$$

这给出上界，并证明等号由所述曲线达到。 $\square$

说明 3.8 (光滑定理没有包含哪些存在性结论)。这里明确假设了最优映射存在、插值映射可逆以及所需的流正则性。一般 Benamou–Brenier 理论允许弱连续性方程与有限动能速度，还涉及弱解的表示和逼近。本讲只在上述光滑可实现范围内证明 (3.12)，不以这份短证明代替一般弱理论。

一维高斯满足上述假设。沿其最优插值，

$$X_t = m_t + \sigma_t G, \quad m_t = (1-t)m_0 + tm_1, \quad \sigma_t = (1-t)\sigma_0 + t\sigma_1,$$

所以中间分布仍然是高斯，速度为

$$v_t(x) = (m_1 - m_0) + \frac{\sigma_1 - \sigma_0}{\sigma_t}(x - m_t). \quad (3.13)$$

应区分两种插值：这里是粒子位置的线性插值；把两个密度按 $(1-t)\rho_0 + t\rho_1$ 相加得到的混合分布通常不同。后面的 Flow Matching 也将从粒子路径出发，但任意选取的起终点配对不一定是最优运输配对。

4 Wasserstein 梯度流：目标怎样决定动力学

现在已有两个组成部分：一个需要降低的分布泛函 $\mathcal{F}$ ，以及用动能衡量改变分布的几何。接下来照着有限维梯度下降的思路，寻找单位移动代价下最有效的下降方向。自由能将给出一个特别重要的答案：势能产生输运，熵产生扩散。

4.1 第一变分与搬动粒子的导数

先考虑密度泛函

$$\mathcal{F}(\rho) = \int_{\mathbb{R}^d} H(x, \rho(x)) \, dx.$$

假设 $\rho > 0$ 光滑, $H \in C^2(\mathbb{R}^d \times (0, \infty))$, 且 $\mathcal{F}(\rho)$ 有限. 对紧支撑扰动 η 且 $\int \eta = 0$, 直接对一元参数求导得到

$$\left. \frac{d}{d\epsilon} \mathcal{F}(\rho + \epsilon\eta) \right|_{\epsilon=0} = \int \partial_r H(x, \rho(x)) \eta(x) \, dx.$$

因为扰动有紧支撑且 ρ 在该紧集上有正下界, 小 ϵ 保留非负性, 局部有界性也保证求导可以移入积分。

定义 4.1 (第一变分). 若密度扰动的一阶变化可写成

$$\left. \frac{d}{d\epsilon} \mathcal{F}(\rho + \epsilon\eta) \right|_{\epsilon=0} = \int \xi_\rho(x) \eta(x) \, dx, \quad \int \eta = 0,$$

则称 $\xi_\rho = \delta\mathcal{F}/\delta\rho$ 为第一变分. 它只在相差一个空间常数的意义下唯一, 因为常数与所有零积分扰动的配对均为零. 对于上述积分泛函, 可以取 $\xi_\rho(x) = \partial_r H(x, \rho(x))$ 。

命题 4.2 (搬动位置产生的泛函变化). 设 $\rho > 0$ 为 C^1 概率密度, $H \in C^2(\mathbb{R}^d \times (0, \infty))$, 且 $\mathcal{F}(\rho)$ 有限. 对 $w \in C_c^\infty(\mathbb{R}^d; \mathbb{R}^d)$, 令 $\mu_\epsilon = (\text{Id} + \epsilon w)_\#(\rho \, dx)$, 其密度为 ρ_ϵ . 则

$$\left. \frac{d}{d\epsilon} \mathcal{F}(\rho_\epsilon) \right|_{\epsilon=0} = \int \rho(x) \nabla \xi_\rho(x) \cdot w(x) \, dx. \quad (4.1)$$

证明直觉 密度方向 η 与粒子速度 w 是两种不同的坐标. 搬动粒子时, 连续性方程告诉我们对应的密度变化为 $\eta = -\nabla \cdot (\rho w)$. 将这个变化代入第一变分, 再分部积分, 梯度就从密度变量转到了空间变量。

证明. 小 ϵ 时, $\text{Id} + \epsilon w$ 是微分同胚. 变量代换公式与

$$(\text{Id} + \epsilon w)^{-1}(x) = x - \epsilon w(x) + o(\epsilon), \quad \det(\text{Id} + \epsilon Dw) = 1 + \epsilon \nabla \cdot w + o(\epsilon)$$

给出

$$\rho_\epsilon(x) = \rho(x) - \epsilon \nabla \cdot (\rho w)(x) + o(\epsilon).$$

在包含 w 支撑的固定紧集上, 余项可以一致控制; 在该紧集外, 密度保持不变. 因此

$$\left. \frac{d}{d\epsilon} \mathcal{F}(\rho_\epsilon) \right|_0 = - \int \xi_\rho \nabla \cdot (\rho w) \, dx = \int \rho \nabla \xi_\rho \cdot w \, dx.$$

最后一步没有边界项, 因为 w 有紧支撑. □

4.2 配方得到最速下降速度

在有限维欧氏空间，对目标 $F(x)$ 选择速度 v 时，

$$\nabla F(x) \cdot v + \frac{1}{2}|v|^2 = \frac{1}{2}|v + \nabla F(x)|^2 - \frac{1}{2}|\nabla F(x)|^2,$$

所以最优速度是 $-\nabla F(x)$ 。式 (4.1) 给出了分布空间中完全相同的计算。在 $\int \rho |\nabla \xi_\rho|^2 < \infty$ 时，

$$\begin{aligned} \int \rho \nabla \xi_\rho \cdot v \, dx + \frac{1}{2} \int \rho |v|^2 \, dx \\ = \frac{1}{2} \int \rho |v + \nabla \xi_\rho|^2 \, dx - \frac{1}{2} \int \rho |\nabla \xi_\rho|^2 \, dx. \end{aligned}$$

因此在 $L^2(\rho)$ 速度中，唯一的最小化者为

$$v = -\nabla \xi_\rho.$$

代回连续性方程，得到光滑形式的 Wasserstein 梯度流：

$$\partial_t \rho_t = \nabla \cdot \left(\rho_t \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho_t) \right). \quad (4.2)$$

这是一个动力学的选择。上述最速下降计算本身并不证明偏微分方程的存在性；但只要存在满足下列条件的解，就能严格验证它确实在降低目标。

命题 4.3 (光滑梯度流的耗散恒等式)。 设 $\rho_t > 0$ 是 (4.2) 的概率密度解，关于时间为 C^1 、空间为 C^2 ， $\mathcal{F}(\rho) = \int H(x, \rho) \, dx$ 在所考虑的有限时间内有限， $H \in C^3(\mathbb{R}^d \times (0, \infty))$ ， $\xi_t = \partial_r H(x, \rho_t(x))$ 。假设每个紧时间区间 I 上有 $G_I \in L^1(\mathbb{R}^d)$ ，满足

$$|\xi_t(x) \partial_t \rho_t(x)| \leq G_I(x) \quad (t \in I),$$

且 $\int \rho_t |\nabla \xi_t|^2 < \infty$ 。再假设无穷远的通量满足

$$\lim_{R \rightarrow \infty} \int_{\partial B_R} (1 + |\xi_t|) \rho_t |\nabla \xi_t| \, dS = 0.$$

则

$$\frac{d}{dt} \mathcal{F}(\rho_t) = - \int \rho_t |\nabla \xi_t|^2 \, dx \leq 0. \quad (4.3)$$

若在紧的周期空间上工作，光滑性与严格正性可替代上述无穷远条件。

证明直觉 目标的一阶变化等于“目标想让粒子往哪里走”与“粒子实际怎样走”的内积。最速下降恰好令这两个方向相反，于是内积成为一个负的平方。证明中只会用到链式法则和分部积分。

证明. 由支配条件，

$$\frac{d}{dt} \mathcal{F}(\rho_t) = \int \xi_t \partial_t \rho_t \, dx = \int \xi_t \nabla \cdot (\rho_t \nabla \xi_t) \, dx.$$

先在 B_R 内分部积分，得到

$$\int_{B_R} \xi_t \nabla \cdot (\rho_t \nabla \xi_t) \, dx = - \int_{B_R} \rho_t |\nabla \xi_t|^2 \, dx + \int_{\partial B_R} \xi_t \rho_t \nabla \xi_t \cdot n \, dS.$$

令 $R \rightarrow \infty$ ，最后一项由通量假设消失，前两项分别由支配条件与有限耗散积分收敛，便得到结论。□

例如对 $\mathcal{F}_f(\rho) = D_f(\rho\|\pi) = \int \pi f(\rho/\pi)$, 若 $f \in C^2((0, \infty))$, $\rho, \pi > 0$, 第一变分为

$$\frac{\delta \mathcal{F}_f}{\delta \rho} = f'(\rho/\pi),$$

相应方程是

$$\partial_t \rho = \nabla \cdot (\rho \nabla f'(\rho/\pi)).$$

不同的散度一般产生不同的、甚至非线性的扩散。KL 的对数结构恰好使下一节中的扩散项简化成普通 Laplacian。

4.3 自由能把势能下降与熵扩散放在同一个方程中

设 $V: \mathbb{R}^d \rightarrow \mathbb{R}$ 光滑且

$$0 < Z = \int e^{-V(x)} dx < \infty, \quad \pi(x) = Z^{-1} e^{-V(x)}.$$

再假设 $\int |V(x)|\pi(x) dx < \infty$, 从而目标本身的势能与熵积分均有限。我们希望获得 π 的样本, 但可能无法计算 Z 。定义自由能

$$\mathcal{F}(\rho) = \int V(x)\rho(x) dx + \int \rho(x) \log \rho(x) dx, \quad (4.4)$$

在这里假设两项都有限。由 $\log \pi = -V - \log Z$, 直接得到

$$\text{KL}(\rho\|\pi) = \mathcal{F}(\rho) + \log Z, \quad \mathcal{F}(\rho) = \text{KL}(\rho\|\pi) - \log Z. \quad (4.5)$$

因此自由能的最小值为 $-\log Z$, 最小化者为 π 。归一化常数只给目标增加一个与 ρ 无关的常数, 不影响下降方向。

命题 4.4 (自由能的梯度流是 Langevin 的 Fokker-Planck 方程)。在第一变分与耗散所需的上述正则条件下, 自由能的 Wasserstein 梯度流为

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla V) + \Delta \rho_t = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\pi} \right). \quad (4.6)$$

并且

$$\frac{d}{dt} \text{KL}(\rho_t\|\pi) = - \int \rho_t \left| \nabla \log \frac{\rho_t}{\pi} \right|^2 dx. \quad (4.7)$$

证明直觉 只让粒子沿 $-\nabla V$ 运动, 会把质量推向势能低的地方, 更接近求最小值。采样还需要分布保有适当的分散程度。熵项的第一变分是 $\log \rho + 1$, 它给出速度 $-\nabla \log \rho$, 使质量从密度高的地方向外扩散。把两者相加, 便得到既考虑能量又考虑分散程度的概率演化。

证明. 分别对势能与熵求第一变分:

$$\xi_\rho = V + \log \rho + 1 = \log(\rho/\pi) + 1 - \log Z.$$

空间梯度消去常数, 所以 (4.2) 给出

$$\partial_t \rho = \nabla \cdot (\rho \nabla V + \rho \nabla \log \rho) = \nabla \cdot (\rho \nabla V) + \Delta \rho,$$

其中用到了 $\rho \nabla \log \rho = \nabla \rho$ 。最后由 (4.5), KL 与自由能的时间导数相同; 应用耗散恒等式即得 (4.7)。□

此时可以先把 (4.6) 看作一个确定的密度方程。下一部分从 Brownian motion 与 Itô 积分开始，证明它也是随机粒子系统

$$dX_t = -\nabla V(X_t) dt + \sqrt{2} dW_t$$

所诱导的分布演化。这样，“熵的几何梯度”和“给粒子加入随机扰动”会得到同一个扩散项。JKO 的工作严格建立了这类自由能与 Fokker-Planck 演化之间的变分联系 [4]。

把 $\rho = \pi$ 代入可见通量为零，因此 π 是平稳密度。在连通空间内，若 $\rho > 0$ 且耗散为零，则 $\nabla \log(\rho/\pi) = 0$ ，归一化迫使 $\rho = \pi$ 。耗散恒等式本身还没有给出收敛速度；定量速度需要进一步控制目标值与耗散项的关系。

4.4 回到 ELBO：固定目标以后，才谈它的分布动力学

固定观测 x 与模型参数 θ ，令潜变量空间中的目标为

$$\pi_x(z) = p_\theta(z | x), \quad V_x(z) = -\log p_\theta(x, z), \quad Z_x = p_\theta(x).$$

当联合密度严格为正且满足上面的正则、可积条件时，

$$\mathcal{F}_x(q) = \int V_x q + \int q \log q = -\mathcal{L}(\theta, q; x) = \text{KL}(q \| \pi_x) - \log p_\theta(x).$$

因此，如果选择 Wasserstein 几何来优化 q ，就得到

$$\partial_t q_t = \Delta_z q_t - \nabla_z \cdot (q_t \nabla_z \log p_\theta(x, z)), \quad (4.8)$$

并由 (4.7) 得到

$$\frac{d}{dt} \mathcal{L}(\theta, q_t; x) = \int q_t \left| \nabla_z \log \frac{q_t}{\pi_x} \right|^2 dz \geq 0.$$

这一联系包含三个明确的选择：固定后验目标，选择 Wasserstein 几何，再让整个分布沿其梯度流演化。VAE 中对有限维参数 η 做梯度优化，一般并不等于这个无限维分布梯度流。ELBO 与 Fokker-Planck 的联系由目标和几何共同建立，不能只凭两者都涉及概率就视为同一概念。

4.5 JKO 步：把连续下降变成一个隐式运输更新

欧氏梯度下降可以用一个隐式最小化步骤表达：

$$x^{k+1} \in \arg \min_x \left\{ F(x) + \frac{|x - x^k|^2}{2h} \right\}.$$

把平方距离换成 W_2^2 ，便得到分布空间的 JKO 步 [4, 2]：

$$\mu^{k+1} \in \arg \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{F}(\mu) + \frac{W_2^2(\mu, \mu^k)}{2h} \right\}. \quad (4.9)$$

其中 $\mathcal{F}(\mu)$ 表示密度存在时的相应泛函，并按需要在其他测度上扩展为 $+\infty$ 。第一项鼓励降低目标，第二项控制一次移动的距离。

命题 4.5 (一个 JKO 步的耗散与光滑一阶条件)。假设 (4.9) 的最小化者 μ^{k+1} 存在, 且相关目标值有限。则

$$\mathcal{F}(\mu^{k+1}) + \frac{W_2^2(\mu^{k+1}, \mu^k)}{2h} \leq \mathcal{F}(\mu^k). \quad (4.10)$$

进一步, 若 $\mu^{k+1} = \rho^{k+1} dx$ 有正的光滑密度, 第一变分满足 (4.1), 且存在从 μ^{k+1} 到 μ^k 的最优运输映射 T , 则

$$T(x) = x + h \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho^{k+1})(x) \quad \mu^{k+1} \text{ 几乎处处}. \quad (4.11)$$

证明直觉 最优的新分布不应在任何微小搬动下继续改善目标。证明时无需预先知道 W_2^2 的导数: 保留旧的最优配对, 只搬动它的新端点, 就得到一个显式可微的比较函数。这个比较函数在零位移处同样取最小值, 一阶条件便足够。

证明. 在最小化中取竞争者 $\mu = \mu^k$, 直接得到 (4.10)。对一阶条件, 记 $\mu^+ = \mu^{k+1}$, $\xi^+ = \delta \mathcal{F} / \delta \rho(\rho^{k+1})$ 。取 $w \in C_c^\infty(\mathbb{R}^d; \mathbb{R}^d)$, $\mu_\epsilon = (\text{Id} + \epsilon w)_\# \mu^+$ 。原最优耦合 $(x, T(x))$ 可以改成 $(x + \epsilon w(x), T(x))$, 因此

$$W_2^2(\mu_\epsilon, \mu^k) \leq \int |x + \epsilon w(x) - T(x)|^2 d\mu^+(x).$$

定义显式比较函数

$$\tilde{J}(\epsilon) = \mathcal{F}(\mu_\epsilon) + \frac{1}{2h} \int |x + \epsilon w(x) - T(x)|^2 d\mu^+(x).$$

它不小于 μ_ϵ 的 JKO 目标, 而后者不小于最小值; 在 $\epsilon = 0$ 时, 两次不等式都取等号。故 $\tilde{J}$ 在零点取最小值。对这个可微函数求导, 并用 (4.1), 得到

$$0 = \tilde{J}'(0) = \int \left(\nabla \xi^+(x) + \frac{x - T(x)}{h} \right) \cdot w(x) d\mu^+(x).$$

对所有紧支撑光滑 w 成立, 就得到 (4.11)。□

注意 T 在此处把新分布运回旧分布, 所以 (4.11) 中出现加号。真正向前的步进满足 $x - T(x) = -h \nabla \xi^+(x)$, 正是使用新位置计算梯度的隐式下降。本命题证明了一个光滑最小化步骤的性质; JKO 最小化者的一般存在性与 $h \rightarrow 0$ 的收敛属于进一步的变分分析。

到这里, “在分布空间中做计算” 已经有了具体的数学含义: 散度与 ELBO 给出目标, 最优运输衡量移动代价, 第一变分把目标转成速度, 连续性方程把速度转成分布演化。对 KL, 自由能的下降恰好分成势能输运与熵扩散。接下来引入随机粒子和归一化生灭, 便可以进一步把这些分布操作变成可实现的算法。

5 Brownian motion: 平方根尺度如何改变微积分

前一章用速度场推动粒子, 从而在概率分布空间中移动。本章加入另外两种操作: 给位置添加随机增量, 以及改变不同位置的相对权重。我们先把随机增量对应的微积分建立起来, 再从同一个测试函数计算推出分布方程。这也补全原课程粒子系统讲义中从链式法则到分布演化的主线 [2, 6]。

5.1 Brownian motion 与信息的时间顺序

定义 5.1 (Brownian motion)。在概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 上, 令 $(\mathcal{F}_t)_{t \geq 0}$ 为满足通常条件的滤过, 即随时间增加的信息族, 取完备且右连续的版本。一个 m 维标准 Brownian motion W_t 满足: $W_0 = 0$, 样本路径几乎必然连续, W_t 对 $\mathcal{F}_t$ 可测, 且对任意 $0 \leq s < t$,

$$W_t - W_s \sim \mathcal{N}(0, (t-s)I_m), \quad W_t - W_s \text{ 独立于 } \mathcal{F}_s.$$

过程 H_t 称为适应的, 是指 H_t 对 $\mathcal{F}_t$ 可测: 时刻 t 的选择只能使用已经出现的信息。

独立于整个 $\mathcal{F}_s$ 比仅仅独立于 W_s 更强。我们可以让 $\mathcal{F}_0$ 包含随机初值, 但它不能预先透露未来的 Brownian 增量。一个可积适应过程 M_t 称为鞅, 是指对 $s < t$ 有 $\mathbb{E}[M_t | \mathcal{F}_s] = M_s$: 给定当前信息, 对未来值的最佳均值预测仍是当前值。在一维情形, 对固定 t 与 $h > 0$,

$$W_{t+h} - W_t \stackrel{d}{=} \sqrt{h} Z, \quad Z \sim \mathcal{N}(0, 1), \quad \mathbb{E}|W_{t+h} - W_t|^2 = h. \quad (5.1)$$

这说明增量的均方根尺度为 $\sqrt{h}$, 并不是沿每条样本路径都有一个固定常数控制全部小增量。确定性速度在时间 h 内贡献 $O(h)$ 位移; Brownian 噪声的典型位移则是 $O(\sqrt{h})$ 。因此, 后者的平方仍然与时间 h 同阶。

证明直觉 普通链式法则忽略二阶项, 是因为 n 个长度约 $1/n$ 的时间步中, 每步的位移平方约为 $1/n^2$, 相加后消失。Brownian 每步的位移平方约为 $1/n$, 相加后留下有限量。Itô 修正项的来源, 就是这个没有消失的二阶账目。

5.2 二次变差: 把尺度直觉变成一个 L^2 结论

固定 $T > 0$ 。令 $\mathcal{D} = \{0 = t_0 < \dots < t_n = T\}$ 为确定性分割, 记 $h_i = t_{i+1} - t_i$, $|\mathcal{D}| = \max_i h_i$, $\Delta_i W = W_{t_{i+1}} - W_{t_i}$ 。

命题 5.2 (Brownian 二次变差与交叉变差)。当 $|\mathcal{D}| \rightarrow 0$ 时, 对任意坐标 $\alpha, \beta \in \{1, \dots, m\}$,

$$\sum_i \Delta_i W^\alpha \Delta_i W^\beta \longrightarrow \delta_{\alpha\beta} T \quad \text{于 } L^2(\mathbb{P}). \quad (5.2)$$

若 H_i 对 $\mathcal{F}_{t_i}$ 可测且 $|H_i| \leq C$, 则还有

$$\mathbb{E} \left| \sum_i H_i (\Delta_i W^\alpha \Delta_i W^\beta - \delta_{\alpha\beta} h_i) \right|^2 \leq 2C^2 T |\mathcal{D}|. \quad (5.3)$$

证明直觉 每个平方增量的均值就是这一小段时间的长度。中心化之后, 不同时段之间的交叉项为零; 总方差因此是各段长度平方之和。分割越细, 这个总方差越小。即使系数取决于过去, 条件期望仍能消掉下一步的中心化增量。

证明. 一维高斯变量满足 $\mathbb{E}Z^2 = 1$, $\mathbb{E}Z^4 = 3$, 故

$$\mathbb{E}(\Delta_i W^\alpha)^2 = h_i, \quad \text{Var}((\Delta_i W^\alpha)^2) = 2h_i^2.$$

当 $\alpha \neq \beta$ 时, 同一时段内两坐标独立, 乘积均值为零, 二阶矩为 h_i^2 。因此, 令 $U_i = \Delta_i W^\alpha \Delta_i W^\beta - \delta_{\alpha\beta} h_i$, 有

$$\mathbb{E}[U_i | \mathcal{F}_{t_i}] = 0, \quad \mathbb{E}[U_i^2 | \mathcal{F}_{t_i}] \leq 2h_i^2.$$

对 $i < j$, $H_i U_i H_j$ 已经对 $\mathcal{F}_{t_j}$ 可测, 故 $\mathbb{E}[H_i U_i H_j U_j] = 0$ 。于是

$$\mathbb{E} \left| \sum_i H_i U_i \right|^2 = \sum_i \mathbb{E}[H_i^2 U_i^2] \leq 2C^2 \sum_i h_i^2 \leq 2C^2 T |\mathcal{D}|.$$

取 $H_i = 1$ 即得第一个结论。 $\square$

若 H_t 是有界、连续、适应的过程, 取 $H_i = H_{t_i}$, 再加上普通 Riemann 和的收敛, 可得

$$\sum_i H_{t_i} \Delta_i W^\alpha \Delta_i W^\beta \longrightarrow \delta_{\alpha\beta} \int_0^T H_t dt \quad \text{于概率意义.}$$

这正是稍后 Taylor 展开需要的带权版本。上面的命题对任意确定性细分给出 L^2 收敛, 没有声称所有依赖样本路径的分割都同时成立。例如, 沿二进分割, 方差界可求和, Chebyshev 不等式与 Borel–Cantelli 引理进一步给出几乎必然收敛。结合路径连续性, 这也说明 Brownian 路径几乎必然没有有限全变差: 否则平方增量之和不超过最大增量乘全变差, 必然趋于零, 与极限 T 矛盾。

5.3 Itô 积分: 先决定系数, 再观察下一步噪声

定义 5.3 (简单过程的 Itô 积分). 设 H_i 是对 $\mathcal{F}_{t_i}$ 可测的 $d \times m$ 随机矩阵, 且 $\mathbb{E} \|H_i\|_F^2 < \infty$ 。对简单过程

$$H_t = \sum_{i=0}^{n-1} H_i \mathbf{1}_{(t_i, t_{i+1}]}(t),$$

定义

$$\int_0^T H_t dW_t := \sum_{i=0}^{n-1} H_i (W_{t_{i+1}} - W_{t_i}). \quad (5.4)$$

这里 $\|A\|_F^2 = \text{Tr}(AA^\top)$ 是矩阵 Frobenius 范数的平方。区间内尚未结束时, 将最后一个增量截在当前时刻, 就定义了整个积分过程。

命题 5.4 (Itô 等距与零均值). 简单过程的积分是连续的平方可积鞅, 且

$$\mathbb{E} \int_0^T H_t dW_t = 0, \quad \mathbb{E} \left\| \int_0^T H_t dW_t \right\|^2 = \mathbb{E} \int_0^T \|H_t\|_F^2 dt. \quad (5.5)$$

证明直觉 系数可以随机, 也可以根据过去不断调整; 它只要没有偷看下一步噪声, 下一步增量的条件均值就仍是零。于是不同时间段的贡献在 L^2 中正交, 求和后的能量等于每一步能量之和。这个等式使积分能从简单过程稳定地延拓出去。

证明. 记 $V_i = H_i \Delta_i W$ 。因为 H_i 在区间左端已知,

$$\mathbb{E}[V_i | \mathcal{F}_{t_i}] = 0, \quad \mathbb{E}[\|V_i\|^2 | \mathcal{F}_{t_i}] = h_i \text{Tr}(H_i H_i^\top).$$

对 $i < j$ 条件于 $\mathcal{F}_{t_j}$, 得到 $\mathbb{E}\langle V_i, V_j \rangle = 0$ 。展开平方后, 只剩对角项, 便得到等距公式和零均值。在任意两个确定时刻之间细分, 同样的条件期望计算给出鞅性质; 路径连续性则直接来自 Brownian 路径连续性。 $\square$

现在以 $\|H\|_{H^2}^2 = \mathbb{E} \int_0^T \|H_t\|_F^2 dt$ 为范数，将简单适应过程取闭包。若 $H^{(k)} \rightarrow H$ 于这个范数，则等距公式保证 $\int H^{(k)} dW$ 是 L^2 Cauchy 列，其极限与逼近方式无关。这就定义了 $\int H dW$ 。平方可积的连续适应过程属于这个闭包：先截断过程的大小，再用左端取值的分段常数过程逼近，即可用控制收敛证明。

延拓后的积分仍有连续鞅版本。这里使用的 Doob 最大不等式为

$$\mathbb{E} \sup_{0 \leq t \leq T} \|M_t\|^2 \leq 4\mathbb{E} \|M_T\|^2, \quad (5.6)$$

其中 M 为从零出发的连续平方可积鞅。对右连续且有左极限的平方可积鞅，同一个最大不等式也成立，后文的跳跃鞅将使用这个版本。因此积分过程的逼近也在 L^2 的一致范数下收敛。这个不等式的核心很短：记 $M^* = \sup_{t \leq T} \|M_t\|$ ，在首次越过 λ 的时刻停下，对次鞅 $\|M_t\|$ 取条件期望，得到 $\lambda \mathbb{P}(M^* \geq \lambda) \leq \mathbb{E}[\|M_T\| \mathbf{1}_{\{M^* \geq \lambda\}}]$ 。对 λ 积分并用 Cauchy-Schwarz，便有 $\mathbb{E}(M^*)^2 \leq 2\mathbb{E}[M^* \|M_T\|] \leq 2\|M^*\|_{L^2} \|M_T\|_{L^2}$ ；先截断 M^* 再取极限可避免预先假设其可积。

若 H 仅局部平方可积，就在 $\int_0^t \|H_s\|_F^2 ds$ 或过程大小首次超过整数阈值时停止，再将各段积分拼接。这样得到的是局部鞅；没有额外可积性时，不能直接写它的期望为零。另一方面，即使积分是平方可积鞅，随机系数也不保证积分服从高斯分布。

例 5.5 (为什么不能随意更换取值点)。由恒等式

$$W_T^2 = 2 \sum_i W_{t_i} \Delta_i W + \sum_i (\Delta_i W)^2$$

及前面的两个收敛结论，一维情形下

$$\int_0^T W_t dW_t = \frac{1}{2}(W_T^2 - T). \quad (5.7)$$

右端均值为零，却不是高斯变量。若改用右端取值 $W_{t_{i+1}}$ ，两种和的差恰为平方增量之和，极限相差 T 。取值规则因此是积分定义的一部分。

5.4 Itô 公式：Taylor 二阶项的累积

定理 5.6 (Itô 公式)。设 W 为 m 维 Brownian motion, $X_t \in \mathbb{R}^d$ 具有积分表示

$$X_t = X_0 + \int_0^t \beta_s ds + \int_0^t \Gamma_s dW_s, \quad \Gamma_s \in \mathbb{R}^{d \times m}.$$

本讲采用以下充分条件： X_0 对 $\mathcal{F}_0$ 可测且几乎必然有限， β, Γ 是连续适应过程，随机积分按上述局部化定义，且 $u \in C^{1,2}([0, T] \times \mathbb{R}^d)$ ，即对时间有连续一阶导数、对空间有连续二阶导数。那么几乎必然，对全部 $t \in [0, T]$ 同时成立

$$\begin{aligned} u(t, X_t) - u(0, X_0) &= \int_0^t \left[\partial_s u + \langle \beta_s, \nabla u \rangle + \frac{1}{2} \text{Tr}(\Gamma_s \Gamma_s^\top \nabla^2 u) \right] (s, X_s) ds \\ &\quad + \int_0^t \nabla u(s, X_s)^\top \Gamma_s dW_s. \end{aligned} \quad (5.8)$$

随机积分一般先按局部意义理解；若其被积函数满足平方可积条件，就是真正的零均值鞅。

证明直觉 先对一个小时间段做普通 Taylor 展开，再将所有小段相加。一阶随机项成为 Itô 积分；二阶随机项的条件平均留下 $\frac{1}{2} \text{Tr}(\Gamma \Gamma^\top \nabla^2 u) dt$ 。漂移的平方以及漂移与噪声的交叉项都消失。真正需要证明的并非形式上的 $dW dW = dt$ ，而是这几类和分别收敛到什么。

证明。先在系数和所用导数有界的情形证明，再局部化。

第一步：一维 *Brownian motion*。先只取 $X = W$ 。对空间做二阶 Taylor 展开，对时间使用积分形式的一阶展开，得到

$$\begin{aligned} u(T, W_T) - u(0, W_0) &= \sum_i \int_{t_i}^{t_{i+1}} \partial_t u(s, W_{t_{i+1}}) ds + \sum_i \partial_x u(t_i, W_{t_i}) \Delta_i W \\ &\quad + \frac{1}{2} \sum_i \partial_{xx} u(t_i, W_{t_i}) (\Delta_i W)^2 + \sum_i R_i. \end{aligned}$$

时间项由路径连续性收敛到 $\int_0^T \partial_t u(t, W_t) dt$ 。随机一阶项由等距和左端逼近收敛到 $\int_0^T \partial_x u(t, W_t) dW_t$ 。由命题 5.2，二阶项收敛到 $\frac{1}{2} \int_0^T \partial_{xx} u(t, W_t) dt$ 。最后，空间二阶导数在每个紧集上一致连续；若 ω_u 为该紧集上的连续模，则

$$\left| \sum_i R_i \right| \leq \frac{1}{2} \omega_u \left(\max_i |\Delta_i W| \right) \sum_i (\Delta_i W)^2 \rightarrow 0 \quad \text{于概率意义.}$$

这里第一个因子趋于零，第二个因子由二次变差结论在概率意义下有界。这也表明只需空间 C^2 ，不必额外假设三阶导数存在。

第二步：多维 *Brownian motion* 与冻结系数。多维 Taylor 二阶项为 $\frac{1}{2} \sum_{\alpha, \beta} \partial_{\alpha\beta} u \Delta_i W^\alpha \Delta_i W^\beta$ 。带权交叉变差极限为零，带权平方变差留下对角线之和，即 $\frac{1}{2} \Delta u$ 。在一个区间左端，把 β, Γ 冻结成当时已知的随机常数。条件于左端信息，对函数

$$(s, w) \mapsto u(s, x + \beta(s - s_0) + \Gamma w)$$

应用刚才的 Brownian 公式，得到该区间上 $\partial_s u + \beta \cdot \nabla u + \frac{1}{2} \text{Tr}(\Gamma \Gamma^\top \nabla^2 u)$ 。未来 Brownian 增量独立于左端信息，故这一条件计算合法。对简单适应系数，逐段相加便证明了公式。

第三步：逼近连续系数。用左端简单过程 $\beta^{(k)}, \Gamma^{(k)}$ 逼近，并令 $X^{(k)} = X_0 + \int \beta^{(k)} ds + \int \Gamma^{(k)} dW$ 。有界性、连续性、等距和最大不等式给出

$$\mathbb{E} \sup_{t \leq T} \|X_t^{(k)} - X_t\|^2 \rightarrow 0.$$

在已对 $X^{(k)}$ 证明的公式中，普通积分项用控制收敛，随机积分项用 Itô 等距，即可逐项取极限。先对每个有理 t 成立，再利用两边连续性，得到对所有 t 同时成立。一般局部条件下，在 $\|X\| + \|\beta\| + \|\Gamma\|$ 首次达到阈值时停止，并截断 u 。停止后的系数只额外出现一个跳点，仍能作同样的 L^2 简单过程逼近。令阈值趋于无穷，便得到所述公式。□

因此常用的符号规则

$$dW_t^\alpha dW_t^\beta = \delta_{\alpha\beta} dt, \quad dt dW_t = 0, \quad (dt)^2 = 0$$

是上述极限定理的记忆法，不是普通实数无穷小的乘法规则。

6 SDE 与 Fokker–Planck：从轨道到概率分布

6.1 SDE 是一个积分方程

写

$$dX_t = b(t, X_t) dt + \Sigma(t, X_t) dW_t \quad (6.1)$$

的准确含义是

$$X_t = X_0 + \int_0^t b(s, X_s) ds + \int_0^t \Sigma(s, X_s) dW_s. \quad (6.2)$$

其中 $b: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ 是漂移, $\Sigma: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times m}$ 是扩散系数矩阵。它不是将不存在的普通导数 $\dot{W}_t$ 代入 ODE。

定理 6.1 (全局 Lipschitz 系数下的存在唯一性)。设 b, Σ 连续, 且存在 $L, K < \infty$, 使得对全部 t, x, y ,

$$\begin{aligned} \|b(t, x) - b(t, y)\|^2 + \|\Sigma(t, x) - \Sigma(t, y)\|_F^2 &\leq L^2 \|x - y\|^2, \\ \|b(t, 0)\|^2 + \|\Sigma(t, 0)\|_F^2 &\leq K^2. \end{aligned}$$

若 X_0 对 $\mathcal{F}_0$ 可测且 $\mathbb{E}\|X_0\|^2 < \infty$, 则对于给定的 Brownian motion, 存在唯一的连续适应解, 满足 $\mathbb{E}\sup_{t \leq T} \|X_t\|^2 < \infty$ 。这里唯一性指路径唯一性: 使用同一个初值和同一条 Brownian 路径的两个解几乎必然完全相同。

证明直觉 Picard 迭代仍然有效, 只是比较两次迭代时, 要用 Itô 等距控制随机积分。Lipschitz 条件保证输入轨道的误差不会被系数过度放大。每多迭代一次, 都多积一次时间; 反复积分产生的阶乘最终压过所有固定常数。

证明. 令 $X_t^{(0)} = X_0$, 并递归定义

$$X_t^{(k+1)} = X_0 + \int_0^t b(s, X_s^{(k)}) ds + \int_0^t \Sigma(s, X_s^{(k)}) dW_s.$$

假设蕴含线性增长界 $\|b(t, x)\|^2 + \|\Sigma(t, x)\|_F^2 \leq 2K^2 + 2L^2 \|x\|^2$ 。由 Cauchy–Schwarz、Itô 等距及最大不等式, 每次迭代均连续、适应, 且有有限的 L^2 最大范数。记 $D_k(t) = \mathbb{E}\sup_{u \leq t} \|X_u^{(k+1)} - X_u^{(k)}\|^2$ 。对 $k \geq 1$, 上述不等式给出

$$D_k(t) \leq 2L^2(T+4) \int_0^t D_{k-1}(s) ds.$$

若 $C_T = 2L^2(T+4)$, $A_T = \sup_{t \leq T} D_0(t) < \infty$, 则归纳得到

$$D_k(T) \leq A_T \frac{(C_T T)^k}{k!}.$$

由于 $\sum_k \sqrt{D_k(T)} < \infty$, 迭代在 L^2 的一致范数下收敛到连续适应过程 X 。Lipschitz 条件与等距保证两个积分也收敛, 所以极限满足积分方程。对两个解作完全相同的估计, 差异函数 $D(t)$ 满足 $D(t) \leq C_T \int_0^t D(s) ds$; Grönwall 不等式给出 $D = 0$ 。若尚未知道待比较的解有全局二阶矩, 先在两条轨道离开同一个大球时停止, 作相同估计, 再令球半径趋于无穷即可。□

这里列的是容易检查的充分条件, 并不要求每一个后续模型都必须全局 Lipschitz。如果只满足局部条件, 就需要另外证明解不会在有限时间逃向无穷; 本定理没有省略这一步来覆盖它们。

6.2 OU 过程：先解一个能逐项核验的模型

命题 6.2 (Ornstein–Uhlenbeck 过程)。设 $\kappa > 0$ 、 $g > 0$ 、 $\bar{x} \in \mathbb{R}$ ，并考虑

$$dX_t = -\kappa(X_t - \bar{x}) dt + g dW_t.$$

其解为

$$X_t = \bar{x} + e^{-\kappa t}(X_0 - \bar{x}) + g \int_0^t e^{-\kappa(t-s)} dW_s. \quad (6.3)$$

若 $X_0 = x_0$ 确定，则

$$X_t \sim \mathcal{N}\left(\bar{x} + e^{-\kappa t}(x_0 - \bar{x}), \frac{g^2}{2\kappa}(1 - e^{-2\kappa t})\right). \quad (6.4)$$

若 X_0 随机、对 $\mathcal{F}_0$ 可测且平方可积，则

$$\begin{aligned} \mathbb{E}X_t &= \bar{x} + e^{-\kappa t}(\mathbb{E}X_0 - \bar{x}), \\ \text{Var}(X_t) &= e^{-2\kappa t} \text{Var}(X_0) + \frac{g^2}{2\kappa}(1 - e^{-2\kappa t}). \end{aligned} \quad (6.5)$$

平稳分布为 $\mathcal{N}(\bar{x}, g^2/(2\kappa))$ ，且上述初始分布的边缘分布随 $t \rightarrow \infty$ 收敛到它。

证明直觉 漂移不断擦去初值的影响，噪声不断注入新方差。过去注入的噪声也会被漂移衰减，所以长时间后的总方差是一个有限的指数衰减积分。稳态方差来自收缩和随机扰动的平衡。

证明. 对 $u(t, x) = e^{\kappa t}(x - \bar{x})$ 应用 Itô 公式。它对 x 是线性的，因此二阶导数为零，漂移恰好抵消：

$$d(e^{\kappa t}(X_t - \bar{x})) = ge^{\kappa t} dW_t.$$

积分后即得显式解。被积函数是确定性的，故随机积分是高斯变量：它是独立高斯增量的确定系数线性组合的 L^2 极限。由等距，其方差为

$$g^2 \int_0^t e^{-2\kappa(t-s)} ds = \frac{g^2}{2\kappa}(1 - e^{-2\kappa t}).$$

该积分独立于 $\mathcal{F}_0$ ，从而得到均值方差公式。取 $X_0 \sim \mathcal{N}(\bar{x}, g^2/(2\kappa))$ ，独立高斯项相加后分布不变。一般初值的衰减项趋于零，而高斯噪声项的方差趋于 $g^2/(2\kappa)$ ，给出所述分布收敛。□

若随机 X_0 不是高斯变量，有限时刻的 X_t 通常也不是高斯变量；式 (6.5) 仍然成立。例如 $\kappa = 1$ 、 $\bar{x} = 2$ 、 $g = 1/\sqrt{2}$ 时，平稳分布正是 $\mathcal{N}(2, 1/4)$ 。

6.3 先用测试函数，再写密度方程

对概率测度 μ ，简记 $\mu(\varphi) = \int \varphi d\mu$ 。令 $\mu_t = \text{Law}(X_t)$ ，并定义生成算子

$$\mathcal{L}_t \varphi(x) = \langle b(t, x), \nabla \varphi(x) \rangle + \frac{1}{2} \sum_{i,j=1}^d a_{ij}(t, x) \partial_{ij} \varphi(x), \quad a = \Sigma \Sigma^\top. \quad (6.6)$$

定理 6.3 (SDE 对应的 Fokker–Planck 方程)。在定理 6.1 的条件下, 对每个 $\varphi \in C_c^\infty(\mathbb{R}^d)$,

$$\mu_t(\varphi) - \mu_0(\varphi) = \int_0^t \mu_s(\mathcal{L}_s \varphi) \, ds. \quad (6.7)$$

这称为 $\partial_t \mu_t = \mathcal{L}_t^* \mu_t$ 的弱形式。若 μ_t 存在密度 ρ_t , 则密度在分布意义下满足

$$\partial_t \rho_t = - \sum_{i=1}^d \partial_i (b_i \rho_t) + \frac{1}{2} \sum_{i,j=1}^d \partial_{ij} (a_{ij} \rho_t). \quad (6.8)$$

若进一步 ρ 对时间 C^1 , $b_i \rho$ 对空间 C^1 , $a_{ij} \rho$ 对空间 C^2 , 且这些导数连续, 则该等式逐点成立。

证明直觉 描述一个分布怎样变化, 可以先观察所有光滑读数 $\varphi(X_t)$ 怎样变化。Itô 公式给出每条轨道上的读数增量; 取期望去掉零均值噪声后, 就得到分布的确定性演化。最后用分部积分, 把导数从测试函数移到密度上。漂移项移动一次导数, 产生负号; 扩散项移动两次导数, 符号恢复为正。

证明. 对 $u(t, x) = \varphi(x)$ 应用 Itô 公式, 有

$$\varphi(X_t) - \varphi(X_0) = \int_0^t \mathcal{L}_s \varphi(X_s) \, ds + \int_0^t \nabla \varphi(X_s)^\top \Sigma(s, X_s) \, dW_s.$$

因为 $\nabla \varphi$ 支持在紧集上, Σ 在该紧集与有限时间区间上有界, 最后的积分平方可积且均值为零。同样, $\mathcal{L}_s \varphi$ 有界, 故可以交换时间积分和期望, 得到弱形式。若存在密度, 则

$$\begin{aligned} \int \rho_t b_i \partial_i \varphi \, dx &= - \int \varphi \partial_i (b_i \rho_t) \, dx, \\ \int \rho_t a_{ij} \partial_{ij} \varphi \, dx &= \int \varphi \partial_{ij} (a_{ij} \rho_t) \, dx. \end{aligned}$$

紧支持测试函数消去了边界项; 没有经典导数时, 上面两行就是分布导数的定义。再允许对时间作紧支持测试, 便得到时空分布意义下的 PDE。在额外的连续可微条件下, 两个连续函数作为分布相等, 因而逐点相等。□

三个容易混淆的点。噪声维数 m 与位置维数 d 可以不同; 分布方程依赖的是 $a = \Sigma \Sigma^\top$ 。若 $\Sigma = g I_d$ 且 g 不依赖位置, 扩散项为 $\frac{g^2}{2} \Delta \rho$; 因此 $\sqrt{2} \, dW$ 对应 $\Delta \rho$, 标准 dW 对应 $\frac{1}{2} \Delta \rho$ 。若 a 依赖位置, 正确形式是 $\frac{1}{2} \partial_{ij} (a_{ij} \rho)$, 一般不能改写成 $\frac{1}{2} \nabla \cdot (a \nabla \rho)$ 。

密度的存在并不是全局 Lipschitz 假设的自动结论。例如 $\Sigma = 0$ 且初值为一个点时, 分布可以始终是 Dirac 测度。因此弱形式 (6.7) 是更基础的对象。概率质量为一则直接来自 $\mu_t = \text{Law}(X_t)$, 也可在测试函数计算中取 $\varphi = 1$ 得到。若只从密度 PDE 出发在全空间积分, 还需用衰减、零通量或合适的截断论证来消去无穷远边界。

最后回看纯 transport: 令 $\Sigma = 0$, $b = v$, 就有 $\frac{d}{dt} \mu_t(\varphi) = \mu_t(v_t \cdot \nabla \varphi)$, 对应前章的连续性方程 $\partial_t \rho + \nabla \cdot (\rho v) = 0$ 。这与式 (3.9) 是同一个方程。扩散保留同一个漂移项, 只是额外加入了由二次变差产生的二阶算子。

7 生灭与权重: 带归一化反应的粒子系统

移动位置、增加噪声、改变相对权重, 是三种不同的分布操作。前两种操作由轨道上的 SDE 描述; 第三种操作需要给路径附上权重, 或者在粒子群中复制与替换。

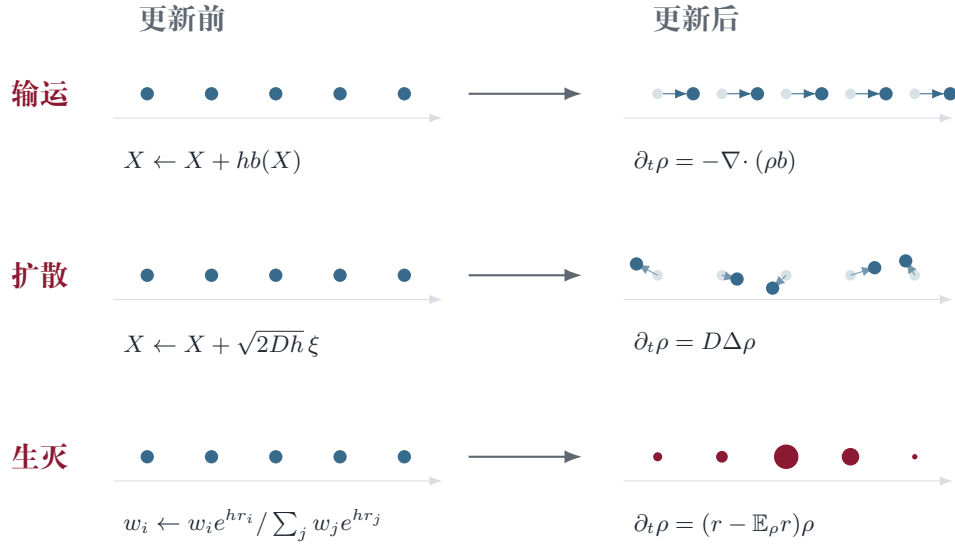


图 1: 同一组粒子的三种更新。输运改变位置, 扩散加入尺度为 $\sqrt{h}$ 的随机增量, 生灭改变质量的分配。第三行用点的大小表示权重; 复制与删除可以把权重重新表示成等权样本。此图仅示意一步操作, 不表示有限个粒子已经满足连续密度方程; 反应本身也不能在原先零质量的区域创造新支持。

7.1 从 Feynman-Kac 权重推导反应项

命题 7.1 (路径加权与归一化反应)。 令 X 解式 (6.1), 初始分布为概率测度 μ_0 , 并满足定理 6.1 的条件。设 $r : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ 连续, 且 $|r(t, x)| \leq R_r$ 。定义路径权重、未归一化测度及其总质量

$$H_t = \exp\left(\int_0^t r(s, X_s) ds\right), \quad \gamma_t(\varphi) = \mathbb{E}[H_t \varphi(X_t)], \quad Z_t = \gamma_t(1). \quad (7.1)$$

则 γ_t 是正的有限测度, $e^{-R_r t} \leq Z_t \leq e^{R_r t}$, 且

$$\frac{d}{dt} \gamma_t(\varphi) = \gamma_t(\mathcal{L}_t \varphi + r_t \varphi). \quad (7.2)$$

归一化后的概率测度 $\mu_t = \gamma_t / Z_t$ 满足

$$\frac{d}{dt} \mu_t(\varphi) = \mu_t(\mathcal{L}_t \varphi) + \mu_t(r_t \varphi) - \mu_t(r_t) \mu_t(\varphi). \quad (7.3)$$

这里可取 $\varphi \in C_c^\infty$, 也可取常数函数 1。若存在密度, 则相应的分布方程为

$$\partial_t \tilde{\rho}_t = \mathcal{L}_t^* \tilde{\rho}_t + r_t \tilde{\rho}_t, \quad (7.4)$$

$$\partial_t \rho_t = \mathcal{L}_t^* \rho_t + (r_t - \mathbb{E}_{\rho_t} r_t) \rho_t. \quad (7.5)$$

证明直觉 短时间内, 位于 x 的路径权重乘上 $1 + r_t(x) dt$ 。因此未归一化质量多出 $r_t \tilde{\rho}_t$ 。随后把全体质量重新缩放为一, 就必须扣掉群体的平均增长率。一个位置是否增加相对质量, 取决于它是否比平均水平增长得快。

证明. 权重满足普通有限变差方程 $dH_t = H_t r(t, X_t) dt$. 对乘积 $H_t \varphi(X_t)$ 使用 Itô 公式; 有限变差项与 Brownian 项的二次交叉变差为零, 故

$$d(H_t \varphi(X_t)) = H_t (\mathcal{L}_t \varphi + r_t \varphi)(X_t) dt + H_t \nabla \varphi(X_t)^\top \Sigma(t, X_t) dW_t.$$

由于 $H_t \leq e^{R_r T}$, 随机积分仍平方可积. 取期望给出未归一化等式. 令 $\varphi = 1$, 得到

$$Z'_t = \gamma_t(r_t) = Z_t \mu_t(r_t).$$

对商 $\gamma_t(\varphi)/Z_t$ 求导, 便得

$$\frac{\gamma'_t(\varphi)}{Z_t} - \frac{\gamma_t(\varphi) Z'_t}{Z_t^2} = \mu_t(\mathcal{L}_t \varphi + r_t \varphi) - \mu_t(\varphi) \mu_t(r_t).$$

密度形式由前一节同样的分部积分得到. 特别地, 取 $\varphi = 1$ 时右侧为零, 验证了归一化后的质量守恒. $\square$

这是 Feynman–Kac 的前向测度表述; 它与常见的后向价值函数表述使用同一个路径权重 [6, 第 7.7 节]. 这里的 μ_t 是加权路径群的分布, 一般不等于单个未加权轨道的 $\text{Law}(X_t)$. 式 (7.5) 虽然总质量为一, 仍应称为带归一化反应项的 Fokker–Planck 方程; 其中包含空间位置上的质量增加与减少.

例 7.2 (没有位置移动时, 生灭就是指数重加权). 当 $b = 0$ 、 $\Sigma = 0$, 且 r 不依赖时间时,

$$\mu_t(dx) = \frac{e^{tr(x)} \mu_0(dx)}{\int e^{tr(y)} \mu_0(dy)}.$$

例如只有 A, B 两个位置, $p_t = \mu_t(\{A\})$, 则 $\dot{p}_t = (r(A) - r(B))p_t(1 - p_t)$. 它可以重新分配两处的比例, 但不会在原支持以外产生一个新位置. 同时, 将 $r_t(x)$ 加上任何仅依赖时间的常数, 归一化演化完全不变.

7.2 从权重到算法: 先区分两个误差来源

设当前用带权粒子表示分布:

$$\mu_k^N = \sum_{i=1}^N w_i^k \delta_{x_i^k}, \quad w_i^k \geq 0, \quad \sum_i w_i^k = 1.$$

一个简单时间步长为 h 的实现, 是先作指数权重步, 再作位置 Euler 步:

$$w_i^{k+1} = \frac{w_i^k e^{hr(t_k, x_i^k)}}{\sum_j w_j^k e^{hr(t_k, x_j^k)}}, \quad (7.6)$$

$$x_i^{k+1} = x_i^k + hb(t_k, x_i^k) + \sqrt{h} \Sigma(t_k, x_i^k) \xi_i^k, \quad \xi_i^k \stackrel{\text{iid}}{\sim} \mathcal{N}(0, I_m). \quad (7.7)$$

各步的新噪声独立于已知历史. 这里明确选择先反应再移动的分裂顺序; 这是连续方程的离散近似, 并不是其精确有限步长解. 指数形式保持权重非负, 无须为了非负性强制 $1 + hr_i \geq 0$.

证明直觉 指数权重的一阶展开扣除平均增长率, Euler 位置步的一阶条件期望产生生成算子. 把这两个一阶计算相加, 就能核对程序是否对应目标分布方程. 时间步长误差来自离散化; 粒子数误差来自用有限点集代表分布, 两者应分别考虑.

一步一致性核验。固定当前有限粒子配置，记 $r_i = r(t_k, x_i^k)$ 。由 r 有界，

$$w_i^{k+1} = w_i^k [1 + h(r_i - \mu_k^N(r))] + O(h^2).$$

对 $\varphi \in C_b^3(\mathbb{R}^d)$ ，位置步的增量 Δ_i 满足

$$\mathbb{E}[\Delta_i | \mathcal{F}_{t_k}] = hb_i, \quad \mathbb{E}[\Delta_i \Delta_i^\top | \mathcal{F}_{t_k}] = ha_i + h^2 b_i b_i^\top.$$

二阶 Taylor 展开及 $\mathbb{E} \|\Delta_i\|^3 = O(h^{3/2})$ 因而给出

$$\mathbb{E}[\varphi(x_i^{k+1}) | \mathcal{F}_{t_k}] = \varphi(x_i^k) + h\mathcal{L}_{t_k}\varphi(x_i^k) + o(h).$$

这里余项针对固定配置；若系数有界，则相应估计可对配置一致。乘以新权重并求和，得到

$$\mathbb{E}[\mu_{k+1}^N(\varphi) | \mathcal{F}_{t_k}] = \mu_k^N(\varphi) + h[\mu_k^N(\mathcal{L}_{t_k}\varphi) + \mu_k^N(r\varphi) - \mu_k^N(r)\mu_k^N(\varphi)] + o(h).$$

这与式 (7.3) 的漂移一致。 $\square$

重采样的无偏性，究竟指什么？ 当权重高度集中时，可给定当前带权点集 $\nu^N = \sum_i w_i \delta_{x_i}$ ，记其全部位置与权重的信息为 $\mathcal{G}$ 。条件于 $\mathcal{G}$ ，独立抽取索引 $A_1, \dots, A_N$ ，满足 $\mathbb{P}(A_j = i | \mathcal{G}) = w_i$ ，然后令新粒子为 x_{A_j} ，权重统一为 $1/N$ 。记 $\hat{\nu}^N = N^{-1} \sum_j \delta_{x_{A_j}}$ ，则

$$\mathbb{E}[\hat{\nu}^N(\varphi) | \mathcal{G}] = \nu^N(\varphi), \quad \text{Var}(\hat{\nu}^N(\varphi) | \mathcal{G}) = \frac{\nu^N(\varphi^2) - \nu^N(\varphi)^2}{N}. \quad (7.8)$$

证明就是给定点集后对 N 个独立抽样的均值、方差求和。因此，这里的无偏是对当前离散测度的条件无偏。自归一化的有限粒子估计本身一般仍有偏，不能据此声称它无偏估计真实目标。给定当前点集，新粒子是从该离散分布条件独立抽出的；消去这个条件后通常相关。例如对不同新粒子 $i \neq j$ ， $\text{Cov}(\varphi(x_{A_i}), \varphi(x_{A_j})) = \text{Var}(\nu^N(\varphi))$ 。它们也不能被直接视为从连续目标分布得到的独立样本。

7.3 固定 N 的成对复制与替换：精确的条件漂移

还可以始终保持等权粒子 $\mu_t^N = N^{-1} \sum_{i=1}^N \delta_{X_t^i}$ 。各粒子在跳跃之间用独立 Brownian 驱动，遵循同一个 SDE。仍采用前面对 b, Σ, r 的假设，且各初值平方可积；不要求初值之间相互独立。在当前配置下，记 $r_i = r(t, X_t^i)$ 。对每个有序对 (i, j) ，以非负速率

$$q_{i \rightarrow j} = \frac{(r_i - r_j)_+}{N}, \quad z_+ = \max\{z, 0\}, \quad (7.9)$$

把接收者 j 的位置替换成供者 i 的位置，其他位置不变。每次事件同时复制一个粒子并删除另一个，故粒子总数严格等于 N 。当 $|r| \leq R_r$ 时，总跳跃速率不超过 $2R_r N$ ，有限时间内不会发生无穷多次跳跃。

命题 7.3 (成对替换的漂移与鞅噪声)。 对上述系统和 $\varphi \in C_c^\infty(\mathbb{R}^d)$,

$$d\mu_t^N(\varphi) = [\mu_t^N(\mathcal{L}_t\varphi) + \mu_t^N(r_t\varphi) - \mu_t^N(r_t)\mu_t^N(\varphi)] dt + dM_t^N(\varphi), \quad (7.10)$$

其中 $M^N(\varphi)$ 是从零出发的平方可积鞅。若 $C_\varphi = \sup_{t \leq T, x} \|\Sigma(t, x)^\top \nabla \varphi(x)\|$ ， $B_\varphi = \|\varphi\|_\infty$ ，则

$$\mathbb{E} \sup_{t \leq T} |M_t^N(\varphi)|^2 \leq \frac{4T}{N} (C_\varphi^2 + 8R_r B_\varphi^2). \quad (7.11)$$

在前面的连续系数假设下，紧支持的 $\nabla \varphi$ 保证 $C_\varphi < \infty$ 。

证明直觉 让高增长率粒子替换低增长率粒子，可以直接增加前者的相对份额。速率中的 $1/N$ 保证全群每单位时间约有 $O(N)$ 次操作；每次操作对平均读数只改变 $O(1/N)$ 。因此漂移是 $O(1)$ ，而累积方差是 $O(N) \times O(1/N^2) = O(1/N)$ 。下面的代数检查表明，漂移恰好就是所需的中心化反应项。

证明. 在固定配置处简记 $\varphi_i = \varphi(X^i)$ 。事件 $i \rightarrow j$ 使经验读数增加 $(\varphi_i - \varphi_j)/N$ ，故跳跃部分的条件漂移为

$$\begin{aligned} \sum_{i,j} \frac{(r_i - r_j)_+}{N} \frac{\varphi_i - \varphi_j}{N} &= \frac{1}{2N^2} \sum_{i,j} (r_i - r_j)(\varphi_i - \varphi_j) \\ &= \frac{1}{N} \sum_i r_i \varphi_i - \left(\frac{1}{N} \sum_i r_i \right) \left(\frac{1}{N} \sum_i \varphi_i \right). \end{aligned}$$

第一行将 (i, j) 与 (j, i) 配对，并使用 $z_+ - (-z)_+ = z$ 。第二行展开四项求和。这正是 $\mu^N(r\varphi) - \mu^N(r)\mu^N(\varphi)$ 。连续运动部分由每个粒子的 Itô 公式给出 $\mu^N(\mathcal{L}_t\varphi)$ 。将实际跳跃数减去其条件速率的时间积分，得到补偿跳跃鞅；再加上 Brownian 积分鞅，便得式 (7.10)。

为核对噪声大小，记 $\langle M^N \rangle$ 为可预测二次变差，即鞅增量条件方差的累积。独立 Brownian 驱动给出的单位时间二次变差为

$$\frac{1}{N^2} \sum_i \left\| \Sigma(t, X^i)^\top \nabla \varphi(X^i) \right\|^2 \leq \frac{C_\varphi^2}{N}.$$

跳跃部分为

$$\frac{1}{N^3} \sum_{i,j} (r_i - r_j)_+ (\varphi_i - \varphi_j)^2 \leq \frac{8R_r B_\varphi^2}{N}.$$

连续鞅与纯跳跃鞅的交叉二次变差为零。因此 $\mathbb{E} |M_T^N|^2 = \mathbb{E} \langle M^N \rangle_T$ 不超过两个上界的时间积分。再用 Doob 最大不等式，便得到结论。□

这是一条有限 N 的精确随机方程。它显示，固定时间区间内每个测试读数的鞅噪声具有 $N^{-1/2}$ 均方根尺度。但要证明整个经验分布逼近确定性的 μ_t ，还须控制初始误差与非线性漂移的稳定性；上述噪声界本身没有证明完整的粒子极限，也没有证明粒子相互独立。第 8 节将补全这一环：定理 8.9 证明纯生灭的固定粒子数复制替换极限，定理 8.10 证明三种操作组合后的带权粒子极限。当 $R_r > 0$ 时，一个明确的时钟实现是：用速率 $2R_r N$ 的独立 Poisson 时钟产生候选事件，每次均匀选择一个有序对 (i, j) ，再以 $(r_i - r_j)_+ / (2R_r)$ 的概率接受替换。候选速率乘选中概率再乘接受概率，正好等于式 (7.9)。每次要把位置推进到候选时刻后，再计算接受概率；若 $R_r = 0$ 则没有替换。这避免了把随位置变化的速率错误地冻结成下一次事件的恒定速率。一般 SDE 的位置推进仍需数值近似，相应时间误差另行处理。

7.4 回到采样：密度依赖的增长率

原生灭采样工作选择增长率与当前密度和目标密度的偏差有关 [7]。例如

$$r[\rho_t](x) = -\lambda \log \frac{\rho_t(x)}{\pi(x)}, \quad \lambda \geq 0.$$

中心化后，当前相对过多的位置减少份额，相对不足的位置增加份额。目标 π 的未知归一化常数只给增长率增加一个常数，因此会消去。与 Langevin 扩散组合，可以同时探索位置与调整已有

区域之间的比例。它与前章 KL 的 Wasserstein 梯度流的联系 [4]，将在综合部分通过同一个能量耗散公式进一步解释。

说明 7.4 (实现与定理的边界)。 这里的 $r[\rho]$ 依赖未知密度，且未必有界。前面针对给定有界增长率 Feynman–Kac 命题，不是这个非线性方程存在唯一性的证明。另外，经验测度 $N^{-1} \sum_i \delta_{X^i}$ 是原子的，不能直接对它写 $\log \rho(X^i)$ 。实际算法需要额外的密度估计或经过定义的替代量；相应近似与正则化也需单独分析。

同一条计算链。 对位置作 transport，用普通一阶链式法则，得到连续性方程。给位置加入 Brownian diffusion，二次变差留下 Itô 二阶项，得到 Fokker–Planck 方程。再给路径指数权重或在群体中复制与替换，得到反应项；除以总质量后，增长率自动中心化。从始至终，测试函数把粒子层面的操作翻译为分布层面的方程。

8 从粒子到 PDE：三种操作的严格对应

前两章已经推导了生成算子。本节把这个计算补成粒子与分布方程之间的严格结论，并区分三个对象：一组实际粒子的经验测度、一个粒子的概率分布，以及粒子数趋于无穷后的极限。它们有关联，却不能在推导中直接互换。

证明直觉 测试函数 φ 可以看作一个测量仪器。粒子移动时，仪器的读数改变；粒子出生或死亡时，求和中的项数改变。先精确计算平均读数的变化，再说明随机误差如何消失，才得到了粒子算法与确定性 PDE 的对应。对随机经验测度直接删掉噪声项，会跳过这个对应中最关键的一步。

本节固定有限时间区间 $[0, T]$ ，空间为 $\mathbb{R}^d$ 。记 C_b^k 为函数及其至多 k 阶空间导数连续且有界的函数类， C_c^k 还要求紧支撑。概率测度曲线 $t \mapsto \mu_t$ 称为窄连续，是指对每个有界连续函数 φ ， $t \mapsto \mu_t(\varphi)$ 连续。除最后明确讨论的密度依赖问题外，系数 b, Σ, r 都是预先给定的确定性函数。

定义 8.1 (测度意义下的分布演化)。令 $a = \Sigma \Sigma^\top$ ，且 $\mathcal{L}_t$ 为式 (6.6) 的算子。窄连续概率测度曲线 μ 满足带归一化反应的分布方程，是指对所有 $\varphi \in C_c^\infty(\mathbb{R}^d)$ 和 $t \in [0, T]$ ，

$$\mu_t(\varphi) - \mu_0(\varphi) = \int_0^t \left[\mu_s(\mathcal{L}_s \varphi) + \mu_s(r_s \varphi) - \mu_s(r_s) \mu_s(\varphi) \right] ds. \quad (8.1)$$

若密度存在，其对应的 PDE 是

$$\partial_t \rho = -\nabla \cdot (b\rho) + \frac{1}{2} \sum_{i,j=1}^d \partial_{ij} (a_{ij} \rho) + \left(r - \int r \rho \, dx \right) \rho. \quad (8.2)$$

删去不使用的系数，就得到下面各个操作的弱方程。

弱方程允许原子初值，也允许演化始终没有密度。将 $\varphi(x)$ 换成时间与空间的测试函数、对时间分部积分，再把空间导数移到测度上，就得到时空分布形式。只有在相关导数连续存在时，才进一步得到逐点成立的经典 PDE。本节对每个结论列出足够且便于核验的条件，不把形式计算当作额外正则性的证明。

8.1 输运：有限粒子本身就满足弱连续性方程

前面的定理 3.5 建立了流与连续性方程的联系。这里把粒子算法中的量词说清楚：**有限个确定性运动的粒子，已经精确满足测度形式的 PDE；增加粒子数的作用，是逼近所希望的初始分布。**本节采用方便验证的充分条件： $b, D_x b$ 在 $[0, T] \times \mathbb{R}^d$ 上连续，且

$$B := \sup_{t,x} |b(t, x)| < \infty, \quad L := \sup_{t,x} \|D_x b(t, x)\| < \infty.$$

记 $\Phi_{s,t}(x)$ 为从时刻 s 、位置 x 出发的 ODE 解。Picard 迭代与全局 Lipschitz 性给出唯一性，有界速度排除了有限时间逃逸。因此正向与反向流都存在，并满足 $\Phi_{s,t}^{-1} = \Phi_{t,s}$ 及 $|\Phi_{s,t}(x) - x| \leq B|t - s|$ 。

定理 8.2 (输运粒子与弱 PDE 的精确对应)。给定任意位置 $x_1, \dots, x_N$ 和固定权重 $w_i \geq 0$ ， $\sum_i w_i = 1$ ，令

$$X_t^i = \Phi_{0,t}(x_i), \quad \mu_t^N = \sum_{i=1}^N w_i \delta_{X_t^i}.$$

则对每个 $\varphi \in C_c^1(\mathbb{R}^d)$, 逐路径精确成立

$$\mu_t^N(\varphi) - \mu_0^N(\varphi) = \int_0^t \mu_s^N(b(s, \cdot) \cdot \nabla \varphi) \, ds. \quad (8.3)$$

更一般地, 对任意概率测度 μ_0 , 推前 $\mu_t = (\Phi_{0,t})_{\#} \mu_0$ 是弱连续性方程 $\partial_t \mu_t + \nabla \cdot (b \mu_t) = 0$ 的唯一窄连续概率测度解。这些结论均不要求初始测度有密度或有限二阶矩。

证明直觉 弱 PDE 测量的是所有光滑观测量的变化。每个粒子的读数变化由普通链式法则精确给出, 乘上固定权重再求和就得到 PDE。要证明唯一性, 把一个终点观测量沿特征线向后搬运; 沿流不变的观测量能够直接把终点分布读回初始分布。

证明. 有限粒子的恒等式。 对每条轨迹, 链式法则给出

$$\varphi(X_t^i) - \varphi(x_i) = \int_0^t b(s, X_s^i) \cdot \nabla \varphi(X_s^i) \, ds.$$

乘 w_i 求和即得式 (8.3)。该计算对每组初始点成立, 所以随机初值也不需要取期望才能得到此式。对一般 μ_0 , 把同一恒等式关于初始点积分即可。因为被积函数绝对值不超过 $B \|\nabla \varphi\|_{\infty}$, Fubini 定理适用。对任意 $f \in C_b(\mathbb{R}^d)$, $f(\Phi_{0,t}(x))$ 关于 t 连续且被 $\|f\|_{\infty}$ 控制, 支配收敛给出 $t \mapsto \mu_t(f)$ 连续, 即所需的窄连续性。

允许随时间变化的测试函数。 设 ν_t 是任一窄连续概率测度弱解。如果弱解起初仅对 C_c^{∞} 定义, 先将 C_c^1 函数作空间磨光, 在函数与一阶导数的一致范数中取极限, 即可得到同一恒等式。对空间支撑位于固定紧集内的 $u \in C^1([0, t] \times \mathbb{R}^d)$, 它进一步推出

$$\nu_t(u(t, \cdot)) - \nu_0(u(0, \cdot)) = \int_0^t \nu_s(\partial_s u + b \cdot \nabla u) \, ds. \quad (8.4)$$

为严格验证, 把区间分割成 $[t_k, t_{k+1}]$, 在每段将增量拆成

$$\nu_{t_{k+1}}(u(t_{k+1}) - u(t_k)) + (\nu_{t_{k+1}} - \nu_{t_k})(u(t_k)).$$

第一项用时间微积分, 第二项用固定测试函数的弱方程, 然后求和。当网格趋于零, ∇u 的一致连续性控制第二项的冻结误差。映射 $(a, s) \mapsto \nu_a(\partial_s u(s, \cdot))$ 在紧时间正方形上连续: 这由窄连续性及 $s \mapsto \partial_s u(s, \cdot)$ 的一致范数连续性得到。因此第一项也收敛到所需积分, 证明式 (8.4)。

反向特征线确定唯一解。 固定 $\tau \leq T$ 和 $h \in C_c^1(\mathbb{R}^d)$, 取 $u(s, x) = h(\Phi_{s,\tau}(x))$ 。流关于空间的 C^1 性来自对其积分方程求导; 流的复合恒等式进一步给出

$$\partial_s \Phi_{s,\tau}(x) = -D_x \Phi_{s,\tau}(x) b(s, x), \quad \partial_s u + b \cdot \nabla u = 0.$$

若 $\text{supp } h \subset B_R$, 有界位移保证 $\text{supp } u(s, \cdot) \subset B_{R+BT}$, 故这个时间依赖测试函数确实合法。代入式 (8.4) 得

$$\nu_{\tau}(h) = \nu_0(h \circ \Phi_{0,\tau}).$$

C_c^1 函数区分有限 Borel 测度, 因而 $\nu_{\tau} = (\Phi_{0,\tau})_{\#} \nu_0$ 。这对每个 τ 成立, 唯一性得证。□

推论 8.3 (何时可以把测度方程写成密度方程)。 若 $\mu_0(dx) = \rho_0(x) dx$, 则 μ_t 也有密度, 且几乎处处

$$\rho_t(\Phi_{0,t}(x)) = \rho_0(x) \exp \left[- \int_0^t \nabla \cdot b(s, \Phi_{0,s}(x)) \, ds \right]. \quad (8.5)$$

该密度在分布意义下满足 $\partial_t \rho + \nabla \cdot (b \rho) = 0$ 。

证明. 令 $A_t = D_x \Phi_{0,t}(x)$. 对流的积分方程求导, 得到

$$\dot{A}_t = D_x b(t, \Phi_{0,t}(x)) A_t, \quad A_0 = \text{Id}.$$

反向流提供 A_t 的逆; 其行列式从 1 连续变化而不会过零. Jacobi 行列式公式与迹的循环性给出

$$\frac{d}{dt} \log \det A_t = \text{Tr}(A_t^{-1} \dot{A}_t) = \text{Tr} D_x b(t, \Phi_{0,t}(x)) = \nabla \cdot b(t, \Phi_{0,t}(x)).$$

积分后用 C^1 微分同胚的变量代换公式 $\rho_t(\Phi_{0,t}(x)) \det A_t = \rho_0(x)$, 即得结论. 将有密度的 μ_t 代入弱恒等式, 按分布导数的定义移去测试函数上的导数, 便得到密度 PDE. 这里不要求 ρ_0 本身光滑. $\square$

式 (8.5) 的含义是: 散度记录小体积单元的相对膨胀率, 而质量守恒使密度按相反比例变化. 若初始测度是原子测度, 输运后仍是原子测度; 此时直接使用式 (8.3), 无须引入不存在的密度.

推论 8.4 (有限粒子逼近目标分布的含义). 若 X_0^i 独立同分布于 μ_0 且 $w_i = 1/N$, 则对每个固定 t 和每个固定有界 Borel 函数 φ ,

$$\mathbb{E} \mu_t^N(\varphi) = \mu_t(\varphi), \quad \mathbb{E} |\mu_t^N(\varphi) - \mu_t(\varphi)|^2 = \frac{\text{Var}_{\mu_t}(\varphi)}{N} \leq \frac{\|\varphi\|_\infty^2}{N}.$$

因此这些固定观测量在 L^2 中收敛.

证明. 确定性映射 $\Phi_{0,t}$ 保持样本间的独立性, 所以 $\varphi(X_t^i)$ 独立同分布, 均值为 $\mu_t(\varphi)$. 展开样本均值的平方误差, 异指标的中心化交叉项均为零, 余下 N 个相同方差项, 即得公式. $\square$

这里的抽样误差来自用 μ_0^N 近似 μ_0 ; 有限粒子关于其自身初始测度的弱 PDE 恒等式始终精确成立.

8.2 扩散: 经验测度带鞅, 分布满足确定性方程

给粒子独立加入噪声时, 必须分清随机的有限粒子群与确定性的概率分布. 本节把定理 6.3 的单粒子推导提升为有限粒子恒等式, 再在纯扩散情形证明 PDE 解反过来确实由 Brownian 粒子表示.

命题 8.5 (经验测度的精确鞅方程). 设 $b: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$, $\Sigma: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times m}$ 连续、有界, 且关于空间一致全局 Lipschitz. 令 $K_\Sigma = \sup_{t,x} \|\Sigma(t, x)\|_F$. 取独立同分布、平方可积的初值 $X_0^i \sim \mu_0$, 以及相互独立且独立于全部初值的 m 维 Brownian motion W^i . 定义

$$dX_t^i = b(t, X_t^i) dt + \Sigma(t, X_t^i) dW_t^i, \quad \mu_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_t^i}, \quad \mu_t = \text{Law}(X_t^1).$$

用式 (6.6) 的 $\mathcal{L}_t$ 与 $a = \Sigma \Sigma^\top$. 对固定 $\varphi \in C_c^\infty(\mathbb{R}^d)$, 几乎必然对全部 $t \in [0, T]$ 有

$$\mu_t^N(\varphi) - \mu_0^N(\varphi) = \int_0^t \mu_s^N(\mathcal{L}_s \varphi) ds + M_t^{N, \varphi}, \quad (8.6)$$

$$M_t^{N, \varphi} = \frac{1}{N} \sum_{i=1}^N \int_0^t \nabla \varphi(X_s^i)^\top \Sigma(s, X_s^i) dW_s^i. \quad (8.7)$$

这是平方可积连续鞅，其二次变差及最大估计为

$$\langle M^{N,\varphi} \rangle_t = \frac{1}{N} \int_0^t \mu_s^N \left(\left\| \Sigma(s, \cdot)^\top \nabla \varphi \right\|^2 \right) ds, \quad (8.8)$$

$$\mathbb{E} \sup_{t \leq T} \left| M_t^{N,\varphi} \right|^2 \leq \frac{4TK_\Sigma^2 \|\nabla \varphi\|_\infty^2}{N}. \quad (8.9)$$

确定性曲线 μ_t 则满足完整的时空弱形式：对空间支撑包含于一个固定紧集的 $\psi \in C^{1,2}([0, T] \times \mathbb{R}^d)$ 且 $\psi(T, \cdot) = 0$,

$$\int \psi(0, x) \mu_0(dx) + \int_0^T \int \left(\partial_t \psi + \sum_i b_i \partial_i \psi + \frac{1}{2} \sum_{i,j} a_{ij} \partial_{ij} \psi \right) (t, x) \mu_t(dx) dt = 0. \quad (8.10)$$

此外，对每个固定 t ,

$$\mathbb{E} \left| \mu_t^N(\varphi) - \mu_t(\varphi) \right|^2 = \frac{\text{Var}_{\mu_t}(\varphi)}{N} \leq \frac{\|\varphi\|_\infty^2}{N}. \quad (8.11)$$

证明直觉 每个粒子的噪声读数都是零均值，但一次实际模拟中的平均读数仍有波动。平均 N 个正交噪声把方差缩小到 $1/N$ ，并没有使有限 N 的噪声完全消失。因此，对真实概率分布取期望可以精确得到确定性 PDE；对有限粒子经验测度，则应保留鞅，或者证明它在所考察的尺度上趋于零。

证明. 定理 6.1 保证强解存在唯一。对每条轨道应用 Itô 公式后求平均，即得式 (8.6)。被积函数由 $K_\Sigma \|\nabla \varphi\|_\infty$ 控制，故随机积分平方可积。不同 W^i 在联合滤过下的交叉变差为零，因此鞅的二次变差是各项二次变差之和，给出式 (8.8)。再用 Itô 等距及 Doob 不等式 (5.6) 得到最大估计。

对单粒子使用时间相关的 Itô 公式，取期望；所有被积项有界，Fubini 定理可用，且鞅均值为零。这给出式 (8.10)。另一方面，Picard 构造表明每个 X^i 只由其自身 (X_0^i, W^i) 决定。输入独立同分布，故粒子过程也独立同分布。展开 N 个中心化读数的平方，交叉期望均为零，即得式 (8.11)。□

式 (8.10) 无需假设密度存在。若另外已知 $\mu_t(dx) = \rho_t(x) dx$ ，则在空间分布意义下

$$\begin{aligned} \int \rho b_i \partial_i \psi dx &= - \int \psi \partial_i (b_i \rho) dx, \\ \int \rho a_{ij} \partial_{ij} \psi dx &= - \int \partial_j \psi \partial_i (a_{ij} \rho) dx = \int \psi \partial_j \partial_i (a_{ij} \rho) dx. \end{aligned}$$

没有经典导数时，右端积分理解为分布与测试函数的配对。连同时间分部积分，这正是

$$\partial_t \rho = -\nabla \cdot (b\rho) + \frac{1}{2} \sum_{i,j} \partial_{ij} (a_{ij} \rho), \quad \rho|_{t=0} = \mu_0$$

的分布形式与弱初值条件。位置相关的 a 必须留在两次导数内部。

这里两个 $1/N$ 结论的用途不同：式 (8.9) 控制一个测试函数上的时间一致鞅误差；式 (8.11) 控制固定时刻的采样误差。它们均未声称对所有测试函数同时有相同界。若系数依赖整个粒子群，独立驱动噪声仍可使交叉变差为零，但粒子轨道通常相关，此时不能沿用独立样本的方差恒等式。

命题 8.6 (Brownian 粒子与热方程的双向对应)。设 $\kappa > 0$, μ_0 为任意概率测度。令 $X_t = X_0 + \sqrt{2\kappa} W_t$, 其中 $X_0 \sim \mu_0$ 独立于 W 。则 $t > 0$ 时其分布具有光滑密度

$$\rho_t(x) = \int G_t(x-y) \mu_0(dy), \quad G_t(z) = \frac{e^{-\|z\|^2/(4\kappa t)}}{(4\pi\kappa t)^{d/2}}. \quad (8.12)$$

它满足 $\partial_t \rho = \kappa \Delta \rho$, 且 $\rho_t(x) dx \Rightarrow \mu_0$ 当 $t \downarrow 0$ 。反过来, 初值为 μ_0 的热方程在连续概率测度曲线中有唯一弱解, 就是式 (8.12), 因而可由上述 Brownian 粒子表示。这里连续是指对每个有界连续函数 f , $t \mapsto \mu_t(f)$ 连续。

证明直觉 Brownian 粒子在每个正时间把一个点摊成高斯云; 一般初值就是把这些高斯云混合起来。热方程唯一性的关键也很直观: 每个 Fourier 频率都按一个确定的指数速度衰减, 所以给定全部初始频率后, 没有选择另一条分布演化的余地。

证明. 条件于 $X_0 = y$, X_t 的密度为 $G_t(x-y)$, 由 Fubini 得到卷积表示及总质量为一。对任意 $\varepsilon > 0$, 热核的各阶时间、空间导数在 $[\varepsilon, T] \times \mathbb{R}^d$ 上有界, 故可在积分下逐次求导。直接计算有

$$\partial_t G_t(z) = \left(-\frac{d}{2t} + \frac{\|z\|^2}{4\kappa t^2} \right) G_t(z) = \kappa \Delta G_t(z).$$

这证明正时间的光滑性及热方程。又因 $X_t \rightarrow X_0$ 几乎必然, 对有界连续 f 用控制收敛便得弱初值。

为证明唯一性, 设 ν_t 为另一个连续概率测度弱解。在时空弱形式中逼近区间 $[0, t]$ 的指示函数, 并用时间连续性处理端点, 便得 $\nu_t(\varphi) - \mu_0(\varphi) = \kappa \int_0^t \nu_s(\Delta \varphi) ds$ 。固定 $\zeta \in \mathbb{R}^d$, 记 $h_t(\zeta) = \int e^{i\zeta \cdot x} \nu_t(dx)$ 。复值测试等价于分别对实部、虚部测试。取 $\chi \in C_c^\infty$, $0 \leq \chi \leq 1$, 在单位球上等于 1, 并令 $\chi_R(x) = \chi(x/R)$ 。对 $\chi_R e^{i\zeta \cdot x}$ 使用弱方程, 其中

$$\Delta(\chi_R e^{i\zeta \cdot x}) = -\|\zeta\|^2 \chi_R e^{i\zeta \cdot x} + e^{i\zeta \cdot x} (2i\zeta \cdot \nabla \chi_R + \Delta \chi_R).$$

括号项的一致上界为 $C(\|\zeta\|/R + R^{-2})$ 。因 ν_s 总质量为一, 积分误差也趋于零; 其余项由控制收敛处理。令 $R \rightarrow \infty$ 得

$$h_t(\zeta) - h_0(\zeta) = -\kappa \|\zeta\|^2 \int_0^t h_s(\zeta) ds, \quad h_t(\zeta) = e^{-\kappa \|\zeta\|^2 t} h_0(\zeta).$$

后一个式子由前一个标量线性积分方程唯一确定。它恰好是热核卷积的特征函数。对任意 $\varphi \in C_c^\infty$, Fourier 反演及 Fubini 将特征函数相等转为测试积分相等, 故 $\nu_t = G_t * \mu_0$ 。□

8.3 生灭: 先写未归一化质量, 再证明概率归一化

本小节先固定粒子的位置, 只改变质量。设 $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$, $r \in C_b([0, T] \times \mathbb{R}^d)$ 且 $|r| \leq R_r$ 。增长率 r 是预先给定的函数; 它不依赖未知的当前密度。以下测试函数均可取实值 $\varphi \in C_b(\mathbb{R}^d)$, 无需空间导数, 也不要求 μ_0 有密度。

命题 8.7 (指数权重给出反应方程的唯一测度解)。定义

$$H_t(x) = \exp\left(\int_0^t r(s, x) \, ds\right), \quad \gamma_t(dx) = H_t(x)\mu_0(dx), \quad \mu_t = \frac{\gamma_t}{\gamma_t(1)}.$$

则 γ 是下式唯一的弱连续正有限测度解, μ 是下式唯一的弱连续概率测度解:

$$\gamma_t(\varphi) = \mu_0(\varphi) + \int_0^t \gamma_s(r_s \varphi) \, ds, \quad (8.13)$$

$$\mu_t(\varphi) = \mu_0(\varphi) + \int_0^t [\mu_s(r_s \varphi) - \mu_s(r_s)\mu_s(\varphi)] \, ds. \quad (8.14)$$

若 μ_0 有密度, 这就是 $\partial_t \tilde{\rho} = r \tilde{\rho}$ 与 $\partial_t \rho = (r - \mathbb{E}_\rho r) \rho$ 的分布意义解。

证明直觉 每个位置的绝对质量独立地按增长率 $r_t(x)$ 演化, 所以其解是指数乘子。总质量也随之改变; 除以总质量, 恰好减去平均增长率。反过来, 把归一化过程失去的总质量乘回来, 就能把非线性概率方程变回线性方程。

证明. 由 $e^{-R_r t} \leq H_t \leq e^{R_r t}$, 可在积分号内求导, 得到第一式。记 $Z_t = \gamma_t(1)$, 则 $Z'_t = \gamma_t(r_t)$ 且 $Z_t \geq e^{-R_r t} > 0$; 商法则给出第二式。若两条正有限测度曲线均解第一式, 其差 Δ_t 满足

$$\|\Delta_t\|_{\text{TV}} \leq R_r \int_0^t \|\Delta_s\|_{\text{TV}} \, ds.$$

这里总变差范数可对所有 $\|\varphi\|_\infty \leq 1$ 的连续有界函数取上确界得到。弱连续性保证有限时间内的总质量有界, 故 Gronwall 不等式给出 $\Delta = 0$ 。对第二式的任意概率测度解 ν , 令

$$A_t = \exp\left(\int_0^t \nu_s(r_s) \, ds\right).$$

乘积法则表明 $A_t \nu_t$ 解第一式, 故 $A_t \nu_t = \gamma_t$ 。取测试函数 1 得 $A_t = Z_t$, 从而 $\nu_t = \mu_t$ 。整个论证针对测度成立, 原子初值也包括在内。□

有限个位置上的指数权重是精确解。 取 $X_0^1, \dots, X_0^N \stackrel{\text{iid}}{\sim} \mu_0$, 并令

$$w_i(t) = \frac{H_t(X_0^i)}{\sum_j H_t(X_0^j)}, \quad \hat{\mu}_t^N = \sum_i w_i(t) \delta_{X_0^i}.$$

直接求导有 $\dot{w}_i = w_i(r_t(X_0^i) - \hat{\mu}_t^N(r_t))$ 。因此对每一组初始样本, $\hat{\mu}^N$ 都精确满足式 (8.14), 只是初值换成经验测度。这里使用的是 $\int_0^t r(s, x) \, ds$ 的精确指数, 并未将增长率冻结成离散时间步。此外, 对每个固定 $t \leq T$,

$$\mathbb{E}|\hat{\mu}_t^N(\varphi) - \mu_t(\varphi)|^2 \leq \frac{4e^{4R_r T} \|\varphi\|_\infty^2}{N}. \quad (8.15)$$

为证明该界, 写 $\theta = \mu_t(\varphi)$ 、 $\psi = \varphi - \theta$, 则

$$\hat{\mu}_t^N(\varphi) - \theta = \frac{N^{-1} \sum_i H_t(X_0^i) \psi(X_0^i)}{N^{-1} \sum_i H_t(X_0^i)}.$$

分母逐样本不小于 $e^{-R_r t}$; 分子的各项独立、均值为零, 且绝对值不超过 $2e^{R_r t} \|\varphi\|_\infty$ 。计算样本均值的方差便得结论。

真正的复制与死亡：随机总数对应未归一化方程

命题 8.8 (分枝系统的期望质量与反应方程)。从上述 N 个独立初始位置出发，每个存活粒子在位置 x 以速率 $r_+(t, x)$ 产生一个同位置的新粒子，以速率 $r_-(t, x)$ 死亡，其中 $r_+ = \max(r, 0)$ 、 $r_- = \max(-r, 0)$ ；不同粒子使用独立的事件时钟。记存活集合为 $\mathcal{A}_t$ ，总数为 K_t ，并定义

$$\Gamma_t^N = \frac{1}{N} \sum_{u \in \mathcal{A}_t} \delta_{x_u}.$$

该系统在有限时间内不爆炸，并且

$$\Gamma_t^N(\varphi) = \Gamma_0^N(\varphi) + \int_0^t \Gamma_s^N(r_s \varphi) \, ds + B_t^N(\varphi), \quad \langle B^N(\varphi) \rangle_t = \frac{1}{N} \int_0^t \Gamma_s^N(|r_s| \varphi^2) \, ds, \quad (8.16)$$

其中 B^N 为零初值平方可积鞅。因此 $\mathbb{E}\Gamma_t^N = \gamma_t$ ，且

$$\mathbb{E}|\Gamma_t^N(\varphi) - \gamma_t(\varphi)|^2 \leq \frac{2e^{2R_r T} \|\varphi\|_\infty^2}{N}. \quad (8.17)$$

证明直觉 一次出生增加 $\varphi(x)/N$ ，一次死亡减少同样的量。两者的平均变化相减，产生 $r\varphi$ ；两者的方差贡献相加，产生 $|r|\varphi^2$ 。所以反应项来自事件的条件均值，有限粒子的额外噪声由补偿计数鞅明确记录。

证明. 忽略死亡，把每个粒子的出生率提高到 R_r ，得到支配 K_t 的 Yule 纯出生过程 Y_t 。在 Y 首次达到 m 时停下，生成算子作用于 k 给出 $\mathbb{E}Y_{T \wedge \tau_m} \leq Ne^{R_r T}$ ，故 $\mathbb{P}(\tau_m \leq T) \leq Ne^{R_r T}/m$ 。令 $m \rightarrow \infty$ 得非爆炸。死亡次数不超过初始数加出生次数，因此原系统也没有有限时间内的无穷次事件。对 k^2 的截停生成算子恒等式先用 Gronwall 不等式，得到不依赖 m 的界 $\mathbb{E}Y_{t \wedge \tau_m}^2 \leq (N^2 + N)e^{2R_r t}$ 。由于 $Y_{t \wedge \tau_m} = \min(Y_t, m)$ 随 m 递增，可在 k 、 k^2 的截停恒等式两侧用单调收敛去掉停时。有限的极限矩因此满足 $m'_1 = R_r m_1$ 、 $m'_2 = 2R_r m_2 + R_r m_1$ ，解得

$$\mathbb{E}Y_t = Ne^{R_r t}, \quad \mathbb{E}Y_t^2 = N^2 e^{2R_r t} + N(e^{2R_r t} - e^{R_r t}).$$

每个出生、死亡计数减去其强度的时间积分后都是补偿计数鞅。把它们分别乘以 $\varphi(x)/N$ 、 $-\varphi(x)/N$ 并求和，漂移为 $\Gamma^N(r\varphi)$ ，二次变差率为 $N^{-2} \sum_{u \in \mathcal{A}_s} |r_s(x_u)| \varphi(x_u)^2$ 。这证明式 (8.16)。由 $K_s \leq Y_s$ ，其期望二次变差至多为 $R_r T e^{R_r T} \|\varphi\|_\infty^2 / N$ ，故局部鞅实际上平方可积。取期望并使用命题 8.7 的唯一性，即得 $\mathbb{E}\Gamma_t^N = \gamma_t$ 。

最后， N 个初始祖先各自的整个后代群相互独立。每群对未归一化读数的贡献不超过其后代数乘 $\|\varphi\|_\infty$ ，其二阶矩由从一个粒子开始的 Yule 过程界住，不超过 $2e^{2R_r T} \|\varphi\|_\infty^2$ 。对这 N 个独立群的贡献取平均，方差相除以 N ，便得最后一式。□

必须留意， $\Gamma_t^N(1) = K_t/N$ 是随机总质量，灭绝时为零。随机归一化的期望不能换成期望质量的比值。例如 $N = 1$ ，初始等概率位于 A, B ，且 $r(A) = a > 0$ 、 $r(B) = 0$ 。全部后代始终留在初始位置，故归一化粒子群落在 A 的平均比例始终是 $1/2$ ；但确定性反应解给出的比例是 $e^{at}/(1 + e^{at})$ 。这也解释了为什么需要分别说明绝对质量、概率归一化以及粒子数极限。

固定粒子数的复制与替换：完整的粒子极限证明

现在保持 N 个等权粒子。位置只在替换时改变：有序对 (i, j) 以速率 $(r_t(X^i) - r_t(X^j))_+/N$ 把 j 替换成 i 。记其经验测度为 ν_t^N ，初值仍为独立的 μ_0 样本。总事件率不超过 $2R_r N$ ，所以该纯跳跃系统可以由 Poisson 候选时钟严格构造。

定理 8.9 (有相互依赖的粒子仍以均方根尺度逼近反应解)。对上述系统、每个固定 $t \leq T$ 与 $\varphi \in C_b(\mathbb{R}^d)$,

$$\mathbb{E}|\nu_t^N(\varphi) - \mu_t(\varphi)|^2 \leq \frac{4(1 + 8R_r T)e^{4R_r T}\|\varphi\|_\infty^2}{N}. \quad (8.18)$$

这里 μ_t 是命题 8.7 的唯一归一化反应解。

证明直觉 已知复制漂移正确、鞅噪声小，还不足以得到粒子极限，因为漂移含有粒子群自身的平均增长率。证明分两次消去漂移：向后测试函数吸收每个位置的增长，积分因子补回群体的总增长。消去之后只剩初始抽样误差和一个可计算二次变差的鞅，复制产生的相关性便包含在这个鞅里。

证明. 固定终点 t , 记 $K = \|\varphi\|_\infty$ 、 $\theta = \mu_t(\varphi)$ 、 $\psi = \varphi - \theta$, 并令

$$h_s(x) = \psi(x) \exp\left(\int_s^t r(u, x) \, du\right), \quad A_s^N = \exp\left(\int_0^s \nu_u^N(r_u) \, du\right), \quad 0 \leq s \leq t.$$

于是 $\partial_s h_s = -r_s h_s$, $\|h_s\|_\infty \leq 2Ke^{R_r(t-s)}$, 且 $e^{-R_r s} \leq A_s^N \leq e^{R_r s}$ 。命题 7.3 的有序对求和, 在这里对任意连续有界读数同样成立。在跳跃间对 h_s 求导, 在跳跃时加上读数增量, 使得

$$d\nu_s^N(h_s) = [\nu_s^N(\partial_s h_s + r_s h_s) - \nu_s^N(r_s)\nu_s^N(h_s)] \, ds + dM_s^N(h) = -\nu_s^N(r_s)\nu_s^N(h_s) \, ds + dM_s^N(h).$$

这里 $M^N(h)$ 表示时间依赖读数的补偿跳跃鞅。 A^N 连续且有有限变差, 故乘积法则没有额外跳跃或二次变差项, 得到

$$A_t^N(\nu_t^N(\varphi) - \theta) = \nu_0^N(H_t \psi) + \mathcal{M}_t^{N,t}, \quad \mathcal{M}_s^{N,t} = \int_0^s A_u^N \, dM_u^N(h). \quad (8.19)$$

精确计算这个鞅的二次变差: 令 $r_i = r(s, X_{s-}^i)$, 则

$$\begin{aligned} \frac{d}{ds} \langle \mathcal{M}^{N,t} \rangle_s &= \frac{(A_s^N)^2}{N^3} \sum_{i,j} (r_i - r_j)_+ (h_s(X_{s-}^i) - h_s(X_{s-}^j))^2 \\ &\leq \frac{8R_r}{N} (A_s^N)^2 \|h_s\|_\infty^2 \leq \frac{32R_r K^2 e^{2R_r t}}{N}. \end{aligned}$$

因此 $\mathcal{M}^{N,t}$ 平方可积, 且 $\mathbb{E}|\mathcal{M}_t^{N,t}|^2 \leq 32R_r t K^2 e^{2R_r t}/N$ 。另一方面, $\mu_0(H_t \psi) = 0$, 初值独立给出

$$\mathbb{E}|\nu_0^N(H_t \psi)|^2 \leq \frac{4K^2 e^{2R_r t}}{N}.$$

鞅终值条件于初始信息的期望为零, 所以它与初始项的交叉期望也为零。在式 (8.19) 中先计算右侧二阶矩, 再用逐路径的 $A_t^N \geq e^{-R_r t}$, 便有

$$\mathbb{E}|\nu_t^N(\varphi) - \theta|^2 \leq \frac{(4 + 32R_r t)e^{4R_r t} K^2}{N},$$

这就证明所述界。整个证明仅在初始时刻使用独立性, 没有把复制后的粒子视为独立。 $\square$

以上结论针对固定时刻、固定观测量, 给出了从真实粒子操作到确定性反应方程的完整极限。加入输运和扩散后, 路径上的指数权重仍然有效; 下一小节把三种操作放进同一个测度方程中证明。

8.4 三种操作同时使用：带权随机粒子的完整对应

下面给出一个无需处理相互作用闭合问题的组合实现：先独立生成位置轨道，沿每条轨道累积权重，再在当前时刻归一化。这个实现同时包含输运、扩散和生灭，而且可以直接证明它的粒子极限。

定理 8.10 (带权扩散粒子、反应方程与粒子数误差)。 设 b, Σ 在 $[0, T] \times \mathbb{R}^d$ 上连续、有界，且对空间一致全局 Lipschitz， r 连续并满足 $|r| \leq R_r$ 。初值 $X_0^1, \dots, X_0^N$ 独立同分布于 $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ ，Brownian motion $W^1, \dots, W^N$ 相互独立且独立于初值。各 X^i 解式 (6.1)。定义

$$\begin{aligned} H_t^i &= \exp\left(\int_0^t r(s, X_s^i) ds\right), & Z_t^N &= \frac{1}{N} \sum_{i=1}^N H_t^i, \\ w_t^i &= \frac{H_t^i}{\sum_j H_t^j}, & \mu_t^N &= \sum_{i=1}^N w_t^i \delta_{X_t^i}. \end{aligned} \quad (8.20)$$

则对每个 $\varphi \in C_c^\infty(\mathbb{R}^d)$ ，有限 N 的精确方程为

$$\mu_t^N(\varphi) - \mu_0^N(\varphi) = \int_0^t \left[\mu_s^N(\mathcal{L}_s \varphi) + \mu_s^N(r_s \varphi) - \mu_s^N(r_s) \mu_s^N(\varphi) \right] ds + M_t^{N,w}(\varphi), \quad (8.21)$$

$$M_t^{N,w}(\varphi) = \sum_{i=1}^N \int_0^t w_s^i \nabla \varphi(X_s^i)^\top \Sigma(s, X_s^i) dW_s^i. \quad (8.22)$$

若 $C_\varphi = \sup_{s,x} \|\Sigma(s, x)^\top \nabla \varphi(x)\|$ ，则

$$\mathbb{E} \sup_{t \leq T} |M_t^{N,w}(\varphi)|^2 \leq \frac{4Te^{2R_r T} C_\varphi^2}{N}. \quad (8.23)$$

令 X 为同一 SDE 的一个副本， $H_t = \exp(\int_0^t r(s, X_s) ds)$ ，并定义

$$\gamma_t(\varphi) = \mathbb{E}[H_t \varphi(X_t)], \quad Z_t = \mathbb{E} H_t, \quad \mu_t = \gamma_t / Z_t. \quad (8.24)$$

则 μ 满足弱方程 (8.1)。对每个固定 $t \in [0, T]$ 和有界 Borel 函数 φ ，还有

$$\mathbb{E} |\mu_t^N(\varphi) - \mu_t(\varphi)|^2 \leq \frac{4e^{4R_r T} \|\varphi\|_\infty^2}{N}. \quad (8.25)$$

因而每个固定有界测试读数均在 L^2 中收敛到所述 PDE 的测度解。

证明直觉 位置的 Itô 公式提供输运与扩散；权重的普通乘法法则提供生灭。分母虽然随机，却只通过时间积分变化，所以没有额外的二次变差修正。证明粒子极限时，先看归一化之前的独立路径，再用分母确定下界控制比值。不需要把归一化后的权重误认为相互独立，也不需要把随机比值的期望换成期望的比值。

证明. 第一步：有限粒子的精确随机方程。 各 H^i 连续且为有限变差过程，满足

$$dH_t^i = H_t^i r(t, X_t^i) dt, \quad e^{-R_r T} \leq H_t^i \leq e^{R_r T}.$$

所以 Z^N 也具有有限变差，且

$$dZ_t^N = Z_t^N \mu_t^N(r_t) dt, \quad dw_t^i = w_t^i (r(t, X_t^i) - \mu_t^N(r_t)) dt.$$

应用乘法法则，有

$$d(w_t^i \varphi(X_t^i)) = w_t^i d\varphi(X_t^i) + \varphi(X_t^i) dw_t^i.$$

其中没有交叉变差项，因为 w^i 具有有限变差。代入 Itô 公式，再对 i 求和，得到式 (8.21)。

第二步：控制不能直接删去的鞅项。不同 Brownian 坐标的交叉变差为零；即使被积系数 w^i 依赖整个粒子群，这个正交性也仍然成立。因此

$$d\langle M^{N,w}(\varphi) \rangle_t = \sum_i (w_t^i)^2 \left\| \Sigma(t, X_t^i)^\top \nabla \varphi(X_t^i) \right\|^2 dt.$$

由 $w_t^i \leq e^{2R_r T}/N$ 以及 $\sum_i w_t^i = 1$ ， $\sum_i (w_t^i)^2 \leq e^{2R_r T}/N$ 。故该鞅平方可积，Itô 等距与 Doob 最大不等式给出式 (8.23)。

第三步：识别确定性极限方程。对一个副本计算 $H_t \varphi(X_t)$ 。有限变差乘法法则与 Itô 公式给出

$$d(H_t \varphi(X_t)) = H_t(\mathcal{L}_t \varphi + r_t \varphi)(X_t) dt + H_t \nabla \varphi(X_t)^\top \Sigma(t, X_t) dW_t.$$

有界性保证随机积分平方可积且均值为零，时间积分可与期望交换。于是 $\frac{d}{dt} \gamma_t(\varphi) = \gamma_t(\mathcal{L}_t \varphi + r_t \varphi)$ ， $Z'_t = Z_t \mu_t(r_t)$ 。对 $\gamma_t(\varphi)/Z_t$ 求导，即得式 (8.1)。有界连续测试函数的窄连续性由路径连续性与控制收敛得到。

第四步：真正证明粒子数趋于无穷的收敛。固定 t 和 φ ，令 $m = \mu_t(\varphi)$ ，并写成一个中心化的独立样本平均：

$$\mu_t^N(\varphi) - m = \frac{N^{-1} \sum_i H_t^i(\varphi(X_t^i) - m)}{Z_t^N}.$$

各项 $Y_i = H_t^i(\varphi(X_t^i) - m)$ 独立同分布， $\mathbb{E} Y_i = \gamma_t(\varphi) - m Z_t = 0$ ，且 $|Y_i| \leq 2e^{R_r T} \|\varphi\|_\infty$ 。分母逐路径满足 $Z_t^N \geq e^{-R_r T}$ ，故

$$\mathbb{E} |\mu_t^N(\varphi) - m|^2 \leq e^{2R_r T} \mathbb{E} \left| \frac{1}{N} \sum_i Y_i \right|^2 = \frac{e^{2R_r T}}{N} \mathbb{E} Y_1^2 \leq \frac{4e^{4R_r T} \|\varphi\|_\infty^2}{N}.$$

这证明了所需收敛，不曾假设归一化后的随机权重独立。□

为什么也可以在经验弱方程中严格取极限？对 $\varphi \in C_c^\infty$ ， $\mathcal{L}_s \varphi$ 、 $r_s \varphi$ 与 r_s 的上界均对 s 一致。式 (8.25) 对每个确定的测试函数都成立，也可在固定 s 用于这些随 s 确定变化的函数。对非线性乘积，直接使用

$$\begin{aligned} & |\mu_s^N(r_s) \mu_s^N(\varphi) - \mu_s(r_s) \mu_s(\varphi)| \\ & \leq R_r |\mu_s^N(\varphi) - \mu_s(\varphi)| + \|\varphi\|_\infty |\mu_s^N(r_s) - \mu_s(r_s)|. \end{aligned}$$

由 Cauchy-Schwarz 和 Fubini，时间积分中每一项的 L^1 误差都是 $O(N^{-1/2})$ ；鞅项由式 (8.23) 在 L^2 的时间一致范数中消失。因此经验弱方程确实收敛到确定性的弱方程。这里得到的是逐个测试函数的结论，并非所有有界测试函数上同时成立的全变差收敛。

怎样得到经典密度 PDE？若 $\mu_t(dx) = \rho_t(x) dx$ ，对任意 $\psi \in C_c^\infty([0, T] \times \mathbb{R}^d)$ ，弱方程等价于

$$\int \psi(0, x) \mu_0(dx) + \int_0^T \int \rho_t(x) [\partial_t \psi + \mathcal{L}_t \psi + (r_t - \mu_t(r_t)) \psi] (t, x) dx dt = 0. \quad (8.26)$$

输运项对空间分部积分一次，扩散项两次，生灭项不移动导数，分别给出式 (8.2) 中的一阶、二阶和零阶项。若这些分布导数由连续函数表示，就得到逐点成立的 PDE。不能仅凭粒子构造，省略密度存在和可微性的假设。

三个操作各自留下什么？ 输运留下普通链式法则中的一阶项，扩散留下二次变差产生的二阶项，生灭留下权重乘法法则中的零阶项；概率归一化再扣去平均增长率。有限粒子是否精确满足确定性方程，要看是否仍有鞅项。对独立带权轨道，上面的方差计算已经给出了完整的固定测试函数粒子极限；对固定粒子数的复制替换，纯生灭的证明见上一小节。

上述组合定理处理给定、有界的 $r(t, x)$ ，没有把 $r[\rho](x) = -\lambda \log(\rho(x)/\pi(x))$ 自动包含进来。密度依赖的生灭需要证明非线性方程的适定性，以及密度估计与粒子相互作用的误差。这种区分使我们能准确地知道：哪些对应已在本讲中证明，哪些额外结论还需要新的假设。

9 扩散模型：从加噪路径学习反向动力学

前面把粒子运动写成了分布的演化方程。现在反过来使用这件事：先选一条容易构造的分布路径，再寻找能够沿着它生成数据的粒子动力学。扩散模型选择的路径来自一个已知的加噪过程。我们可以从数据产生任意中间时刻的训练样本，随后学习把这条分布路径反向走过所需要的场。这里的主要参考是 Song 等人的 SDE 框架[10, 第 3 节与附录 A、D]。

先区分三个对象。 一条样本轨迹 $t \mapsto X_t$ 、每个时刻的边缘密度 p_t 、以及两个时刻之间的条件转移核，携带的信息不同。Fokker Planck 方程描述边缘密度；真正的时间反演还必须确定反向条件转移。同一条密度路径可以由不同的 ODE 或 SDE 产生。

9.1 线性加噪：不必模拟整条轨迹就能抽取训练样本

令 $X_0 \sim \mu_{\text{data}} \in \mathcal{P}_2(\mathbb{R}^d)$ ，且 X_0 与 Brownian 运动 W 独立。考虑已知的线性正向 SDE

$$dX_t = f_t X_t dt + g_t dW_t, \quad 0 \leq t \leq T. \quad (9.1)$$

其中 f_t, g_t 是确定性连续函数， $g_t > 0$ 。定义

$$\alpha_t = \exp\left(\int_0^t f_r dr\right), \quad \sigma_t^2 = \alpha_t^2 \int_0^t \frac{g_r^2}{\alpha_r^2} dr. \quad (9.2)$$

这里 σ_t 是给定 X_0 后新增噪声的标准差，不是边缘分布 p_t 的总标准差。

命题 9.1 (线性 SDE 的高斯条件转移)。 对每个固定的 $t > 0$ ，有

$$X_t \mid X_0 = x_0 \sim \mathcal{N}(\alpha_t x_0, \sigma_t^2 \text{Id}_d). \quad (9.3)$$

因此可以通过 $X_0 \sim \mu_{\text{data}}$ 、独立的 $\epsilon \sim \mathcal{N}(0, \text{Id}_d)$ ，直接抽取

$$X_t \stackrel{\text{law}}{=} \alpha_t X_0 + \sigma_t \epsilon. \quad (9.4)$$

证明直觉 先用积分因子去掉确定性的缩放，剩下的是确定性函数对 Brownian 运动的积分。这个积分是高斯变量，方差由 Itô 等距公式给出。训练只需要某个随机时刻的联合样本 (X_0, X_t) ，因此有了这个条件分布，就不必逐步模拟正向 SDE。

证明. 由于 $\dot{\alpha}_t = f_t \alpha_t$, Itô 公式给出

$$d(\alpha_t^{-1} X_t) = \alpha_t^{-1} g_t dW_t.$$

积分后得到

$$X_t = \alpha_t X_0 + \alpha_t \int_0^t \alpha_r^{-1} g_r dW_r.$$

第二项与 X_0 独立, 均值为零, 协方差矩阵为 $\alpha_t^2 \int_0^t \alpha_r^{-2} g_r^2 dr \text{Id}_d$, 即 $\sigma_t^2 \text{Id}_d$. 这就给出了条件转移和直接抽样公式. $\square$

式(9.4)是一条固定时刻的分布等式。把同一个 ϵ 固定后对所有 t 都使用这个式子, 通常不会得到原 SDE 的整条轨迹。这个区别稍后还会在 Flow Matching 中出现。

记高斯条件密度为

$$k_t(x | x_0) = (2\pi\sigma_t^2)^{-d/2} \exp\left(-\frac{\|x - \alpha_t x_0\|^2}{2\sigma_t^2}\right).$$

即使 μ_{data} 没有密度, $t > 0$ 时也有光滑、严格为正的边缘密度

$$p_t(x) = \int k_t(x | x_0) \mu_{\text{data}}(dx_0). \quad (9.5)$$

由定理 6.3, 它满足

$$\partial_t p_t = -\nabla \cdot (b_t p_t) + \frac{g_t^2}{2} \Delta p_t, \quad b_t(x) = f_t x. \quad (9.6)$$

例 9.2 (贯穿本节的一维例子)。取数据分布 $\mathcal{N}(2, 1/4)$, 正向过程为

$$dX_t = -X_t dt + \sqrt{2} dW_t.$$

此时

$$\alpha_t = e^{-t}, \quad \sigma_t^2 = 1 - e^{-2t}, \quad p_t = \mathcal{N}(m_t, c_t), \quad m_t = 2e^{-t}, \quad c_t = 1 - \frac{3}{4}e^{-2t}. \quad (9.7)$$

数据的非零均值逐渐衰减, 方差逐渐趋向 1。对于任何有限的 T , 均值 $2e^{-T}$ 仍不等于零, 所以 p_T 并非精确的标准正态。用标准正态作为采样起点, 是对这个终端边缘分布的近似。

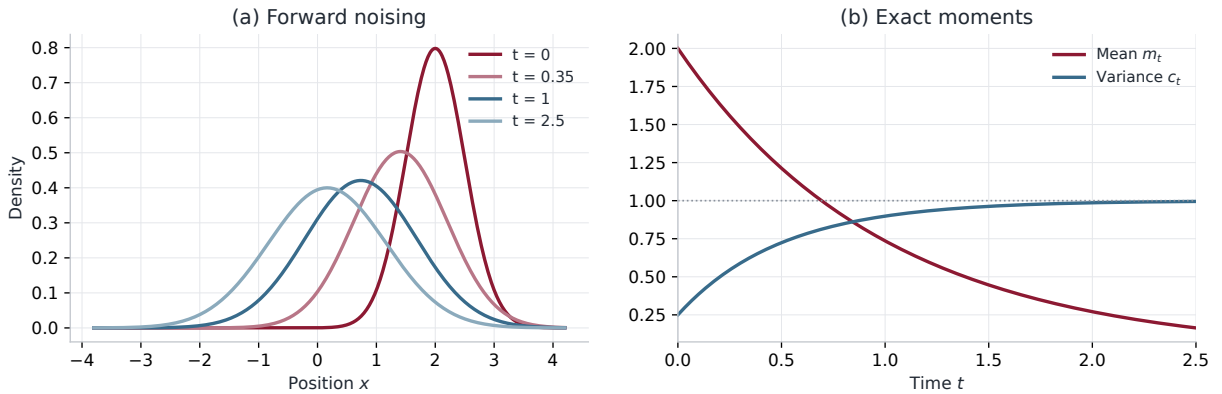


图 2: 从 $\mathcal{N}(2, 1/4)$ 出发的正向 OU 加噪。各时刻的密度、均值与方差均由式(9.7)给出。有限时刻仍保留原始数据的均值和方差信息, 标准正态是长时间极限。

9.2 Score: 把未知边缘的梯度写成可监督的条件期望

定义 9.3 (密度的 score)。对于正密度 p_t , 定义

$$\mathbf{s}_t(x) = \nabla_x \log p_t(x).$$

它是对样本位置 x 求梯度。沿着 $\mathbf{s}_t(x)$ 的方向作步长 $h > 0$ 的小移动, $\log p_t$ 的一阶变化是 $h \|\mathbf{s}_t(x)\|^2$, 在 score 非零时为正。

直接计算 p_t 需要对整个数据分布积分, 然而给定 X_0 后的 score 是已知的:

$$\nabla_x \log k_t(x | X_0) = -\frac{x - \alpha_t X_0}{\sigma_t^2}. \quad (9.8)$$

下面的等式说明, 回归这个可计算目标就能学习边缘 score。

命题 9.4 (Denoising score matching 是条件期望投影)。固定 $t > 0$, 令

$$H_t = \nabla_x \log k_t(X_t | X_0) = -\epsilon / \sigma_t.$$

则

$$\mathbf{s}_t(x) = \mathbb{E}[H_t | X_t = x] = -\frac{1}{\sigma_t} \mathbb{E}[\epsilon | X_t = x]. \quad (9.9)$$

对任意平方可积向量场 h , 有风险分解

$$\mathbb{E} \|h(X_t) - H_t\|^2 = \mathbb{E} \|h(X_t) - \mathbf{s}_t(X_t)\|^2 + \mathbb{E} \|H_t - \mathbf{s}_t(X_t)\|^2. \quad (9.10)$$

第二项与 h 无关, 因此总体回归问题的最优解是 $\mathbf{s}_t$, 唯一性按 p_t 几乎处处相等理解。

证明直觉 一次加噪同时告诉我们干净样本 X_0 、噪声 ϵ 和结果 X_t 。生成时只能看到当前状态 X_t , 因此不能使用那个不可见的 X_0 。平方损失会自动把所有可能的干净来源, 按它们在当前状态下的后验概率平均起来。这正是第一讲中“条件均值是平方损失最优预测器”的向量版本。

证明. 固定 $t > 0$ 时, 高斯核及其关于 x 的各阶导数均有可积的控制, 可以对式(9.5)在积分号下求导。于是

$$\begin{aligned} \nabla \log p_t(x) &= \frac{\int \nabla_x k_t(x | x_0) \mu_{\text{data}}(dx_0)}{p_t(x)} \\ &= \int \nabla_x \log k_t(x | x_0) \frac{k_t(x | x_0) \mu_{\text{data}}(dx_0)}{p_t(x)}. \end{aligned}$$

最后一行中的概率权重正是 $X_0 | X_t = x$ 的条件分布, 得到(9.9)。而 $\mathbb{E} \|H_t\|^2 = d/\sigma_t^2 < \infty$, 条件期望的平方可积性也随之成立。

写成

$$h(X_t) - H_t = [h(X_t) - \mathbf{s}_t(X_t)] + [\mathbf{s}_t(X_t) - H_t].$$

展开平方后, 交叉项为零, 因为给定 X_t , 第一项已经确定, 第二项的条件期望为零。这就证明了风险分解。□

同一等式也给出去噪公式

$$\mathbb{E}[X_0 | X_t = x] = \frac{x + \sigma_t^2 \mathbf{s}_t(x)}{\alpha_t}. \quad (9.11)$$

因此预测噪声、预测干净样本和预测 score 之间有明确的代数联系。这里的条件均值通常不是产生当前样本的那个唯一来源; 同一个带噪位置可能来自很多不同的干净样本。

9.3 反向时钟：为什么漂移中会出现一整份 score 修正

先允许一般的漂移 $b_t(x)$ ，仍令 g_t 只依赖时间。统一采用向前增加的反向时钟

$$s = T - t, \quad Y_s = X_{T-s}, \quad q_s(x) = p_{T-s}(x). \quad (9.12)$$

这样所有 SDE 中的 ds 都是正向时间增量，不需要把 Brownian 方差写成负数。由 $\partial_s q_s = -\partial_t p_t$ ，反向边缘应满足

$$\partial_s q_s = \nabla \cdot (b_{T-s} q_s) - \frac{g_{T-s}^2}{2} \Delta q_s.$$

如果反向过程仍具有噪声强度 g_{T-s} ，它自己的 Fokker Planck 方程却必须是

$$\partial_s q_s = -\nabla \cdot (\bar{b}_s q_s) + \frac{g_{T-s}^2}{2} \Delta q_s.$$

把两式对齐，并使用 $q_s \mathbf{s}_{T-s} = \nabla q_s$ ，得到候选漂移

$$\bar{b}_s(x) = -b_{T-s}(x) + g_{T-s}^2 \mathbf{s}_{T-s}(x). \quad (9.13)$$

证明直觉 正向过程的概率流为 $b_t p_t - (g_t^2/2) \nabla p_t$ 。反向时希望把这个总概率流取负，但新的 SDE 仍然有一份正向扩散。因此漂移要用一整份 $g_t^2 \nabla \log p_t$ 修正，才能同时抵消原来的扩散流并补偿新加入的扩散流。

直接代入可核验

$$-\nabla \cdot ((-b_t + g_t^2 \mathbf{s}_t) p_t) + \frac{g_t^2}{2} \Delta p_t = \nabla \cdot (b_t p_t) - \frac{g_t^2}{2} \Delta p_t.$$

在对应方程适定且初始分布为 p_T 时，这说明候选 SDE 的边缘密度为 p_{T-s} 。到这里，还没有证明它是原过程的完整时间反演。一个边缘密度方程没有告诉我们相邻两个时刻如何耦合；例如加上满足 $\nabla \cdot (q_s w_s) = 0$ 的漂移，也不会改变这个边缘方程。下面用 Bayes 公式把反向转移核本身找出来。

9.4 从 Bayes 短时核严格识别反向生成元

本小节完整证明线性正向过程(9.1)的反演。为使短时展开直接成立，假设 $f, g \in C^1([0, T])$ 且 $g > 0$ 。固定 $\delta > 0$ ，先在 $0 \leq s \leq T - \delta$ 上工作。由于对应的正向时刻 $t \geq \delta$ ，高斯平滑保证 p_t 光滑且正，所有局部求导都有意义。一般扩散的更弱假设可参阅 Anderson 与 Haussmann、Pardoux [8, 9]。

引理 9.5 (线性高斯短时转移的伴随展开)。令 $k_{t-h,t}(z, x)$ 为正向过程从 $t-h$ 到 t 的转移密度，并定义

$$P_{t-h,t}^* r(x) = \int k_{t-h,t}(z, x) r(z) dz.$$

对 $r \in C_c^3(\mathbb{R}^d)$ ，当 $h \downarrow 0$ 时，局部一致地有

$$P_{t-h,t}^* r = r + h \mathcal{L}_t^* r + o(h), \quad \mathcal{L}_t^* r = -\nabla \cdot (f_t x r) + \frac{g_t^2}{2} \Delta r. \quad (9.14)$$

证明. 在区间 $[t-h, t]$ 上重做积分因子计算, 令

$$A_h = \exp \left(\int_{t-h}^t f_r \, dr \right), \quad S_h^2 = \int_{t-h}^t \exp \left(2 \int_r^t f_u \, du \right) g_r^2 \, dr.$$

则 $k_{t-h,t}(z, x)$ 是均值 $A_h z$ 、协方差 $S_h^2 \text{Id}_d$ 的高斯密度。在积分中作变量替换, 得到含 Jacobian 的精确公式

$$P_{t-h,t}^* r(x) = A_h^{-d} \mathbb{E} [r(A_h^{-1}(x - S_h \epsilon))]. \quad (9.15)$$

这里 $\epsilon \sim \mathcal{N}(0, \text{Id}_d)$ 。注意前面的 A_h^{-d} 不能省略: 这个算子在输运密度, 空间的缩放同时改变了单位体积中的质量。

由连续可微性, $A_h = 1 + hf_t + o(h)$ 、 $S_h^2 = hg_t^2 + o(h)$ 。记 $D_h = A_h^{-1}(x - S_h \epsilon) - x$, 则对紧集中的 x 一致有

$$\mathbb{E} D_h = -hf_t x + o(h), \quad \mathbb{E}[D_h D_h^\top] = hg_t^2 \text{Id}_d + o(h), \quad \mathbb{E} \|D_h\|^3 = O(h^{3/2}).$$

Taylor 展开并取期望, 给出

$$\mathbb{E} r(x + D_h) = r(x) - hf_t \langle x, \nabla r(x) \rangle + \frac{hg_t^2}{2} \Delta r(x) + o(h).$$

再乘 $A_h^{-d} = 1 - dhf_t + o(h)$, 一次项合并成 $-hf_t \{dr + x \cdot \nabla r\} + hg_t^2 \Delta r/2$, 即为所需的伴随生成元。□

定理 9.6 (真正的反向过程及其生成元)。 在上述假设下, $Y_s = X_{T-s}$ 相对于自己的过去信息 $\mathcal{G}_s = \sigma(Y_u : 0 \leq u \leq s)$ 是 Markov 过程; 以下对该信息流取通常增广。在 $0 \leq s \leq T - \delta$ 上, 它满足

$$dY_s = [-f_{T-s} Y_s + g_{T-s}^2 s_{T-s}(Y_s)] \, ds + g_{T-s} \, d\bar{W}_s, \quad Y_0 \sim p_T, \quad (9.16)$$

其中 $\bar{W}$ 是相对于反向信息流 $\mathcal{G}_s$ 的 Brownian 运动。因此这里识别的是原过程反演后的路径分布, 不只是某条具有相同边缘的路径。

证明直觉 已知后来位置 x , 较早位置 z 的可能性, 正比于“较早时刻本来处在 z 的概率”乘“从 z 走到 x 的概率”。这就是 Bayes 公式。密度 p_{t-h} 的变化会产生一个修正项; 它恰好消去生成元作用在常数上的部分, 留下反向漂移中的 score。

证明. **第一步: 先确定完整的反向转移核。** 固定 $t = T - s$ 。Bayes 公式给出

$$\bar{k}_{s,s+h}(x, z) = \frac{p_{t-h}(z) k_{t-h,t}(z, x)}{p_t(x)}. \quad (9.17)$$

分母为正, 积分后等于 1。对于任意多个时刻, 把这些反向核与 p_T 相乘, 分子分母中的各时刻密度依次约掉, 所得恰为正向 Markov 联合密度按逆序书写的结果。因此这些核不仅给出二时刻条件分布, 也给出全部有限维联合分布; 反向过程相对于 $\mathcal{G}_s$ 的 Markov 性由此得到。

第二步: 让反向时间间隔缩小。 对 $\varphi \in C_c^\infty(\mathbb{R}^d)$, 反向转移算子满足

$$\bar{P}_{s,s+h} \varphi(x) = \frac{P_{t-h,t}^*(p_{t-h} \varphi)(x)}{p_t(x)}.$$

由正向 Fokker Planck 方程, $p_{t-h} = p_t - h\mathcal{L}_t^*p_t + o(h)$. 将它与引理 9.5 合用, 得到

$$P_{t-h,t}^*(p_{t-h}\varphi) = p_t\varphi + h\{\mathcal{L}_t^*(p_t\varphi) - \varphi\mathcal{L}_t^*p_t\} + o(h),$$

$$\bar{\mathcal{L}}_s\varphi := \lim_{h \downarrow 0} \frac{\bar{P}_{s,s+h}\varphi - \varphi}{h} = \frac{\mathcal{L}_t^*(p_t\varphi) - \varphi\mathcal{L}_t^*p_t}{p_t}.$$

这里 $p_t\varphi$ 有紧支撑, 且在 $t \geq \delta$ 的紧时间区间上导数受控, 所以前一引理及时间方向的展开都可以局部一致地使用。

现在逐项展开, 而不是猜测系数:

$$-\nabla \cdot (b_t p_t \varphi) + \varphi \nabla \cdot (b_t p_t) = -p_t b_t \cdot \nabla \varphi,$$

$$\Delta(p_t \varphi) - \varphi \Delta p_t = 2\nabla p_t \cdot \nabla \varphi + p_t \Delta \varphi.$$

从而

$$\bar{\mathcal{L}}_s\varphi = [-b_t + g_t^2 \nabla \log p_t] \cdot \nabla \varphi + \frac{g_t^2}{2} \Delta \varphi, \quad t = T - s. \quad (9.18)$$

这已经把反向局部漂移和局部协方差都识别出来。

第三步: 从生成元恢复反向 Brownian 噪声。为说明此处确实得到了 SDE, 先证明测试函数对应的鞅恒等式。固定 $s < u$, 记 $t = T - s$, $r = T - u$. Bayes 公式在整个区间上给出

$$\bar{P}_{s,u}\varphi(x) = \frac{1}{p_t(x)} \int k_{r,t}(z, x) p_r(z) \varphi(z) \, dz.$$

显式高斯核满足后向方程 $\partial_r k_{r,t} = -\mathcal{L}_r^* k_{r,t}$. 这一点也可直接核验: 它的均值系数 $A_{r,t}$ 与方差 $S_{r,t}^2$ 满足 $\partial_r A_{r,t} = -f_r A_{r,t}$, $\partial_r S_{r,t}^2 = -A_{r,t}^2 g_r^2$, 对高斯密度求导便得到该方程。于是对上式关于 u 求导, 并对 z 分部积分, 得

$$\begin{aligned} \partial_u \bar{P}_{s,u}\varphi(x) &= \frac{1}{p_t(x)} \int k_{r,t}(z, x) [\mathcal{L}_r^*(p_r\varphi) - \varphi\mathcal{L}_r^*p_r](z) \, dz \\ &= \bar{P}_{s,u}(\bar{\mathcal{L}}_u\varphi)(x). \end{aligned}$$

由于 φ 有紧支撑, 求导和分部积分只涉及紧集中的受控系数。在 $t \geq \delta$ 的时间区间上, $\bar{\mathcal{L}}_u\varphi$ 也一致有界。积分后让下端点趋于 s , 得到

$$\bar{P}_{s,u}\varphi - \varphi = \int_s^u \bar{P}_{s,r}(\bar{\mathcal{L}}_r\varphi) \, dr.$$

应用 Markov 性, 右边是在给定 Y_s 后未来生成元积分的条件期望, 故 $\varphi(Y_s) - \varphi(Y_0) - \int_0^s \bar{\mathcal{L}}_r\varphi(Y_r) \, dr$ 是鞅。现在把过程停在大球内, 用带截断的坐标函数 x_i 和乘积 $x_i x_j$, 再让截断球增大, 得到连续局部鞅

$$M_s = Y_s - Y_0 - \int_0^s \bar{b}_r(Y_r) \, dr, \quad \langle M^i, M^j \rangle_s = \delta_{ij} \int_0^s g_{T-r}^2 \, dr.$$

这里 $\langle M^i, M^j \rangle$ 表示连续局部鞅的二次协变差: 它是从零开始的连续有限变差过程, 使 $M^i M^j - \langle M^i, M^j \rangle$ 成为局部鞅。它也可由细分上的增量乘积之和在概率意义下取极限得到。上面对坐标乘积 $x_i x_j$ 的测试, 正是识别这个补偿项的方法。

此处所需的积分和 Itô 公式是前文 Brownian 版本的直接延伸。对简单适应系数仍定义 $\int H \, dM = \sum_i H_i \Delta_i M$; 先停止为平方可积鞅, 同样利用不同时间段的正交性, 得到等距

$$\mathbb{E} \left\| \int_0^s H_r \, dM_r \right\|^2 = \mathbb{E} \int_0^s \text{Tr}(H_r \, d\langle M \rangle_r H_r^\top).$$

按右侧范数取闭包，再作局部化，就定义了这里的随机积分。Taylor 展开中将 Brownian 的 $\delta_{ij} \, dr$ 换成 $d\langle M^i, M^j \rangle_r$ ，便得到同样的 Itô 公式；一阶项是上述积分，二阶项是二次协变差极限，余项仍由二阶导数的连续模控制。因此下文既可对 M 积分，也可对由它构造的连续鞅使用 Itô 公式。

原过程连续，所以反向过程在这个有限时间区间内也有界于某个随机大球；局部化不会遗漏任何有限时刻。令

$$\bar{W}_s = \int_0^s g_{T-r}^{-1} \, dM_r.$$

它是连续局部鞅，且 $\langle \bar{W}^i, \bar{W}^j \rangle_s = \delta_{ij}s$ 。这一性质确实刻画 Brownian 运动：对任意 $\xi \in \mathbb{R}^d$ ，Itô 公式说明 $\exp\{i\langle \xi, \bar{W}_s \rangle + \|\xi\|^2 s/2\}$ 是鞅；其模在固定时间区间上有确定的上界，故可移去局部化。于是增量的条件特征函数为

$$\mathbb{E} \left[e^{i\langle \xi, \bar{W}_u - \bar{W}_s \rangle} \mid \mathcal{G}_s \right] = e^{-\|\xi\|^2(u-s)/2}.$$

这给出独立于过去的中心高斯增量及其方差，证明 $\bar{W}$ 是 Brownian 运动，并得到(9.16)。□

反向噪声是相对于反向信息流重新构造的 Brownian 运动。它一般不能通过把原来的 W_t 简单换成 W_{T-s} 来获得。当 μ_{data} 没有密度时， $s = T$ 对应的 score 也可能不存在；上述证明对每个 $\delta > 0$ 都成立，端点应以 $\delta \downarrow 0$ 的分布极限处理。

说明 9.7 (空间依赖扩散矩阵的额外项)。 对于 $dX_t = b_t(X_t) \, dt + \Sigma_t(X_t) \, dW_t$ ，记 $a_t = \Sigma_t \Sigma_t^\top$ 。在保证密度和反向生成元存在的正则性条件下，同一个伴随计算给出

$$\bar{b}_s^i(x) = -b_t^i(x) + \frac{1}{p_t(x)} \sum_j \partial_j (a_t^{ij}(x) p_t(x)), \quad t = T - s. \quad (9.19)$$

即 $\bar{b}_s = -b_t + \text{div } a_t + a_t s_t$ ，其中 $(\text{div } a_t)_i = \sum_j \partial_j a_t^{ij}$ 。原因是 $\frac{1}{2} \sum_{ij} \partial_{ij} (a_t^{ij} p_t \varphi)$ 的乘积展开，还要对 a_t^{ij} 求导。在本章的 $a_t = g_t^2 \text{Id}_d$ 且 g_t 不依赖位置的情形， $\text{div } a_t = 0$ ，所以没有这一额外项。

9.5 Probability flow ODE：用半份 score 修正搬运同一条密度路径

利用 $\nabla p_t = p_t s_t$ ，正向 Fokker Planck 方程可以重写成

$$\partial_t p_t = -\nabla \cdot (p_t v_t^{\text{PF}}), \quad v_t^{\text{PF}}(x) = b_t(x) - \frac{g_t^2}{2} s_t(x). \quad (9.20)$$

这已经是连续性方程。因此，在这个速度场的 ODE 适定、连续性方程的解唯一时，

$$\dot{U}_t = v_t^{\text{PF}}(U_t), \quad U_0 \sim \mu_{\text{data}} \quad (9.21)$$

具有与正向 SDE 相同的边缘密度 p_t 。若只在 $[\delta, T]$ 上有正则性，则从 $U_\delta \sim p_\delta$ 开始理解该结论。

证明直觉 ODE 没有另加一份 Brownian 扩散。它只需要把原 SDE 的总概率流直接当作输运流，因此只出现半份 $g_t^2 s_t$ 。反向 SDE 中的整份修正与这里的半份修正，对应的是两个不同的粒子模型。

这里使用定理 3.5 中的流与连续性方程对应。所用的唯一性条件可以具体理解为：速度场关于位置 C^1 ，其 Lipschitz 常数和线性增长系数在时间上可积。若 $\Phi_{r,t}$ 是 ODE 的流，对终端测试函数 φ ， $\psi_r(x) = \varphi(\Phi_{r,t}(x))$ 满足 $\partial_r \psi_r + v_r^{\text{PF}} \cdot \nabla \psi_r = 0$ 。把它代入连续性方程的弱形式，得到 $\int \psi_r \, d\text{Law}(X_r)$ 不随 r 改变，因而 $\int \varphi(x) p_t(x) \, dx = \int \varphi(\Phi_{0,t}(x)) \mu_{\text{data}}(dx)$ 。这直接证明了 ODE 推前的分布等于 p_t 。

采样时从 p_T 出发, 使用同一反向时钟 $s = T - t$, 速度为

$$\frac{dU_s}{ds} = -b_{T-s}(U_s) + \frac{g_{T-s}^2}{2} s_{T-s}(U_s). \quad (9.22)$$

对于一般空间依赖的 a_t , 正向 probability flow 速度则为 $b_t - \frac{1}{2} \operatorname{div} a_t - \frac{1}{2} a_t s_t$. 这与说明 9.7 中的整份修正相对应。

例 9.8 (用平稳 OU 过程检查全部符号). 若 $p_t = \mathcal{N}(0, 1)$ 且 $b(x) = -x$, $g^2 = 2$, 则 $s(x) = -x$. 反向 SDE 的漂移为 $x + 2(-x) = -x$, 仍然是平稳 OU 过程. 正向和反向 probability flow ODE 的速度都为零. 这三个过程每个时刻都是标准正态, 但路径并不相同: ODE 的样本保持不动, OU 样本仍在运动. 更精确地, 平稳 OU 的两时刻协方差为 $e^{-|t-u|}$, 而静止 ODE 的协方差恒为 1. 因此, 仅仅检查每个时刻的密度, 无法判断路径是否真的被反演。

对于例 9.2 的非平稳数据分布, $s_t(x) = -(x - m_t)/c_t$, 所以正向 probability flow 速度为

$$v_t^{\text{PF}}(x) = -m_t + \frac{1 - c_t}{c_t}(x - m_t).$$

其解可以显式写成

$$U_t = m_t + \sqrt{\frac{c_t}{c_0}}(U_0 - m_0), \quad U_0 \sim \mathcal{N}(2, 1/4). \quad (9.23)$$

对右侧求导, 使用 $\dot{m}_t = -m_t$, $\dot{c}_t = 2(1 - c_t)$ 即可核验速度. 这条确定性路径和正向 OU 路径共享全部高斯边缘。

9.6 训练与采样: 两段计算分别使用什么信息

训练. 选择时间采样分布和正权重 $w(t)$, 训练

$$\mathcal{L}_{\text{DSM}}(\theta) = \mathbb{E}_{t, X_0, \epsilon} \left[w(t) \|s_\theta(t, \alpha_t X_0 + \sigma_t \epsilon) + \epsilon / \sigma_t\|^2 \right]. \quad (9.24)$$

一次随机梯度更新依次抽取数据 X_0 、时刻 t 和独立高斯噪声 ϵ , 形成 X_t , 再用已知目标 $-\epsilon / \sigma_t$ 做回归. 不需要计算 p_t , 也不需要在一训练更新中求解 ODE 或 SDE. 若改为预测噪声 $\epsilon_\theta = -s_\theta$, 对应的权重是 $w(t) / \sigma_t^2$, 可由平方损失直接换算。

当 $t \downarrow 0$ 时, $\sigma_t \downarrow 0$, 条件 score 的二阶矩可能发散. 为了让训练目标有定义, 可以在 $[\delta, T]$ 上抽取时间, 或选取使 $\mathbb{E}_t[w(t) / \sigma_t^2] < \infty$ 的权重. 命题 9.4 对时间积分后仍成立。

采样. 训练完成后, 抽取新的初始噪声 $Y_0 \sim \nu_T$, 其中 ν_T 通常取容易抽样的正态分布. 在反向时钟上逐步更新. 若 $s_{k+1} - s_k = h_k > 0$, $t_k = T - s_k$, Euler 型反向 SDE 更新为

$$Y_{k+1} = Y_k + h_k [-b_{t_k}(Y_k) + g_{t_k}^2 s_\theta(t_k, Y_k)] + g_{t_k} \sqrt{h_k} \xi_k, \quad \xi_k \sim \mathcal{N}(0, \text{Id}_d). \quad (9.25)$$

这里每一步的 ξ_k 都重新独立抽取. 若使用 probability flow ODE, 则每步评估 $-b_{t_k} + (g_{t_k}^2 / 2) s_\theta$, 并交给 ODE 求解器积分. 两种方式都只使用当前状态、当前时刻和训练好的网络, 不再访问原来的干净样本。

连续层面的等价结论要求真 score 与正确的初始边缘. 实际误差有三个不同来源: 以 ν_T 替换 p_T 的初始分布误差, 以 s_θ 替换 s 的场估计误差, 以及用有限步更新替换连续动力学的离散误差. 如果停在正向时刻 $\delta > 0$, 理想输出也是 p_δ , 还需与目标 μ_{data} 区分. 这些差异不能由“两个连续方程具有相同边缘”这一等式自动消除。

10 Flow Matching: 先设计分布路径, 再回归边缘速度

扩散模型从一个 SDE 得到分布路径, 再求用于生成的反向场。Flow Matching 可以直接从两端样本构造插值路径, 再学习产生该路径的速度。共同的核心仍是分布的演化方程, 而训练的回归目标由路径本身给出 [11, 12]。本节重新使用 $t \in [0, 1]$, 时间方向从噪声走向数据。这里的 α_t, β_t 表示插值系数, 与上一节的加噪时钟分别定义。

10.1 耦合与条件速度: 把很多随机路径平均成一个场

设 $\gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ 是噪声分布 μ_0 与数据分布 μ_1 的任意耦合, 抽取 $(Z, X_1) \sim \gamma$ 。选取 $\alpha, \beta \in C^1([0, 1])$, 满足

$$(\alpha_0, \beta_0) = (1, 0), \quad (\alpha_1, \beta_1) = (0, 1).$$

定义随机插值路径和它的样本速度

$$I_t = \alpha_t Z + \beta_t X_1, \quad \dot{I}_t = \dot{\alpha}_t Z + \dot{\beta}_t X_1, \quad \mu_t = \text{Law}(I_t). \quad (10.1)$$

在有密度时记 $\mu_t(dx) = \rho_t(x) dx$ 。两端分布由构造保证: $\mu_{t=0} = \mu_0$, $\mu_{t=1} = \mu_1$ 。

命题 10.1 (条件平均速度满足弱连续性方程)。定义

$$v_t(x) = \mathbb{E}[\dot{I}_t \mid I_t = x]. \quad (10.2)$$

则对每个 $\varphi \in C_c^\infty(\mathbb{R}^d)$,

$$\frac{d}{dt} \int \varphi d\mu_t = \int \langle \nabla \varphi(x), v_t(x) \rangle \mu_t(dx) \quad (10.3)$$

在时间上几乎处成立。因此 μ_t 满足弱连续性方程; 若有密度, 则写为

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0. \quad (10.4)$$

此外 $\int_0^1 \int \|v_t\|^2 d\mu_t dt \leq \mathbb{E} \int_0^1 \|\dot{I}_t\|^2 dt < \infty$ 。

证明直觉 同一个位置可能被许多训练插值路径经过, 而这些路径指向不同的终点。只观察当前位置的生成器无法区分这些隐藏来源。它需要的速度, 是所有经过当前位置的路径速度的条件平均。连续性方程只计算经过每个位置的净质量流, 这个平均恰好保留了净流。

证明. 样本路径可微, 链式法则给出

$$\frac{d}{dt} \varphi(I_t) = \langle \nabla \varphi(I_t), \dot{I}_t \rangle.$$

由于 $\nabla \varphi$ 有界、 α, β 的导数在紧区间上有界, 且 Z, X_1 有有限二阶矩, 右侧在时间与概率上可积。因此可先积分再取期望, 得到

$$\mathbb{E} \varphi(I_b) - \mathbb{E} \varphi(I_a) = \int_a^b \mathbb{E} \langle \nabla \varphi(I_t), \dot{I}_t \rangle dt.$$

给定 I_t 再取条件期望, 右侧等于 $\int_a^b \mathbb{E} \langle \nabla \varphi(I_t), v_t(I_t) \rangle dt$, 这证明了弱形式。最后由条件 Jensen 不等式, $\mathbb{E} \|v_t(I_t)\|^2 \leq \mathbb{E} \|\dot{I}_t\|^2$; 对时间积分即得能量界。□

弱连续性方程本身不保证 v_t 的 ODE 对每个初值都适定。若 v_t 关于位置 C^1 , Lipschitz 常数和线性增长系数在时间上可积, 则可用定理 3.5 的反向测试函数论证唯一性, 从而

$$\dot{U}_t = v_t(U_t), \quad U_0 \sim \mu_0 \implies \text{Law}(U_t) = \mu_t. \quad (10.5)$$

这里断言的是 U_t 与 I_t 在每个时刻的分布相同。它没有断言 ODE 粒子会跟随抽样时配给它的那条插值路径。

10.2 为什么无需知道边缘速度就能训练它

边缘速度(10.2)是未知的条件期望。但每次抽取 (Z, X_1) 和独立的 t 后, I_t 与 $\dot{I}_t$ 都可以直接计算。因此训练一个向量场 $v_\theta(t, x)$, 使用 conditional flow matching 损失

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, (Z, X_1) \sim \gamma} \left\| v_\theta(t, I_t) - \dot{I}_t \right\|^2, \quad t \sim \text{Unif}[0, 1]. \quad (10.6)$$

命题 10.2 (条件回归与边缘速度回归只差一个常数)。在上述二阶矩条件下, 对平方可积的 v_θ , 有

$$\mathcal{L}_{\text{CFM}}(\theta) = \underbrace{\mathbb{E}_t \int \|v_\theta(t, x) - v_t(x)\|^2 \mu_t(dx)}_{\mathcal{L}_{\text{FM}}(\theta)} + \underbrace{\mathbb{E} \left\| \dot{I}_t - v_t(I_t) \right\|^2}_{C_\gamma}. \quad (10.7)$$

其中 C_γ 与模型参数 θ 无关。

证明直觉 监督目标可以含有当前位置无法解释的随机性。只要这个随机目标的条件均值是所需的速度, 平方损失就会把无法解释的部分留在一个常数项里。这与 denoising score matching 的证明使用同一个 L^2 投影结构。

证明. 令 $A = v_\theta(t, I_t) - v_t(I_t)$, $B = v_t(I_t) - \dot{I}_t$. 展开 $\|A + B\|^2$ 后, 交叉项满足

$$\mathbb{E} \langle A, B \rangle = \mathbb{E} \langle A, \mathbb{E}[B | t, I_t] \rangle = 0.$$

因为给定 (t, I_t) 后 A 已确定, 而 B 的条件均值为零。其余两项分别是 $\mathcal{L}_{\text{FM}}$ 与 C_γ , 得到结论。□

因此 CFM 的总体最小化器是边缘速度, 虽然训练中从未显式计算边缘密度或条件期望。若参数化目标可微且导数可与期望交换, 两种损失的条件梯度也相同。这一结论只比较训练目标, 不会把有限网络容量或有限次优化自动变成精确的速度场。

10.3 最常用的直线插值: 条件速度为何看似在端点奇异

取独立耦合 $Z \sim \mathcal{N}(0, \text{Id}_d)$, $X_1 \sim \mu_1$, 并令

$$I_t = (1 - t)Z + tX_1, \quad \dot{I}_t = X_1 - Z. \quad (10.8)$$

给定终点 $X_1 = x_1$, 在 $t < 1$ 时有

$$I_t | X_1 = x_1 \sim \mathcal{N}(tx_1, (1 - t)^2 \text{Id}_d).$$

由 $z = (x - tx_1)/(1 - t)$, 沿着这条条件路径的速度为

$$u_t(x | x_1) = x_1 - \frac{x - tx_1}{1 - t} = \frac{x_1 - x}{1 - t}, \quad t < 1. \quad (10.9)$$

边缘速度则是

$$v_t(x) = \mathbb{E}[u_t(I_t | X_1) | I_t = x] = \frac{\mathbb{E}[X_1 | I_t = x] - x}{1 - t}. \quad (10.10)$$

平均所用的是 $X_1 | I_t = x$ 的后验权重，而不是对所有终点无差别地平均。

更一般地，当 $\alpha_t \neq 0$ 时， $I_t = \alpha_t Z + \beta_t X_1$ 的条件速度为

$$u_t(x | x_1) = \frac{\dot{\alpha}_t}{\alpha_t}(x - \beta_t x_1) + \dot{\beta}_t x_1.$$

这可以直接由插值映射求导得到，无需先解一个偏微分方程。

说明 10.3 (端点奇异与实际回归目标)。 式(10.9)在 $t = 1$ 不定义，因为给定 x_1 的条件高斯在终点收缩成一个点。但在实际训练样本 $x = I_t$ 上，分子正好等于 $(1 - t)(X_1 - Z)$ ，所以回归目标始终可以直接写成 $X_1 - Z$ ，其二阶矩有限。数值实现没有必要先形成两个很接近的数相减，再除以 $1 - t$ 。

条件场的形式奇异不代表边缘速度一定发散，下面的高斯例子在整个闭区间上都光滑。对于有原子的目标分布，则不能默认一个处处适定的可逆光滑 ODE 在有限时间把连续密度压成原子；此时需要单独处理终点极限。也可以保留终端噪声 $\sigma_{\min} > 0$ ，使用 $I_t = [1 - (1 - \sigma_{\min})t]Z + tX_1$ ，但终点分布相应变成数据与 $\mathcal{N}(0, \sigma_{\min}^2 \text{Id}_d)$ 的卷积 [11, 第 4 节]。

10.4 同边缘的两种高斯路径，以及为什么直线配对未必是 OT

继续使用数据分布 $\mu_1 = \mathcal{N}(2, 1/4)$ 。令 Z, E 独立标准正态，写成

$$X_1 = 2 + \frac{1}{2}E, \quad I_t = 2t + (1 - t)Z + \frac{t}{2}E. \quad (10.11)$$

于是

$$\rho_t = \mathcal{N}(2t, q_t), \quad q_t = (1 - t)^2 + \frac{t^2}{4} > 0. \quad (10.12)$$

我们可以把其边缘速度和 ODE 都完整算出来。

例 10.4 (训练插值与采样流有相同边缘，却有不同配对)。 对于(10.11)，

$$v_t(x) = 2 + \frac{-(1 - t) + t/4}{q_t}(x - 2t) = 2 + \frac{\dot{q}_t}{2q_t}(x - 2t). \quad (10.13)$$

以 $J_0 = Z$ 初始化其 ODE，解为

$$J_t = 2t + \sqrt{q_t} Z. \quad (10.14)$$

每个时刻 I_t 与 J_t 都服从(10.12)，但

$$\text{Cov}(I_0, I_1) = 0, \quad \text{Cov}(J_0, J_1) = \frac{1}{2}.$$

所以训练中使用的独立噪声数据配对，并未成为采样 ODE 的端点配对。

证明. 记 $A_t = I_t - 2t = (1 - t)Z + (t/2)E$ ， $B = \dot{I}_t - 2 = -Z + E/2$ 。二者联合高斯，且

$$\text{Var}(A_t) = q_t, \quad \text{Cov}(B, A_t) = -(1 - t) + t/4 = \dot{q}_t/2.$$

令 $R = B - (\dot{q}_t/(2q_t))A_t$ ，则 $\text{Cov}(R, A_t) = 0$ 。联合高斯变量的协方差为零意味着独立，故 $\mathbb{E}[R | A_t] = 0$ ，从而得到条件均值(10.13)。

对(10.14)求导,

$$\dot{J}_t = 2 + \frac{\dot{q}_t}{2\sqrt{q_t}}Z = 2 + \frac{\dot{q}_t}{2q_t}(J_t - 2t),$$

正好是该 ODE, 且 $q_0 = 1$ 保证 $J_0 = Z$ 。其每时刻的均值和方差为 $2t, q_t$, 与 I_t 相同。最后, $I_1 = 2 + E/2$ 与 $I_0 = Z$ 独立, 而 $J_1 = 2 + Z/2$, 得到两个不同的端点协方差。□

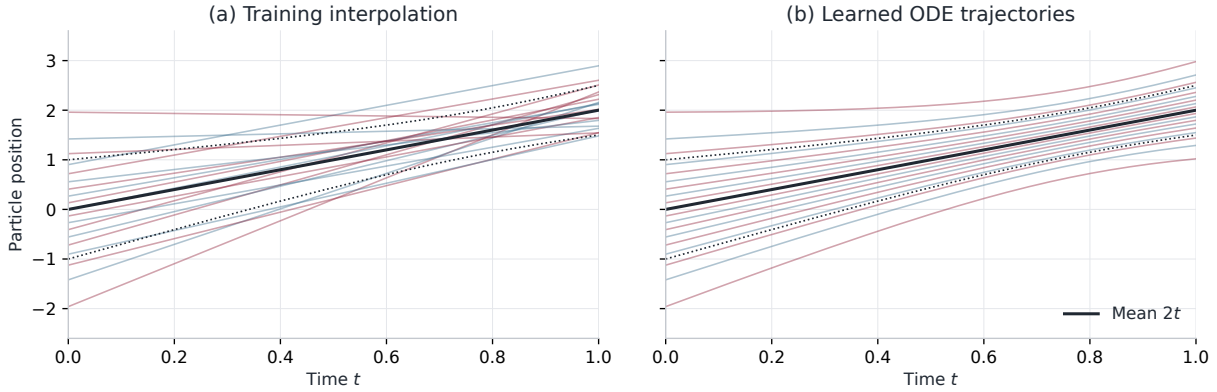


图 3: 独立配对的训练插值 I_t (左) 与条件平均速度产生的 ODE 轨迹 J_t (右)。两者的理论边缘均为 $\mathcal{N}(2t, q_t)$, 但跨时刻的耦合不同。黑实线为均值, 黑色点划线为均值上下一个标准差。彩色曲线只示意有限条轨迹; 理论同边缘不要求两组有限样本的经验分布恰好相等。

这个例子也区分了三件常被混在一起的事: 条件插值路径是直线, 抽样耦合 γ 是最优输运耦合, 以及整条边缘路径是 Wasserstein 测地线。第一件事没有自动推出后两件事。在本例中, 独立耦合的平方输运代价为

$$\mathbb{E}|X_1 - Z|^2 = 2^2 + \frac{1}{4} + 1 = \frac{21}{4}.$$

对任意具有相同两端分布的耦合, Cauchy Schwarz 给出 $\text{Cov}(Z, X_1) \leq \sqrt{1 \cdot (1/4)} = 1/2$, 所以

$$\mathbb{E}|X_1 - Z|^2 = 4 + 1 + \frac{1}{4} - 2\text{Cov}(Z, X_1) \geq \frac{17}{4}.$$

该下界被 $X_1 = 2 + Z/2$ 达到, 故独立耦合并非最优耦合。最优耦合对应的位移插值为

$$K_t = (1-t)Z + t(2 + Z/2) = 2t + (1-t/2)Z, \quad \text{Law}(K_t) = \mathcal{N}(2t, (1-t/2)^2).$$

它的中间方差与 q_t 不同; 例如 $t = 1/2$ 时分别为 $9/16$ 与 $5/16$ 。有趣的是, 本例的采样 ODE 最终得到的端点映射恰好是最优映射, 但它经过的密度路径仍不是上述位移插值。因此应分别检查耦合、边缘路径和采样流, 不能仅凭“条件轨迹是直线”就赋予整体最优输运性质。

10.5 在同一条分布路径上加入噪声: Flow Matching 与扩散的统一式

连续性方程正是扩散矩阵 $a = 0$ 的 Fokker Planck 方程。如果已经有一条光滑正密度路径 ρ_t 及其速度 v_t , 我们还可以加入 Brownian 噪声, 同时修正漂移, 使这条边缘路径保持不变。

命题 10.5 (可调扩散强度而保持相同边缘)。 假设 $\rho_t > 0$ 光滑, 满足 $\partial_t \rho_t = -\nabla \cdot (\rho_t v_t)$ 。令 $\eta_t \geq 0$ 是确定性连续函数, 定义

$$b_t^\eta(x) = v_t(x) + \eta_t \nabla \log \rho_t(x). \quad (10.15)$$

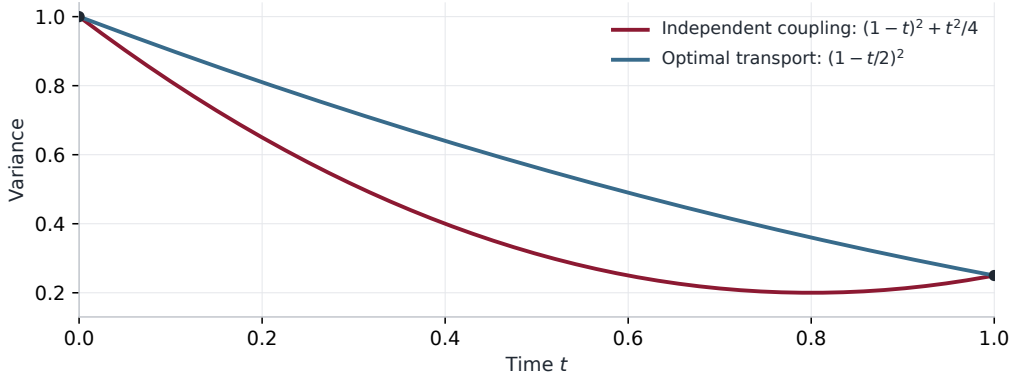


图 4: 同样连接 $\mathcal{N}(0, 1)$ 与 $\mathcal{N}(2, 1/4)$, 独立配对的插值方差 $q_t = (1-t)^2 + t^2/4$ 与最优输运位移插值的方差 $(1-t/2)^2$ 不同。端点相同并未固定中间分布路径。

则同一条 ρ_t 也满足

$$\partial_t \rho_t = -\nabla \cdot (b_t^\eta \rho_t) + \eta_t \Delta \rho_t. \quad (10.16)$$

因此, 在相应 SDE 与 Fokker Planck 方程适定的条件下,

$$dU_t^\eta = [v_t(U_t^\eta) + \eta_t \nabla \log \rho_t(U_t^\eta)] dt + \sqrt{2\eta_t} dW_t, \quad U_0^\eta \sim \rho_0, \quad (10.17)$$

对每种 η 都具有边缘密度 ρ_t 。

证明直觉 新增噪声会产生向外扩散的概率流 $-\eta_t \nabla \rho_t$ 。为了保留原来的净流, 在漂移中加入 $\eta_t \nabla \log \rho_t$, 它产生的质量流正好是 $+\eta_t \nabla \rho_t$ 。这两个新增部分抵消, 所以分布的演化不变, 粒子的轨迹却改变了。

证明. 由于 $\rho_t \nabla \log \rho_t = \nabla \rho_t$, 且 η_t 不依赖位置,

$$\begin{aligned} -\nabla \cdot ([v_t + \eta_t \nabla \log \rho_t] \rho_t) + \eta_t \Delta \rho_t &= -\nabla \cdot (v_t \rho_t) - \eta_t \Delta \rho_t + \eta_t \Delta \rho_t \\ &= -\nabla \cdot (v_t \rho_t) = \partial_t \rho_t. \end{aligned}$$

这证明了密度方程。再由相同的初始分布和方程解的唯一性, SDE 的边缘就是 ρ_t 。例如, 当漂移关于位置全局 Lipschitz、具有线性增长, 且相应系数正则时, 可以使用前文 SDE 的适定性与弱 Fokker Planck 表示来作这一识别。□

这里用 η_t 表示可调扩散强度, 以免与前面的标准高斯噪声 ϵ 混淆。常数情形 $\eta_t \equiv \epsilon$ 就是噪声 $\sqrt{2\epsilon} dW_t$ 与漂移修正 $\epsilon \nabla \log \rho_t$ 。若将时间反向为 $s = 1 - t$, 相应反向 SDE 的漂移为

$$-b_{1-s}^\eta + 2\eta_{1-s} \nabla \log \rho_{1-s} = -v_{1-s} + \eta_{1-s} \nabla \log \rho_{1-s}.$$

这与前面的反向时钟公式完全一致。这族关系是 stochastic interpolants 统一输运与扩散的一项基本结构 [12, 第 2 节]。

反过来, 给定原来的扩散 SDE, 取 $v_t = b_t - (g_t^2/2)\mathbf{s}_t$ 和 $\eta_t = g_t^2/2$, 式(10.17)就恢复原 SDE; 取 $\eta_t = 0$ 则得到 probability flow ODE。因此扩散与输运的联系可以直接在概率流层面写出。若采用非零 η 采样, 除了速度还需要这条密度路径的 score; 仅训练 v_θ 并不会自动提供它。

在一维高斯例子中，这个可调噪声过程是

$$dU_t^\eta = \left[2 + \left(\frac{\dot{q}_t}{2q_t} - \frac{\eta_t}{q_t} \right) (U_t^\eta - 2t) \right] dt + \sqrt{2\eta_t} dW_t. \quad (10.18)$$

其均值满足 $\dot{m}_t = 2$ ，方差满足

$$\frac{d}{dt} \text{Var}(U_t^\eta) = 2 \left(\frac{\dot{q}_t}{2q_t} - \frac{\eta_t}{q_t} \right) \text{Var}(U_t^\eta) + 2\eta_t.$$

代入 $\text{Var}(U_t^\eta) = q_t$ ，右侧恰为 $\dot{q}_t$ ，再次核验了每个系数。初始变量和噪声都是高斯，动力学又是线性的，故相同均值与方差已经确定了整个高斯边缘。

10.6 Flow Matching 的训练与采样分别做什么

对于独立直线插值，一次训练更新只包含以下计算：

1. 抽取数据 x_1 、独立标准高斯 z 与随机时刻 $t \in (0, 1)$ 。
2. 形成输入 $x_t = (1-t)z + tx_1$ ，以及监督目标 $u = x_1 - z$ 。
3. 计算 $\|v_\theta(t, x_t) - u\|^2$ ，对参数作随机梯度更新。

若使用其他耦合或插值，只需相应更换联合抽样方式和 $\dot{I}_t$ 。训练阶段直接抽取路径上的状态，不必反复求解当前网络定义的 ODE。

采样阶段重新抽取 $U_0 \sim \mu_0$ ，从 $t = 0$ 向 $t = 1$ 求解 $\dot{U}_t = v_\theta(t, U_t)$ 。例如一个显式更新为

$$U_{k+1} = U_k + h_k v_\theta(t_k, U_k).$$

每一步只评估当前状态处的速度，没有预先配给这条轨迹的数据终点 x_1 。若额外学习了 $\nabla \log \rho_t$ ，也可以采用(10.17)中的随机采样器。精确场保证的边缘等式，与学习误差、终点正则化和数值积分误差，仍需要分别分析。

从分布空间看两类训练。 Denoising score matching 回归的是一个条件 score，其条件均值决定反向扩散所需的场。Conditional flow matching 回归的是一条可抽样路径的速度，其条件均值决定连续性方程的质量流。两者都把不可直接计算的边缘场，变成可监督的 L^2 投影问题。训练完成后，再选择 ODE 或 SDE 将这个场变成采样算法。

11 把几何、粒子操作与学习目标连起来

11.1 同一个目标，可以通过移动质量和重新分配质量下降

前面已经看到，分布可以通过速度场改变，也可以通过中心化的生长率改变。现在把它们放进同一个最速下降问题。设 $\mathcal{F}$ 是分布上的可微泛函，第一变分为 $\xi_\rho = \delta \mathcal{F} / \delta \rho$ 。考虑瞬时变化

$$\dot{\rho} = -\nabla \cdot (\rho v) + \rho r, \quad \mathbb{E}_\rho[r] = 0. \quad (11.1)$$

零均值约束保证反应不改变总质量。在边界项消失和所有积分有限时，泛函的方向导数为

$$\dot{\mathcal{F}} = \int \xi_\rho \dot{\rho} = \mathbb{E}_\rho[\nabla \xi_\rho \cdot v + \xi_\rho r].$$

若任意增大 v 或 r 都不需要付出代价，下降速度可以被任意放大。因此，为这两种更新各加一个二次代价。

命题 11.1 (输运与生灭的联合下降方向)。设 $\rho > 0$, $\xi_\rho \in L^2(\rho)$ 且 $\nabla \xi_\rho \in L^2(\rho)$ 。在满足 $\mathbb{E}_\rho r = 0$ 的 $L^2(\rho)$ 向量场和标量场中,

$$\min_{v,r} \left\{ \mathbb{E}_\rho[\nabla \xi_\rho \cdot v + \xi_\rho r] + \frac{1}{2} \mathbb{E}_\rho[\|v\|^2 + r^2] \right\} \quad (11.2)$$

的解为

$$v = -\nabla \xi_\rho, \quad r = -(\xi_\rho - \mathbb{E}_\rho \xi_\rho).$$

若对应演化存在光滑解, 并可对泛函求导与分部积分, 则

$$\partial_t \rho = \nabla \cdot (\rho \nabla \xi_\rho) - \rho(\xi_\rho - \mathbb{E}_\rho \xi_\rho), \quad (11.3)$$

$$\frac{d}{dt} \mathcal{F}(\rho_t) = -\mathbb{E}_{\rho_t}[\|\nabla \xi_{\rho_t}\|^2] - \text{Var}_{\rho_t}(\xi_{\rho_t}). \quad (11.4)$$

证明直觉 输运可以沿空间梯度把质量往低处搬, 生灭可以把高能量位置的质量重新分给低能量位置。前者读取 $\nabla \xi$, 后者读取 ξ 相对于其平均值的偏差。这两种信息分别产生耗散中的梯度平方与方差。中心化并非人为修补, 而是“总质量必须保持为 1”这个约束产生的投影。

证明. 记 $m = \mathbb{E}_\rho \xi_\rho$ 。因为 $\mathbb{E}_\rho r = 0$, 有 $\mathbb{E}_\rho[\xi_\rho r] = \mathbb{E}_\rho[(\xi_\rho - m)r]$ 。对式 (11.2) 的两部分分别配方:

$$\begin{aligned} & \mathbb{E}_\rho[\nabla \xi_\rho \cdot v + \xi_\rho r] + \frac{1}{2} \mathbb{E}_\rho[\|v\|^2 + r^2] \\ &= \frac{1}{2} \mathbb{E}_\rho[\|v + \nabla \xi_\rho\|^2] + \frac{1}{2} \mathbb{E}_\rho[(r + \xi_\rho - m)^2] - \frac{1}{2} \mathbb{E}_\rho[\|\nabla \xi_\rho\|^2] - \frac{1}{2} \text{Var}_\rho(\xi_\rho). \end{aligned}$$

两个平方同时为零给出最优解, 且该 r 满足零均值约束。把最优解代入 (11.1) 得到演化式。再把 v, r 代回方向导数, 利用 $\mathbb{E}_\rho[\xi_\rho(\xi_\rho - m)] = \text{Var}_\rho(\xi_\rho)$, 得到耗散恒等式。□

若 $\pi = Z^{-1}e^{-V}$, 取 $\mathcal{F}(\rho) = \text{KL}(\rho|\pi)$, 则 $\xi_\rho = \log(\rho/\pi) + 1$, 常数项在空间梯度和中心化中均消失。于是 (11.3) 变成

$$\partial_t \rho = \nabla \cdot (\rho \nabla V) + \Delta \rho - \rho \left(\log \frac{\rho}{\pi} - \text{KL}(\rho|\pi) \right). \quad (11.5)$$

前两项可以用 Langevin 粒子实现, 最后一项可以用权重或克隆与替换实现。这正是生灭加速采样的一种分布层面动机[7]。耗散式说明两类更新都在降低同一个 KL; 它本身尚未给出有限粒子数下的统一收敛速率。此外, 反应率中包含 ρ , 实现时必须说明怎样估计密度比, 不能把原子经验测度直接放进对数。

11.2 只改变权重也有一个可以解完的模型

命题 11.2 (有限状态上的 KL 生灭流)。设 $p_i^0 > 0$, $\pi_i > 0$, 且 $\sum_i p_i^0 = \sum_i \pi_i = 1$, 状态个数有限。方程

$$\dot{p}_i = -p_i \left(\log \frac{p_i}{\pi_i} - \sum_j p_j \log \frac{p_j}{\pi_j} \right)$$

的解为

$$p_i(t) = \frac{(p_i^0)^{e^{-t}} \pi_i^{1-e^{-t}}}{\sum_j (p_j^0)^{e^{-t}} \pi_j^{1-e^{-t}}}. \quad (11.6)$$

它保持正性与总质量, 并在 $t \rightarrow \infty$ 时收敛到 π 。

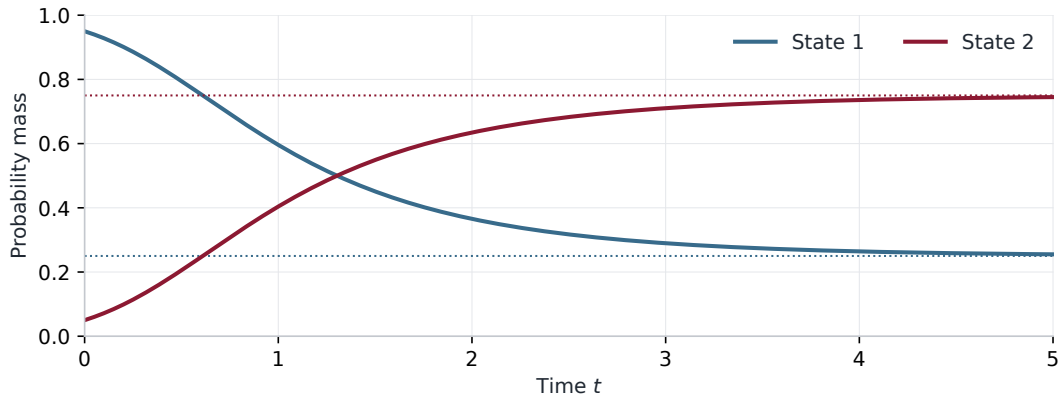


图 5: 两个固定状态上的纯生灭更新: 初始质量为 $(0.95, 0.05)$, 目标为 $(0.25, 0.75)$ 。实线由式 (11.6) 精确计算, 点线为目标质量。整个过程中状态的位置不动, 变化的是质量分配。初始正性很重要: 若某个状态初始质量为零, 仅靠乘法权重更新会一直保持为零。

证明. 令 $A_i = \log(p_i^0/\pi_i)$ 、 $\theta = e^{-t}$, 则候选解为 $p_i = \pi_i e^{\theta A_i} / \sum_j \pi_j e^{\theta A_j}$ 。对其对数求导, 利用 $\dot{\theta} = -\theta$ 得到

$$\frac{\dot{p}_i}{p_i} = -\theta A_i + \theta \sum_j p_j A_j.$$

另一方面, $\log(p_i/\pi_i) = \theta A_i - \log \sum_j \pi_j e^{\theta A_j}$, 其中心化正好为 $\theta(A_i - \sum_j p_j A_j)$, 故满足方程。初值、正性与归一化可直接检查。当 $t \rightarrow \infty$ 时 $\theta \rightarrow 0$, 有限求和中每一项收敛, 极限即为 π_i 。□

这个例子说明, 分布的改变不仅可以通过把每一个样本搬过去实现。但生灭也有边界: 在原先没有支持的位置, 它不能凭空放入粒子。把它与输运或扩散组合, 才同时拥有探索位置与调整质量的能力。

11.3 生成模型在这套语言中学到了什么?

方法	训练或设计时的输入	学到或指定的对象
变分推断与 ELBO	未归一化联合密度, 以及可重参数化的候选分布	用静态 KL 目标选择 q ; ELBO 提供可计算的目标函数。
Langevin 采样	目标势能 V 或其梯度	指定漂移 $-\nabla V$ 和噪声 $\sqrt{2}$, 让自由能下降。
带生灭的粒子法	基础动力学、评价函数或密度比估计	指定位置更新及相对增长率, 产生带反应项的 FP。
扩散模型	数据及可直接采样的加噪条件分布	回归边缘 score, 再构造反向漂移或概率流速度。
Flow Matching	两个端点的样本及选定的耦合与插值	回归条件速度的平均, 从而产生给定边缘路径。

这张表把两个设计问题分开了。泛函下降通过一个固定目标选择动力学; 生成式建模也可以先选择一条可训练的分布路径, 再寻找实现它的动力学。因此, 一般的 Flow Matching 不自动等于最优运输, 一般的生成式扩散也不自动等于某个固定 KL 的梯度流。它们共享的是分布演化的语言, 以及从局部更新产生最终样本的计算方式。

11.4 学对一个速度场以后，为什么还要分析数值误差？

为了给“网络拟合得好，所以最后样本分布也好”一个明确含义，先看一个基础稳定性估计。它同时覆盖 ODE 与使用同一加性噪声的 SDE。

命题 11.3 (从系数误差到终点分布误差)。在 $[0, T]$ 上考虑

$$dX_t = b_t(X_t) dt + G_t dW_t, \quad d\hat{X}_t = \hat{b}_t(\hat{X}_t) dt + G_t dW_t,$$

其中 G_t 仅依赖时间，两个过程使用同一个 Brownian motion，初值有有限二阶矩且与 Brownian 增量独立。假设两方程有非爆炸解及有限二阶矩， $\hat{b}_t$ 对空间一致 L Lipschitz，且

$$e_t^2 := \mathbb{E}[\|b_t(X_t) - \hat{b}_t(X_t)\|^2]$$

在时间上可积。令 $\mu_t = \text{Law}(X_t)$ 、 $\hat{\mu}_t = \text{Law}(\hat{X}_t)$ 。则可选择初始耦合使

$$W_2^2(\mu_T, \hat{\mu}_T) \leq e^{(2L+1)T} \left(W_2^2(\mu_0, \hat{\mu}_0) + \int_0^T e_t^2 dt \right). \quad (11.7)$$

$G_t = 0$ 时同一结论适用于 ODE。

证明直觉 把理想过程和近似过程放在同一个概率空间，用同一份随机噪声驱动。这样它们的差只来自初值和漂移的不同。Lipschitz 条件控制已有误差会被放大多少；训练误差控制每个时刻又注入多少新误差。

证明. 取二阶最优耦合作为初始耦合，并独立地加入 Brownian motion。记 $D_t = X_t - \hat{X}_t$ 。由于噪声系数相同且不依赖位置，随机项相消， D_t 是绝对连续的，且几乎处处有 $\dot{D}_t = b_t(X_t) - \hat{b}_t(\hat{X}_t)$ 。加减 $\hat{b}_t(X_t)$ 并用 $2uv \leq u^2 + v^2$ ，先逐路径得到

$$\begin{aligned} \frac{d}{dt} \|D_t\|^2 &\leq 2L \|D_t\|^2 + 2 \|D_t\| \|b_t(X_t) - \hat{b}_t(X_t)\| \\ &\leq (2L+1) \|D_t\|^2 + \|b_t(X_t) - \hat{b}_t(X_t)\|^2. \end{aligned}$$

先使用积分形式的 Gronwall 不等式，再取期望；非负性与 e_t^2 的可积性允许用 Tonelli 换序，得到

$$\mathbb{E} \|D_T\|^2 \leq e^{(2L+1)T} \mathbb{E} \|D_0\|^2 + \int_0^T e^{(2L+1)(T-t)} e_t^2 dt.$$

任意耦合的平均平方距离都是 W_2^2 的上界，最后再把指数统一放大到 $e^{(2L+1)T}$ 即得结论。□

对 Flow Matching, e_t^2 可以取为理想边缘分布下的速度回归误差。对反向扩散，在同一反向时钟下，漂移误差为 $g_t^2(\hat{s}_t - s_t)$ ；对概率流 ODE，相应系数为 $g_t^2/2$ 。这个 W_2 上界需要命题中的空间正则性；若端点附近的系数变得奇异，应先在截断的时间区间应用，再单独处理端点误差。它是一个基础稳定性结论，并不直接给出所有生成模型的最优统计速率。

完整实现还包含时间离散。设 $\hat{\mu}_T^h$ 是时间步长为 h 的数值算法产生的分布， $\tilde{\mu}_T$ 是使用实际初始分布、但使用理想系数的连续过程。三角不等式给出

$$W_2(\mu_T, \hat{\mu}_T^h) \leq W_2(\mu_T, \tilde{\mu}_T) + W_2(\tilde{\mu}_T, \hat{\mu}_T) + W_2(\hat{\mu}_T, \hat{\mu}_T^h). \quad (11.8)$$

三个量依次来自初始分布近似、函数学习、数值积分。这一拆分中的中间过程使用实际初值，因此若单独估计第二项，系数误差应在 $\tilde{\mu}_t$ 下计算，不能直接换成原训练路径 μ_t 下的误差。也可以

直接对 μ_T 与 $\hat{\mu}_T$ 应用上述稳定性命题：这样就在原理想路径 μ_t 下同时控制初始误差与函数误差，再加最后一项离散误差。如果还要用 N 个生成样本估计某个观测量，再加上有限样本平均的误差。这些误差依赖不同的资源：增加训练数据、增加时间步数、增加独立样本数分别控制不同的部分。附录给出 Euler 与 Euler-Maruyama 的基础误差证明。

读完本讲后应能完成的一条推导。从一个可执行的粒子更新出发，先计算测试函数的平均变化，写出生成元，再通过伴随得到密度演化方程；反过来，从所需分布路径出发，找出能实现它的速度、噪声和反应率，并辨认其中哪些量已知、哪些量要通过条件期望回归学习。这条往返关系把概率、优化、数值分析与生成式模型放在了一起。

A 时间离散：误差从哪里来？

正文的分布方程描述连续时间动力学。计算机只能执行有限次更新，因此即使系数完全正确，也需要分析时间离散。这里给出两种最基本的估计，作为连续模型与算法之间的补充[6]。

A.1 ODE 的 Euler 方法

考虑 $\dot{X}_t = b(t, X_t)$ 和步长 $h = T/K$ 的 Euler 方法 $\bar{X}_{k+1} = \bar{X}_k + hb(t_k, \bar{X}_k)$, $t_k = kh$ 。假设在精确轨迹与数值迭代共同所在的区域， b 对空间 L Lipschitz、对时间 M Lipschitz，且 $\|b\| \leq B$ ，初值相同。准确解的一步余项为

$$\begin{aligned} R_k &= X_{t_{k+1}} - X_{t_k} - hb(t_k, X_{t_k}) \\ &= \int_{t_k}^{t_{k+1}} [b(s, X_s) - b(t_k, X_{t_k})] ds. \end{aligned}$$

因为 $\|X_s - X_{t_k}\| \leq B(s - t_k)$ ，有 $\|R_k\| \leq (M + LB)h^2/2$ 。若 $e_k = \|X_{t_k} - \bar{X}_k\|$ ，则

$$e_{k+1} \leq (1 + Lh)e_k + Ch^2.$$

从 $e_0 = 0$ 迭代，并用 $(1 + Lh)^K \leq e^{LT}$ ，得到 $\max_{k \leq K} e_k \leq C_T h$ ； $L = 0$ 时直接求和也得到同阶结论。局部的 h^2 误差累积约 $1/h$ 次，产生全局的 h 误差。

A.2 SDE 的 Euler–Maruyama 方法

对 $dX_t = b(t, X_t) dt + \Sigma(t, X_t) dW_t$ ，令

$$\bar{X}_{k+1} = \bar{X}_k + hb(t_k, \bar{X}_k) + \Sigma(t_k, \bar{X}_k)(W_{t_{k+1}} - W_{t_k}). \quad (\text{A.1})$$

生成代码时最后一项可写成 $\sqrt{h}\Sigma(t_k, \bar{X}_k)\xi_k$ ，其中 ξ_k 独立服从 $\mathcal{N}(0, I)$ 。但做强误差分析时，我们用同一条 *Brownian* 路径耦合理想解与数值解。

命题 A.1 (基础强误差界)。设 b, Σ 对空间一致全局 Lipschitz，具有一致线性增长，且存在常数 C 使

$$\|b(t, x) - b(s, x)\|^2 + \|\Sigma(t, x) - \Sigma(s, x)\|_F^2 \leq C(1 + \|x\|^2)|t - s|.$$

设初值有有限二阶矩，且理想解与数值解初值相同。则在固定 T 上，存在与 $h \leq 1$ 无关的 C_T ，使

$$\max_{k \leq K} \mathbb{E} \|X_{t_k} - \bar{X}_k\|^2 \leq C_T h. \quad (\text{A.2})$$

特别地，均方根误差及其诱导的 W_2 上界均为 $O(\sqrt{h})$ 。

证明直觉 在时间区间长度为 h 的一小步内，随机轨迹的典型位移为 $\sqrt{h}$ 。数值法把这一小步内的系数冻结在左端点，因此系数变化的均方量为 $O(h)$ 。Itô 等距把随机积分误差变成这个均方量的时间积分，随后用 Gronwall 控制误差传播。

证明. 定义左端点 $\eta(t) = t_k$, $t \in [t_k, t_{k+1})$ ，并把数值解连续插值为

$$\bar{X}_t = X_0 + \int_0^t b(\eta(s), \bar{X}_{\eta(s)}) ds + \int_0^t \Sigma(\eta(s), \bar{X}_{\eta(s)}) dW_s.$$

先估计它的矩。对一步更新条件于 $\mathcal{F}_{t_k}$ 取平方期望，Brownian 增量的交叉项为零，线性增长与 $2uv \leq u^2 + v^2$ 给出

$$\mathbb{E} \|\bar{X}_{k+1}\|^2 \leq (1 + Ch)\mathbb{E} \|\bar{X}_k\|^2 + Ch.$$

离散 Gronwall 因而给出 $\sup_k \mathbb{E} \|\bar{X}_k\|^2 \leq C_T$ 。对一小步内的插值再用 Cauchy-Schwarz 和 Itô 等距，得到

$$\sup_{t \leq T} \mathbb{E} \|\bar{X}_t - \bar{X}_{\eta(t)}\|^2 \leq C_T h. \quad (\text{A.3})$$

令 $D_t = X_t - \bar{X}_t$ ，并记 $\delta b_s = b(s, X_s) - b(\eta(s), \bar{X}_{\eta(s)})$ ， $\delta \Sigma_s = \Sigma(s, X_s) - \Sigma(\eta(s), \bar{X}_{\eta(s)})$ 。两条积分方程相减，得到 $D_t = \int_0^t \delta b_s \, ds + \int_0^t \delta \Sigma_s \, dW_s$ 。由 Cauchy Schwarz 与 Itô 等距，

$$\mathbb{E} \|D_t\|^2 \leq 2t \int_0^t \mathbb{E} \|\delta b_s\|^2 \, ds + 2 \int_0^t \mathbb{E} \|\delta \Sigma_s\|_{\mathbb{F}}^2 \, ds.$$

在每个系数差中插入 $(t, \bar{X}_t)$ 和 $(t, \bar{X}_{\eta(t)})$ 。空间 Lipschitz、时间正则性、矩界与 (A.3) 共同给出

$$\mathbb{E} \|D_t\|^2 \leq C_T \int_0^t \mathbb{E} \|D_s\|^2 \, ds + C_T h.$$

由 $D_0 = 0$ 与 Gronwall 得到 $\sup_{t \leq T} \mathbb{E} \|D_t\|^2 \leq C_T h$ ，从而包含所有网格点的结论成立。此处只对系数差的随机积分使用等距；线性增长与二阶矩界保证它平方可积，无需额外假设初值具有四阶矩。□

这个估计是一般 Lipschitz 系数下的基础保证。对常数加性噪声、更多光滑性或特殊线性结构，强误差阶可以更好；对某些光滑观测量，弱误差也可达到 $O(h)$ 。本讲仅使用已经证明的界，不把均方根误差、观测量误差和密度误差当成同一个量。

B 带解答的练习

B.1 为什么积分要选左端点？

题目。对标准一维 Brownian motion 和分割 $0 = t_0 < \dots < t_n = T$ ，分别考虑左端点和右端点求和

$$S_n^L = \sum_k W_{t_k} \Delta W_k, \quad S_n^R = \sum_k W_{t_{k+1}} \Delta W_k.$$

求两者在分割变细时的极限，并解释差别。

解答。恒等式 $W_{t_{k+1}}^2 - W_{t_k}^2 = 2W_{t_k} \Delta W_k + (\Delta W_k)^2$ 求和给出 $S_n^L = \frac{1}{2}(W_T^2 - \sum_k (\Delta W_k)^2)$ 。二次变差收敛到 T ，所以

$$S_n^L \longrightarrow \frac{1}{2}(W_T^2 - T), \quad S_n^R = S_n^L + \sum_k (\Delta W_k)^2 \longrightarrow \frac{1}{2}(W_T^2 + T).$$

右端点的值已经使用了当前增量，因此交叉项不再具有 Itô 积分中的零均值结构。差别恰为不可忽略的二次变差，而不是普通 Riemann 积分中会消失的端点选择误差。

B.2 稳态 OU 的反向过程和概率流

题目。 设 $dX_t = -\lambda X_t dt + \sqrt{2\lambda} dW_t$, $\lambda > 0$, 且 $X_0 \sim \mathcal{N}(0, 1)$ 。写出反向时间漂移和概率流 ODE。

解答。 所有边缘均为标准高斯, score 为 $-x$ 。使用递增反向时钟, 反向漂移为 $\lambda x + 2\lambda(-x) = -\lambda x$, 与正向漂移相同。概率流速度为 $-\lambda x - \lambda(-x) = 0$ 。所以概率流的每个样本保持不动, 却与不断随机运动的 OU 过程有完全相同的单时刻分布。这同时检查了反向漂移的符号和概率流中的 $1/2$ 系数。

B.3 固定生长率为什么需要中心化?

题目。 两个状态的固定生长率为 r_1, r_2 , 第一状态初始概率为 $p_0 \in (0, 1)$ 。给出归一化权重的显式解。

解答。 先更新未归一化质量: $\tilde{p}_1(t) = p_0 e^{r_1 t}$, $\tilde{p}_2(t) = (1 - p_0) e^{r_2 t}$ 。归一化得到

$$p(t) = \frac{p_0 e^{r_1 t}}{p_0 e^{r_1 t} + (1 - p_0) e^{r_2 t}}, \quad \dot{p} = (r_1 - r_2)p(1 - p).$$

因此所有 r_i 同时加上同一个常数不会改变概率演化。中心化反应项正好表达这一点: 真正有意义的是相对生长率。

B.4 边缘分布是否唯一决定速度场?

题目。 在 $\mathbb{R}^2$ 上令 $\rho(x) = (2\pi)^{-1} e^{-\|x\|^2/2}$ 。证明零速度场与旋转速度场 $v(x) = Ax$, $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, 都保持 ρ 不变。

解答。 零速度显然保持分布不变。对旋转速度, $\text{Tr } A = 0$, 而 $x^\top Ax = 0$, 所以

$$\nabla \cdot (\rho Ax) = \rho \text{Tr } A + (Ax) \cdot \nabla \rho = 0 - \rho x^\top Ax = 0.$$

两个 ODE 的轨迹分别是静止和旋转, 边缘分布却相同。这解释了为什么检验连续性方程或 Fokker-Planck 方程可以确认单时刻分布, 而要确定一个给定随机过程的完整反向路径规律, 还需要反向转移核。Flow Matching 的条件回归则在给定插值构造下选定一个具体速度。

参考文献与原课程资料

- [1] 陆一平。 *Statistical Learning: Optimal Transport* 课程资料, 2025。课程主页: <https://2prime.github.io/teaching/2025-Statistical-Learning>; 原课件: <https://2prime.github.io/files/IEMS402/ot.pdf>。
- [2] 陆一平。 *Statistical Learning: 粒子系统与概率空间计算讲义*, 2025。 <https://2prime.github.io/files/IEMS402/particlesystem.pdf>。本讲沿用原课程从粒子动力学理解分布计算的思路, 并补充随机分析、反向过程及生成式训练目标的证明。
- [3] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Foundations and Trends in Machine Learning, 11(5–6):355–607, 2019. <https://arxiv.org/abs/1803.00567>。
- [4] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. <https://doi.org/10.1137/S0036141096303359>。
- [5] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *ICLR*, 2014. 预印本初稿为 2013 年。 <https://arxiv.org/abs/1312.6114>。
- [6] Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*. Cambridge University Press, 2019. 作者提供的阅读版本: <https://users.aalto.fi/~asolin/sde-book/sde-book.pdf>。
- [7] Yulong Lu, Jianfeng Lu, and James Nolen. Accelerating Langevin Sampling with Birth-death. *NeurIPS*, 2019. <https://arxiv.org/abs/1905.09863>。
- [8] Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. [https://doi.org/10.1016/0304-4149\(82\)90051-5](https://doi.org/10.1016/0304-4149(82)90051-5)。
- [9] U. G. Haussmann and E. Pardoux. Time Reversal of Diffusions. *The Annals of Probability*, 14(4):1188–1205, 1986. <https://doi.org/10.1214/aop/1176992362>。
- [10] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. *ICLR*, 2021. <https://arxiv.org/abs/2011.13456>。
- [11] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. *ICLR*, 2023. 预印本初稿为 2022 年。 <https://arxiv.org/abs/2210.02747>。
- [12] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic Interpolants: A Unifying Framework for Flows and Diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. 预印本初稿为 2023 年。 <https://www.jmlr.org/papers/v26/23-1605.html>; <https://arxiv.org/abs/2303.08797>。