About Diffusion and Flow Matching: From Mathematics to Practice

Note based on README

2025 年 11 月 23 日

摘要

本文从数学的角度解释当今最流行的 Text-to-Image (t2i) 框架：Diffusion 和 Rectified Flow。我们将从高层理论解释如何一眼看破在 Stable Diffusion 和 Flux 各版本训练/推理中出现的 Trick。这也可以看作 MIT S186 Diff and Flow Matching 的理论注解。

目录

1 什么是 Flow Matching

Flow matching 的想法很简单：过去的生成算法都是 Encoder-Decoder 架构，训练数据要从 Encoder 流入 Latent Space，推理时再从 Latent Space 出发流到真实图像分布。那为啥不直接学习后者的向量场呢？

下面给出数学描述：

1.1 概率路径与速度场

假设我们定义了一条路径 $p_t(x)$，对每个 t，有一个密度 $p_t(x)$。这条路径的演化被如下 ODE 的 flow 推动：

$$\frac{\partial x}{\partial t} = u(x,t), \tag{1}$$

其中 $u(x,t)$ 是我们要学习的 **速度场**。

良好定义的路径还必须满足概率守恒方程（即连续性方程）：

$$\frac{\partial p_t(x)}{\partial t} + \nabla \cdot (p_t(x)\, u(x,t)) = 0. \tag{2}$$

它其实就是 Flow Matching 版本的 Fokker-Planck 方程，由归一化条件/质量守恒得到。注意这与 Diffusion 不同：Diffusion 的 SDE 自然对应随机变量的演化，所以归一化条件自动满足，其 Fokker-Planck 方程由 SDE 本身导出。

注意

Flow matching 有个问题：最终流向的分布全测度未必是 1，此时根据归一化分布采样就好。

1.2 训练目标 / 损失函数

假设我们从数据分布采样到 $x_1 \sim p_1(x)$，然后在路径上给定中间时刻 t，我们可以定义一个条件分布（“从数据点向前 / 向后推”）$p(x_t \mid x_1, t)$，它有一个已知 / 可计算的速度场（或者说条件速度场）$u_{\text{cond}}(x_t, t; x_1)$。这个速度场通常比较简单 / 显式。

那么我们训练 u_θ 使得：

$$L(\theta) = \mathbb{E}_{t,x_1,x_t\sim p(\cdot|x_1,t)}\left[\|u_\theta(x_t,t) - u_{\text{cond}}(x_t,t;x_1)\|^2\right]. \tag{3}$$

这个损失就是在逼近网络匹配真实的速度场 $u(x,t)$。至于为什么用平方损失，理论上的考量是因为其有较好的数学结构，Hilbert 空间等泛函分析理论能够在此大展拳脚。

1.3 采样

在训练好 u_θ 后，生成过程就简单许多：从初始简单分布（如标准高斯）在 $t=0$ 取样 x_0，然后解下面的常微分方程（ODE）：

$$\frac{dx}{dt} = u_\theta(x,t), \tag{4}$$

将它从 $t=0$ 移动到 $t=1$，最终得到一个接近目标数据分布的样本 x_1。

1.4 路径设计

路径 p_t 的设计是关键。常见的选择包括 **高斯插值路径（Gaussian paths）**，或者设计成与 **最优传输路径（Optimal Transport）** 更贴近，从而使得速度场更为合理、高效。比如在 *Flow Matching* 的原始工作里，就提出了若干路径选项，其中就包括使用最优传输插值的路径。

2 什么是 Rectified Flow

Rectified Flow (RF) 认为：为了减少推理步数，速度场 $u_{\text{cond}}(x_t, t; x_1)$ 应该就是连接起点和终点的直线。

所以训练目标就变成最小化以下损失函数：

$$L = \mathbb{E}_{x_0, x_1, t}\left[\|v_\theta(x_t, t) - (x_1 - x_0)\|^2\right]. \tag{5}$$

数学表述的注意事项

我不喜欢写 $\frac{dx}{dt} = x_1 - x_0$ 这种数学上非常糟糕的表达式。这里面的变量没法解释：虽然对确定的采样条件向量场是直线，但是整个速度场并非直线。

正确的思路是：先用 RF 的 loss 逼近这种“局部直线”的速度场 v，然后再用 v 做推理：

$$\frac{dx}{dt} = v(x, t)$$

问题来了，对两个分布 p_0 和 p_1，真的能通过这种直线流把一个变成另一个吗？熟悉条件概率的人一眼就能看出来，这个 L^2 loss 其实在目标函数取条件期望的时候取得最小值，该最小值是 $X_1 - X_0$ 在 L^2 空间里相对于 X_t 的条件期望:

$$v^\star(x, t) = \mathbb{E}[X_1 - X_0 \mid X_t = x], \tag{6}$$

其中随机变量满足：

$$X_t = (1 - t)X_0 + tX_1, \quad (X_0, X_1) \sim \pi \in \Pi(p_0, p_1).$$

这里 $\Pi(p_0, p_1)$ 表示所有边缘分布分别为 p_0 与 p_1 的联合分布。

所以这个 MSE loss 是有严格下界的，不小于 Hilbert 空间里 $X_1 - X_0$ 到 t 时刻事件域对应的子空间的距离。换言之，由于随机采样的引入，Universal Approximation 在这里并不适用。

那啥时候能把 loss 降到 0 呢？至少理论上我们希望如此。熟悉应用数学的立马就能看出 Flow Matching 就是在做一种最优传输。

2.1 背景：两个分布之间的“耦合（Coupling）”

学调和分析的肯定熟悉 Coupling，各种 Interpolation 定理都要用；学随机过程的必然也不陌生，Markov chain 的极限分布定理就是这么来的。最优传输的 coupling 就是一回事：要把初始高斯分布和最终数据分布给配个对，才方便分布的传输。

给定两个概率分布 p_0 和 p_1，我们想描述“如何把一个随机变量 $X_0 \sim p_0$ 变成另一个随机变量 $X_1 \sim p_1$”。

最一般的方式是考虑它们的联合分布 $\pi(x_0, x_1)$，满足边缘条件：

$$\int \pi(x_0, x_1)\, dx_1 = p_0(x_0), \quad \int \pi(x_0, x_1)\, dx_0 = p_1(x_1).$$

所有满足该条件的联合分布构成一个 Kantorovich 耦合集合 $\Pi(p_0, p_1)$。

2.1.1 Monge 耦合与 Brenier 定理

Monge 耦合 是一种“极端特殊”的耦合形式，它要求对每个起点 x_0，都只对应唯一的终点 $x_1 = T(x_0)$。

当代价函数为平方欧氏距离 $c(x_0, x_1) = \frac{1}{2}\|x_0 - x_1\|^2$，并且 p_0 是绝对连续分布时，有 **Brenier’s Theorem (1987)**：存在唯一的最优 Monge 映射 $T = \nabla\varphi$，其中 φ 是凸函数，使得

$$T_* p_0 = p_1. \tag{7}$$

这个映射 T 实现了最小传输代价，也定义了 2-Wasserstein 空间上的 Geodesic，即最“高效”的流。

2.2 与 Rectified Flow 的关系

我们证明 Brenier 映射 $X_1 = T(X_0)$ 使得 RF loss 的最优速度场 $v^\star(x, t)$ 确实是起点指向终点的线段。

第 1 步： Φ_t 的可逆性

定义 $\psi_t(x) := \frac{1-t}{2}\|x\|^2 + t\,\varphi(x)$。由于 φ 为凸函数，ψ_t 是严格凸的。其梯度 $\Phi_t(x) = \nabla\psi_t(x) = (1-t)x + t\nabla\varphi(x)$ 是拓扑同胚。因此，对于 $t \in [0, 1)$，存在逆映射使得 $X_0 = \Phi_t^{-1}(X_t)$。

第 2 步：计算期望

令 $h(x_0) = X_1 - X_0 = T(x_0) - x_0$，则有：

$$\mathbb{E}[X_1 - X_0 \mid X_t] = (T - \mathrm{Id})(\Phi_t^{-1}(X_t)).$$

于是对任意 x：

$$v^\star(x, t) = T(\Phi_t^{-1}(x)) - \Phi_t^{-1}(x). \tag{8}$$

这表明在最优传输（$T = \nabla\varphi$）下，RF 的 loss 是 0，最优速度场确实是直线。

3 Stable Diffusion 3 中的两种训练策略

3.1 Reflow

在直接用 RF 的 loss 训练模型后，轨迹的直线性没有保证，因为随机采样噪声和样本点对应一个随机的 coupling。根据理论我们需要一个一一对应的 coupling $X_1 = T(X_0)$。此时理论上 loss 能降到 0。实际训练中，只要用直接训练的 RF 模型生成噪声-图像对，并喂进去训练一轮，loss 就能大幅下降。这属于 RF 的 **Self-Instruct** 和 **One-Epoch** 原则。

3.2 U-shaped Time Weight

Benamou–Brenier (2000) 证明了动态版本最优传输问题：

$$W_2^2(p_0, p_1) = \inf_{(p_t, v_t)} \int_0^1 \int \|v_t(x)\|^2 \, p_t(x) \, dx \, dt,$$

约束条件为连续性方程。最优解满足 $v_t^\star(x) = \nabla \varphi_t(x)$。这意味着最优速度场在每个时刻都是梯度场。我们可以采取不同的时间步采样，来使得 loss 下降的同时提高模型的泛化性。

3.3 实验结果 (Stable Diffusion 3)

以下数据来自 2024 年 SD 组的实验《Improving the Training of Rectified Flows》。

表 1: 消融表：一步/两步 FID 的逐项改进 (NFE=1 / NFE=2)

方法	**CIFAR-10**	**AFHQ-64**	**FFHQ-64**
Base [Liu et al., 2022] (A)	12.21 / —	—	—
(A) + EDM init + large batch (B)	7.14 / 3.61	$\downarrow$ ($\sim 8 \to 5$)	$\downarrow$ ($\sim 8 \to 5$)
(B) + Our p_t (C)	5.17 / 3.37	$\downarrow$	$\downarrow$
(C) + Huber (D)	5.24 / 3.34	$\pm$	$\pm$
(C) + LPIPS-Huber (E)	3.42 / 2.95	$\downarrow$	5.21 / 4.26
(C) + LPIPS-Huber$_t$ (F)	3.38 / 2.76	4.11 / 3.12	5.65 / 4.41
(F) + Real data (G)	**3.07 / 2.40**	—	—

结论： 关键贡献点是 EDM 预热、感知加权路径分布 p_t 以及带时间权重的 LPIPS-Huber。仅一轮 ReFlow，一步 FID 改进约 72%。

表 2: 训练总成本对比

方法	合成配对前向次数 (M)	训练前向 (M)	总成本相对值
ReFlow (本文)	395	1,433.6	$\times 1$
CD (Consistency Distillation)	—	5,734.4	$\times 3.1$
CT (Consistency Training)	—	2,867.2	$\times 1.5$

反演与重建： 实验表明，改良版 RF 不仅少步采样优异，同时具备更好的可逆性与重建稳定性（比 EDM 更好），支持高质量反演与编辑。

4 数学注记与理论联系

4.1 最优传输的动态形式 (Dynamic OT)

Benamou & Brenier (2000) 给出的动态表述为：

$$\begin{aligned}\min_{v_t,\rho_t} \quad & \int_0^1 \int \tfrac{1}{2}\|v_t(x)\|^2 \,\rho_t(x)\,dx\,dt \\ \text{s.t.} \quad & \partial_t \rho_t + \nabla\cdot(\rho_t v_t) = 0, \quad \rho_0 = p_0,\ \rho_1 = p_1.\end{aligned}$$

这其实就是一个最优控制问题，且 cost 是能量形式 $\|v\|^2$。

4.2 与强化学习 (RL) 的统一

强化学习可以看作是最优传输的动态随机推广。

表 3: OT 与 RL 的对应关系

版本	数学形式	对应到 RL
Deterministic OT	$\min \int \|v_t\|^2 \,\rho_t$	Deterministic policy
Stochastic OT	$\min \mathbb{E}\int c(x_t,a_t)\,dt \quad \text{s.t. } dx_t = f + \sigma dW_t$	Stochastic policy / MDP
Controlled OT	加入奖励项或控制能量正则	Entropy-regularized RL

- OT 的连续性方程 $\partial_t\rho_t + \nabla\cdot(\rho_t v_t) = 0$ 对应确定性演化。
- RL 的 Fokker-Planck 方程 $\partial_t\rho_t + \nabla\cdot(\rho_t f_t) - \frac{1}{2}\nabla^2 : (\rho_t\sigma\sigma^\top) = 0$ 对应带噪声演化。

这为 Rectified Flow 与 RL 提供了共同的数学框架：它们都在优化“路径分布 + 速度场”的能量泛函。

4.3 Self-Instruct 与 EM 算法

LLM 的 Self-Instruct 实际上对应于数学上的 **硬 EM (Hard EM) + 自蒸馏**：

- **E 步 (生成)**: $z^{(t)} = \arg\max_z \, p_{\theta^{(t)}}(z \mid x)$ (生成伪标签)
- **M 步 (微调)**: $\theta^{(t+1)} = \arg\max_\theta \, \log p_\theta(x, z^{(t)})$ (在伪标签上训练)

Reflow 也是同样的逻辑：生成噪声-图像对（E步），再喂进去训练（M步）。

4.4 One-Epoch 原则与推荐系统

ReFlow 的“一轮足够”原则与推荐系统中的 **LightGCN** 或矩阵分解的 **One-Step Propagation** 思想一致。

LightGCN 的核心：删除非线性和权重矩阵，仅保留邻居平均传播 $h_i^{(k+1)} = \sum_{j\in\mathcal{N}(i)} \frac{1}{\sqrt{d_i d_j}} h_j^{(k)}$。虽然看似简化，但通过图结构的 K 阶邻域传播捕获了复杂模式。这与 RF 的“一次流动逼近最优路径”具有惊人的相似性。

5 补充：关于推理时的 ODE 求解器

从数学角度看，RF 和 Diffusion 的 ODE 形式不同，应采取不同的求解器。

5.1 Heun 方法 (for Rectified Flow)

Heun 方法是 RK2 的一个特例，几何上是“先用 Euler 预测，再用梯形平均修正”：

$$\begin{cases} k_1 = f(x_n, t_n), \\ k_2 = f(x_n + \Delta t\, k_1,\, t_n + \Delta t), \\ x_{n+1} = x_n + \frac{\Delta t}{2}(k_1 + k_2). \end{cases} \tag{9}$$

由于 RF 训练的轨迹几乎是 **直线轨迹**，低阶 Heun 方法已能达到足够高的精度且成本较低。

5.2 DPM Solver (for Diffusion)

DPM Solver 利用了 Diffusion ODE 的半线性结构。扩散 ODE 可写为线性非齐次形式：

$$\frac{dx_t}{dt} = -\frac{1}{2}\beta(t)x_t - \frac{1}{2}\beta(t)\epsilon_\theta(x_t, t).$$

使用积分因子法（Integrating Factor），可以推导出解析解形式的更新公式：

$$x_{t_{n+1}} = e^{-\int_{t_n}^{t_{n+1}} \frac{\beta(s)}{2} ds}\, x_{t_n} - \int_{t_n}^{t_{n+1}} e^{-\int_{s}^{t_{n+1}} \frac{\beta(u)}{2} du}\, \frac{\beta(s)}{2}\, \epsilon_\theta(x_s, s)\, ds. \tag{10}$$

DPM Solver 专门针对此结构优化。但对于 SD3/RF 这类 Flow Matching 模型，DPM Solver 并不适用，应首选 Heun 或 Euler。