
学海无涯

洗衣机

Last Update: 2026 年 4 月 7 日

目录

1 Optimizaiton	1
1.1 概念&定义	1
1.2 分式规划 (Fractional Programming): Dinkelbach 算法	4
1.3 酉不变性	5
1.4 变分不等式	5
1.5 凸的等价刻画	7
1.6 强凸函数的等价刻画	8
1.7 理论依据: 一阶条件下梯度下降为何一定下降?	9
1.8 梯度下降法中的步长分析与收敛速度	10
1.8.1 问题设定与基础条件	10
1.8.2 步长 $\alpha < \frac{2}{L}$ 保证下降	10
1.8.3 步长 $\alpha \leq \frac{1}{L}$ 才能推出 $\mathcal{O}(1/k)$ 收敛	10
1.8.4 误差线性收敛的条件与最优步长	11
1.9 误差收缩与线性化算子谱半径的关系	12
1.10 优化算法收敛性证明框架的两类方法	13
1.10.1 不动点类	13
1.10.2 能量函数类	15
1.11 Douglas-Rachford 分裂算法的等价形式	16
1.11.1 收敛性分析	16
1.12 收敛性分析: 基于Mann迭代的Douglas-Rachford算法	18
1.12.1 DR算法的等价迭代形式	19
1.12.2 最大单调性假设与约束资格	19
1.12.3 Opial引理	19
1.12.4 Mann迭代与非扩张算子	20
1.12.5 DR收敛性定理	20
1.13 全序列收敛的常用技巧	21
1.14 Scaling 缩放技巧	22
1.15 Kurdyka-Łojasiewicz 条件	23

1.16	Farkas 引理	25
1.17	动量法与外推步	25
1.17.1	多项式重球法 (Polyak's Heavy Ball)[16]	25
1.17.2	Nesterov 加速梯度法 (NAG)	25
1.17.3	外推法 (Extrapolation Method)	26
1.17.4	连续时间视角: ODE 推导与渐隐阻尼	26
1.17.5	总结对比	27
2	Diffusion	28
2.1	from DDPM to DDIM	28
2.2	Flow Matching 中边缘速度场公式的推导与理解	30
2.2.1	数学推导: 基于连续性方程	31
2.2.2	直观物理解	32
2.3	从sde到ode	32
2.3.1	方法一	32
2.3.2	方法二	33
2.4	扩散模型核心架构	33
2.4.1	U-net	33
2.5	Diffusion-Distillation DPO	34
2.5.1	Diffusion-Distillation Progressive	34
2.5.2	Distillation-DPO改进方向	35
2.6	基于强化学习的扩散模型	36
2.7	flow mathching + GRPO	37
2.7.1	GRPO的缺陷	37
2.7.2	SDPO	37
2.7.3	Diffusion + SDPO	38
2.8	Toy idea: Diffusion + 连续优化 (ADMM/DRSM)	39
2.9	Optimal transport in flow matching	40
2.9.1	静态最优传输的数学底座	41
2.9.2	Brenier 定理与最优传输映射的还原	45
2.9.3	从几何测地线到流体力学视角	46
2.9.4	Flow Matching框架	48
2.9.5	矫正流 (Rectified Flow) 与深度拉直机制	50
2.10	Score Distillation	52
2.10.1	KL 梯度流与 Score Distillation 的统一	52
2.11	Flow Distillation	54
2.11.1	flow distillation、Bregman散度、Wasserstein空间的统一	55

3 Agent	58
3.1 面向 Lean 4 的语义检索与混合排序	58
3.1.1 问题背景	58
3.1.2 LeanSearch: 语义检索与查询增广	58
3.1.3 LeanExplore: 混合排序与 StatementGroup	59
3.1.4 两条路线对照 (检索方法)	59
3.2 终身学习 LLM Agent: 路线图与 AutoSkill	60
3.2.1 综述: 终身学习 LLM Agent 路线图	60
3.2.2 AutoSkill: 技能自演化	60
4 PDE	61
4.1 分部积分公式	61
参考文献	62

Chapter 1

Optimizaiton

1.1 概念&定义

定义 1.1.1 (凸函数). 令 $\Omega \subset \mathbb{R}^n$ 为凸集, $\theta: \Omega \rightarrow \mathbb{R}$. 若对任意 $u, v \in \Omega$ 和 $\lambda \in [0, 1]$, 都有

$$\theta(\lambda u + (1 - \lambda)v) \leq \lambda \theta(u) + (1 - \lambda) \theta(v),$$

则称 θ 在 Ω 上为凸函数。

定义 1.1.2 (严格凸函数). 在同样记号下, 若对任意 $u, v \in \Omega$ 且 $u \neq v$, 以及 $\lambda \in (0, 1)$, 严格不等式

$$\theta(\lambda u + (1 - \lambda)v) < \lambda \theta(u) + (1 - \lambda) \theta(v)$$

成立, 则称 θ 在 Ω 上为严格凸函数。

定义 1.1.3 (强凸函数). 若存在常数 $\sigma > 0$, 使得对任意 $u, v \in \Omega$ 和 $\lambda \in (0, 1)$,

$$\theta(\lambda u + (1 - \lambda)v) \leq \lambda \theta(u) + (1 - \lambda) \theta(v) - \frac{\sigma}{2} \lambda(1 - \lambda) \|u - v\|^2,$$

则称 θ 在 Ω 上为强凸函数。

备注 1.1.1. 由上述定义可知:

$$\text{强凸} \implies \text{严格凸} \implies \text{凸}.$$

定义 1.1.4 (可微凸函数的梯度不等式). 若 θ 在 Ω 上可微, 则对任意 $u, v \in \Omega$ 有

$$\theta(u) \geq \theta(v) + \nabla \theta(v)^T (u - v)$$

当且仅当 θ 为凸函数。

定义 1.1.5 (次梯度与次微分). 令 $\theta : \Omega \rightarrow \mathbb{R}$, Ω 凸。若对于某向量 $s \in \mathbb{R}^n$, 对任意 $u \in \Omega$, 都有

$$\theta(u) \geq \theta(v) + s^T(u - v),$$

则称 s 为 θ 在 $v \in \Omega$ 处的次梯度 (Subgradient)。所有次梯度构成的集合称为 θ 在 v 处的次微分 (Subdifferential), 记为 $\partial\theta(v)$ 。

定义 1.1.6 (Lipschitz 连续). 令 $F : \Omega \rightarrow \mathbb{R}^m$ 。若存在常数 $L > 0$, 使得对所有 $u, v \in \Omega$,

$$\|F(u) - F(v)\| \leq L \|u - v\|,$$

则称 F 在 Ω 上为 Lipschitz 连续 (Lipschitz continuous)。

定义 1.1.7 (算子类型). 令 $\Omega \subset \mathbb{R}^n$ 为凸集, $F : \Omega \rightarrow \mathbb{R}^n$ 为映射; 亦令 $(\mathcal{H}, \|\cdot\|)$ 为实 Hilbert 空间, $T : \mathcal{H} \rightarrow \mathcal{H}$ 。则:

1. (单调算子 Monotone) 若对任意 $u, v \in \Omega$ 有

$$(u - v)^\top [F(u) - F(v)] \geq 0,$$

则称 F 在 Ω 上为单调算子;

2. (严格单调算子 Strictly Monotone) 若对任意 $u, v \in \Omega$, $u \neq v$ 有

$$(u - v)^\top [F(u) - F(v)] > 0,$$

则称 F 为严格单调算子;

3. (强单调算子 Strongly Monotone) 若存在 $\eta > 0$ 使得对任意 $u, v \in \Omega$,

$$(u - v)^\top [F(u) - F(v)] \geq \eta \|u - v\|^2,$$

则称 F 为强单调算子;

4. (伪单调算子 Pseudo-monotone) 若对任意 $u, v \in \Omega$,

$$(u - v)^\top F(v) \geq 0 \implies (u - v)^\top F(u) \geq 0,$$

则称 F 为伪单调算子;

5. (余强制算子 Co-coercive) 若存在 $\mu > 0$ 使得对任意 $u, v \in \Omega$,

$$(u - v)^\top [F(u) - F(v)] \geq \mu \|F(u) - F(v)\|^2,$$

则称 F 为余强制算子;

6. (非扩张算子 *Nonexpansive*) 若对任意 $x, y \in \mathcal{H}$ 有

$$\|T(x) - T(y)\| \leq \|x - y\|,$$

则称 T 为非扩张算子;

7. (严格非扩张算子 *Strictly Nonexpansive*) 若存在 $\rho \in [0, 1)$ 使得对任意 $x, y \in \mathcal{H}$,

$$\|T(x) - T(y)\| \leq \rho \|x - y\|,$$

则称 T 为严格非扩张算子;

8. (α -平均算子 *α -Averaged*) 若存在 $\alpha \in (0, 1)$ 与非扩张算子 N 使得

$$T = (1 - \alpha)I + \alpha N,$$

则称 T 为 α -平均算子。

1.2 分式规划 (Fractional Programming): Dinkelbach 算法

考虑如下问题

$$\min_{x \in \mathcal{S}} F(x) = \frac{f(x)}{g(x)}$$

其中 $f(x), g$ 是适当的闭凸函数, $g(x) > 0$.

Algorithm 1 Dinkelbach 算法

1: **Input:** 初始值 $\lambda^{(0)} \in \mathbb{R}$, 迭代次数 $k = 0$, 约束条件 $x \in \mathcal{S}$, 精度阈值 $\epsilon > 0$

2: **Output:** 最优解 x^* 和最优值 λ^*

3: **repeat**

4: **Step 1:** 求解子问题:

$$x^{(k)} = \arg \max_{x \in X} \{f(x) - \lambda^{(k)} g(x)\}$$

5: **Step 2:** 计算:

$$\phi(\lambda^{(k)}) = \max_{x \in X} \{f(x) - \lambda^{(k)} g(x)\}$$

6: **Step 3:** 更新:

$$\lambda^{(k+1)} = \frac{f(x^{(k)})}{g(x^{(k)})}$$

7: 更新迭代次数: $k \leftarrow k + 1$

8: **until** $\phi(\lambda^{(k)}) < \epsilon$

性质 1.2.1. ϕ 关于 λ 单调递减: $\lambda_1 < \lambda_2 \Rightarrow \phi(\lambda_1) > \phi(\lambda_2)$.

性质 1.2.2. $\lambda = \lambda^* \Leftrightarrow \phi(\lambda) = 0$.

证明. (\Rightarrow): 令 $\lambda = \lambda^* = F(x^*) = \frac{f(x^*)}{g(x^*)}$. $\forall x \in \mathcal{S}, \lambda^* \leq \frac{f(x)}{g(x)} \Rightarrow f(x) - \lambda^* g(x) \geq 0$, 因此 x^* 恰好取到 $\phi(\lambda^*)$ 的下界 0.

(\Leftarrow): 假设存在 $\lambda' = F(x')$ 是更优解, 因此 $\lambda' = \frac{f(x')}{g(x')} < \lambda \Rightarrow f(x') - \lambda g(x') < 0$, i.e. $\phi(\lambda) < 0$, 矛盾.

□

1.3 酉不变性

酉矩阵保持内积不变

- 酉矩阵保持向量范数不变;
- 酉相似变换不改变矩阵的谱（特征值、奇异值）;

矩阵的F范数可以用奇异值刻画，因此也是酉不变的。

矩阵的诱导2范数等价于矩阵的谱半径，因此也是酉不变的。

1.4 变分不等式

要证明变分不等式问题 $VI(F, \mathbb{R}_+^n)$ 与互补问题

$$u \geq 0, \quad F(u) \geq 0, \quad u^T F(u) = 0$$

的等价性，我们需要分别证明两个方向的蕴含关系。

****1. 变分不等式 \Rightarrow 互补问题**** 假设 u 是 $VI(F, \mathbb{R}_+^n)$ 的解，即

$$\forall v \in \mathbb{R}_+^n, \quad (v - u)^T F(u) \geq 0.$$

需要证明 u 满足互补问题的三个条件。

**** (1) 非负性 $u \geq 0$ 和 $F(u) \geq 0$ **** - 由于 \mathbb{R}_+^n 是可行域，显然 $u \geq 0$ 。 - 取 $v = u + e_i$ (其中 e_i 是第 i 个单位向量)，则 $v \in \mathbb{R}_+^n$ ，代入变分不等式得

$$(v - u)^T F(u) = e_i^T F(u) = F_i(u) \geq 0.$$

这对所有 i 成立，故 $F(u) \geq 0$ 。

**** (2) 正交性 $u^T F(u) = 0$ **** - 取 $v = 0$ ，代入变分不等式得

$$(0 - u)^T F(u) = -u^T F(u) \geq 0 \quad \Rightarrow \quad u^T F(u) \leq 0.$$

- 取 $v = 2u$ ，代入得

$$(2u - u)^T F(u) = u^T F(u) \geq 0.$$

- 结合两者，得 $u^T F(u) = 0$ 。

综上, $\text{VI}(F, \mathbb{R}_+^n)$ 的解 u 满足互补问题。

—
2. 互补问题 \Rightarrow 变分不等式 假设 u 满足互补问题

$$u \geq 0, \quad F(u) \geq 0, \quad u^T F(u) = 0,$$

需要证明 u 是 $\text{VI}(F, \mathbb{R}_+^n)$ 的解。

对任意 $v \in \mathbb{R}_+^n$, 有

$$(v - u)^T F(u) = v^T F(u) - u^T F(u) = v^T F(u).$$

由于 $v \geq 0$ 且 $F(u) \geq 0$, 故 $v^T F(u) \geq 0$ 。因此

$$(v - u)^T F(u) \geq 0,$$

即 u 满足变分不等式。

—
结论 变分不等式 $\text{VI}(F, \mathbb{R}_+^n)$ 和互补问题

$$u \geq 0, \quad F(u) \geq 0, \quad u^T F(u) = 0$$

在解集上完全等价。

1.5 凸的等价刻画

可微函数 $\theta(\cdot)$ 在非空凸集 Ω 上是凸的, 当且仅当其梯度 $\nabla\theta(\cdot)$ 是一个单调算子, 即对 $\nabla\theta(\cdot)$ 满足:

$$(u - v)^T(\nabla\theta(u) - \nabla\theta(v)) \geq 0, \quad \forall u, v \in \Omega.$$

证明关键在于将 $\theta(u + t(u - v))$ 视作一元函数

类似地, $\theta(\cdot)$ 是严格凸的等价于 $\nabla\theta(\cdot)$ 是严格单调的; $\theta(\cdot)$ 是强凸的等价于 $\nabla\theta$ 是强单调的. 但需要注意的是, 若一个映射 $F(\cdot)$ 是单调的, 我们并不能直接得到 $F(\cdot)$ 是某一凸函数 $\theta(\cdot)$ 的梯度; 其他情况也是类似的.

1.6 强凸函数的等价刻画

设函数 $f: \mathbb{R}^n \rightarrow \mathbb{R}$, 我们称 f 是 μ -强凸的 ($\mu > 0$), 若存在常数 $\mu > 0$, 使得以下任一条件成立。若 f 为连续可导函数, 则下列条件等价:

1. 一阶不等式形式 (定义): 对于任意 $x, y \in \mathbb{R}^n$,

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

这是强凸性的经典定义形式, 也称为带二次项的 Jensen 不等式。

2. 梯度强单调性 (gradient strong monotonicity): 对于任意 $x, y \in \mathbb{R}^n$,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2.$$

它反映了 ∇f 是 μ -强单调的向量场。

3. Hessian 有界性 ($f \in \mathcal{C}^2$ 时):

$$\nabla^2 f(x) \succeq \mu I, \quad \forall x \in \mathbb{R}^n.$$

即 $\nabla^2 f(x)$ 的最小特征值不小于 μ , 说明函数在每一点局部至少具有 μ 的曲率。

备注:

- 如果 f 是不可微但闭凸函数, 则强凸性的定义可通过次梯度推广为:

$$f(y) \geq f(x) + \langle g, y - x \rangle + \frac{\mu}{2} \|y - x\|^2, \quad \forall x, y, g \in \partial f(x).$$

- 函数 f 是 μ -强凸的 \Leftrightarrow 函数 $f(x) - \frac{\mu}{2} \|x\|^2$ 是凸的。

这些等价刻画为分析梯度法与一阶方法提供了理论基础。

1.7 理论依据：一阶条件下梯度下降为何一定下降？

定义 1.7.1 (下降方向). 设 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 是定义在 \mathbb{R}^n 上的连续可微函数。如果向量 $0 \neq \mathbf{d} \in \mathbb{R}^n$ 在点 \mathbf{x} 处的函数 f 的方向导数 $f'(\mathbf{x}; \mathbf{d})$ 是负的，即

$$f'(\mathbf{x}; \mathbf{d}) = \nabla f(\mathbf{x})^T \mathbf{d} < 0,$$

则称 \mathbf{d} 是 f 在 \mathbf{x} 处的下降方向。

下降方向是梯度下降的理论基础。

引理 1.7.1 (下降方向的下降性质). 设 f 是定义在 \mathbb{R}^n 上的连续可微函数，且 $\mathbf{x} \in \mathbb{R}^n$ 。假设 \mathbf{d} 是 f 在 \mathbf{x} 处的下降方向。则存在 $\varepsilon > 0$ 使得

$$f(\mathbf{x} + t\mathbf{d}) < f(\mathbf{x})$$

对任意 $t \in (0, \varepsilon]$ 成立。

证明. 由于 $f'(\mathbf{x}; \mathbf{d}) < 0$ ，根据方向导数的定义，有

$$\lim_{t \rightarrow 0^+} \frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t} = f'(\mathbf{x}; \mathbf{d}) < 0.$$

因此，存在 $\varepsilon > 0$ 使得

$$\frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t} < 0$$

对任意 $t \in (0, \varepsilon]$ 成立，这直接意味着所需的结果。 □

1.8 梯度下降法中的步长分析与收敛速度

1.8.1 问题设定与基础条件

设 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 是凸函数, 且 ∇f 是 L -Lipschitz 连续的, 即对任意 x, y 有

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

考虑标准的梯度下降迭代:

$$x_{k+1} = x_k - \alpha \nabla f(x_k),$$

我们将分析步长 α 的取值对算法性能与收敛速度的影响。

1.8.2 步长 $\alpha < \frac{2}{L}$ 保证下降

由 Descent Lemma (L -光滑性), 对于任意 $\alpha > 0$, 有:

$$f(x - \alpha \nabla f(x)) \leq f(x) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(x)\|^2.$$

因此, 只要

$$\alpha < \frac{2}{L},$$

函数值就严格下降: $f(x_{k+1}) < f(x_k)$ 。

1.8.3 步长 $\alpha \leq \frac{1}{L}$ 才能推出 $\mathcal{O}(1/k)$ 收敛

虽然 $\alpha \in (0, 2/L)$ 保证函数值下降, 但要得到函数值误差的 $\mathcal{O}(1/k)$ 收敛, 需要构造如下递推不等式:

$$f(x_k) - f^* \leq \frac{1}{2\alpha} (\|x_{k-1} - x^*\|^2 - \|x_k - x^*\|^2).$$

这是通过结合凸性与光滑性推出的, 其成立的关键在于以下不等式:

$$f(x - \alpha \nabla f(x)) \leq f(x) - \frac{\alpha}{2} \|\nabla f(x)\|^2,$$

其充要条件为:

$$\boxed{\alpha \leq \frac{1}{L}}.$$

这是因为从 Descent Lemma 出发, 有:

$$f(x - \alpha \nabla f(x)) \leq f(x) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(x)\|^2.$$

令其右边不小于 $\frac{\alpha}{c} \|\nabla f(x)\|^2$, 即有:

$$1 - \frac{L\alpha}{2} \geq \frac{1}{c} \Rightarrow \alpha \leq \frac{2}{L} \left(1 - \frac{1}{c}\right).$$

特别地，当 $c = 2$ 时，恰有：

$$f(x - \alpha \nabla f(x)) \leq f(x) - \frac{\alpha}{2} \|\nabla f(x)\|^2 \iff \alpha \leq \frac{1}{L}.$$

这是构造函数值下降量的核心工具，因此：

$$\text{只有在 } \alpha \leq \frac{1}{L} \text{ 时，才能推出 } \mathcal{O}(1/k) \text{ 的收敛速率。}$$

备注 1.8.1. 等价的关键在于为了构造函数值的次线性收敛，我们需要下降项系数为 $\frac{\alpha}{2}$ ，这样才能得到平方范数的递推式

$$f(x^{k+1}) \leq f^* + \frac{1}{2\alpha} (\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2).$$

1.8.4 误差线性收敛的条件与最优步长

若进一步假设 f 是 μ -强凸的，则梯度下降法可以实现线性收敛：

$$\|x_k - x^*\|^2 \leq \left(\frac{L - \mu}{L + \mu} \right)^{2k} \|x_0 - x^*\|^2,$$

其最优步长为：

$$\alpha^* = \frac{2}{L + \mu}.$$

这是误差收敛率最小化下的最优常数步长。

总结

- 在 $0 < \alpha < \frac{2}{L}$ 时，函数值下降，但只有 $\alpha \leq \frac{1}{L}$ 时能构造函数值的 $\mathcal{O}(1/k)$ 收敛递推。
- 在强凸条件下，最优步长为 $\alpha = \frac{2}{L + \mu}$ ，收敛速率为线性。
- 步长选大虽可能加快早期下降，但不能保证长期收敛性质；分析中选 $\alpha \leq \frac{1}{L}$ 最稳妥。

1.9 误差收缩与线性化算子谱半径的关系

在分析强凸 + 光滑函数上梯度下降法的误差收敛时，我们可以借助目标函数在最优点附近的二阶近似，得出如下误差迭代式：

$$e_{k+1} := x_{k+1} - x^* = x_k - x^* - \alpha \nabla f(x_k) = e_k - \alpha(\nabla f(x_k) - \nabla f(x^*)).$$

一、在 x^* 附近的线性化分析：由于 f 在 x^* 处二阶可导，可以对 $\nabla f(x_k)$ 做泰勒展开：

$$\nabla f(x_k) = \nabla f(x^*) + \nabla^2 f(x^*)(x_k - x^*) + \text{高阶项},$$

而 $\nabla f(x^*) = 0$ ，因此有近似：

$$\nabla f(x_k) \approx \nabla^2 f(x^*)e_k.$$

代入误差更新公式得到：

$$e_{k+1} \approx (I - \alpha \nabla^2 f(x^*))e_k.$$

二、误差收缩速率由谱模控制：此时的误差递推是线性系统：

$$e_{k+1} = T e_k, \quad T := I - \alpha \nabla^2 f(x^*).$$

因此，

$$\|e_{k+1}\| \leq \|T\| \cdot \|e_k\| = \rho(\alpha) \|e_k\|,$$

其中 $\rho(\alpha)$ 是算子 T 的谱模：

$$\rho(\alpha) = \max_{\lambda \in \sigma(\nabla^2 f(x^*))} |1 - \alpha\lambda|.$$

三、Hessian 的谱落在 $[\mu, L]$ 区间：由 f 的 μ -强凸性与 L -光滑性可知：

- μ -强凸性 $\Rightarrow \nabla^2 f(x) \succeq \mu I$ ，即 $\lambda_{\min}(\nabla^2 f(x)) \geq \mu$ ；
- L -光滑性 $\Rightarrow \nabla^2 f(x) \preceq L I$ ，即 $\lambda_{\max}(\nabla^2 f(x)) \leq L$ ；

因此：

$$\sigma(\nabla^2 f(x^*)) \subset [\mu, L], \quad \text{从而} \quad \rho(\alpha) = \max_{\lambda \in [\mu, L]} |1 - \alpha\lambda|.$$

四、结论：在一般 C^2 类强凸 + 光滑函数上，梯度下降误差线性化为：

$$e_{k+1} \approx (I - \alpha \nabla^2 f(x^*))e_k,$$

其收敛因子由谱模决定：

$$\|e_{k+1}\| \leq \max_{\lambda \in [\mu, L]} |1 - \alpha\lambda| \cdot \|e_k\|,$$

这为我们选取最优步长 $\alpha = \frac{2}{L+\mu}$ 提供了理论依据。

1.10 优化算法收敛性证明框架的两类方法

1.10.1 不动点类

定理 1.10.1 (Banach不动点定理). 设 (X, d) 是一个完备度量空间, $T : X \rightarrow X$ 是一个压缩映射, 即存在常数 $0 \leq c < 1$, 使得对任意 $x, y \in X$, 有

$$d(T(x), T(y)) \leq c \cdot d(x, y).$$

则 T 在 X 上存在唯一的不动点 $x^* \in X$, 即 $T(x^*) = x^*$. 此外, 对任意初始点 $x_0 \in X$, 迭代序列 $x_{n+1} = T(x_n)$ 将以指数速度收敛到 x^* , 即

$$d(x_n, x^*) \leq \frac{c^n}{1-c} d(x_1, x_0).$$

Banach不动点定理给了我们设计算法的思路。首先我们考虑单个函数的优化问题

$$\min f(x) \tag{1.1}$$

这里我们假设 $f(x)$ 是一个连续可微的函数。那么上述问题的解 x^* 就满足

$$\nabla f(x^*) = 0$$

如果我们想设计一个不动点迭代算法来求解这个问题, 我们可以考虑构造一个映射 $T(x)$, 使得

$$T(x^*) = x^*$$

那么这个 T 就应该是

$$T = Id() - \nabla f()$$

我们期待按照算法

$$x_{k+1} = T(x_k)$$

迭代的序列 x_k 会收敛到 x^* . 因此, 我们需要做的就是验证 T 是一个压缩映射。我们可以通过计算

$$\begin{aligned} \|T(x) - T(y)\|^2 &= \|x - y - \nabla f(x) + \nabla f(y)\|^2 \\ &\leq \|x - y\|^2 + \|\nabla f(x) - \nabla f(y)\|^2 - 2\langle x - y, \nabla f(x) - \nabla f(y) \rangle \end{aligned} \tag{1.2}$$

对于一般的函数 $f(x)$ 并不会使得 T 满足压缩性质。因此, 我们需要对 $f(x)$ 做一些假设。比如说存在 $\mu > 0, L > 0$ 使得

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq L\|x - y\|^2 \tag{1.3}$$

$$\langle x - y, \nabla f(x) - \nabla f(y) \rangle \geq \mu\|x - y\|^2 \tag{1.4}$$

于是(1.2)就变成了

$$d(T(x), T(y))^2 \leq (1 - 2\mu + L)\|x - y\|^2 \tag{1.5}$$

那么我们就得到下面定理

定理 1.10.2. 设 $f(x)$ 满足(1.3)和(1.4), 并且 $2\mu - 1 \leq L \leq 2\mu$, 那么 $T(x) = x - \nabla f(x)$ 是一个压缩映射。并且对于任意的 x_0 , 迭代序列

$$x_{k+1} = T(x_k)$$

收敛到 x^* 。

备注 1.10.1. 很容易看出(1.3)和(1.4)分别对应于Lipschitz连续性和单调性。也就是说我们可以通过对函数的Lipschitz连续性和强单调性来设计一个不动点迭代算法来求解优化问题。同时不动点迭代就是梯度下降法。类似的, 牛顿法也可以理解为不动点迭代。

接下来我们考虑一个更一般的优化问题

$$\min f(x) + g(x) \tag{1.6}$$

为了重点关注算法设计思路, 我们假设 f, g 都是足够光滑的。因此最优解满足

$$\nabla f(x^*) + \nabla g(x^*) = 0 \tag{1.7}$$

根据上式, 我们要将 x^* 设计为函数的不动点。下面有两种设计方式。

- (Forward-Barckward Splitting) 从(1.7)出发, 我们可以将其改写为

$$\nabla f(x^*) = -\nabla g(x^*),$$

这就给了我们一个设计的思路。我们可以将其改写为

$$x^* + \nabla f(x^*) = x^* - \nabla g(x^*),$$

进一步

$$x^* = (I + \nabla f)^{-1}(I - \nabla g)(x^*),$$

于是我们定义

$$T(x) = (I + \nabla f)^{-1}(I - \nabla g)(x),$$

这样我们就得到了一个不动点迭代算法

$$x_{k+1} = (I + \nabla f)^{-1}(I - \nabla g)(x_k).$$

引入中间变量 $y_k = (I - \nabla g)(x_k)$, 我们可以得到

$$y_{k+1} = (I - \nabla g)(x_k) \tag{1.8}$$

$$x_{k+1} = (I + \nabla f)^{-1}y_{k+1}. \tag{1.9}$$

这就是经典的FBS算法. 可以看出(1.8)是对 g 作梯度下降, 而(1.9)是找到沿着 f 作梯度上升后是 y_{k+1} 的点。我们可以将其理解为一个前向后向的迭代算法。

- (Peaceman -Rachford Splitting)

$$\begin{aligned}
0 = \nabla f(x) + \nabla g(x) &\iff 0 = (I + \alpha \nabla f)x - (I - \alpha \nabla g)x \\
&\iff 0 = (I + \alpha \nabla g)x - R_{\alpha \nabla g}(I + \alpha \nabla f)x, \quad R_{\alpha \nabla g} = 2(I + \alpha \nabla g)^{-1} - I \\
&\iff 0 = (I + \alpha \nabla f)x - R_{\alpha \nabla g}z, \quad z = (I + \alpha \nabla g)x \\
&\iff R_{\alpha \nabla g}z = (I + \alpha \nabla f)J_{\alpha \nabla g}z, \quad x = J_{\alpha \nabla g}z \\
&\iff J_{\alpha \nabla f}R_{\alpha \nabla g}z = J_{\alpha \nabla g}z \\
&\iff J_{\alpha \nabla f}R_{\alpha \nabla g}z = \frac{R_{\alpha \nabla g} + I}{2}z \\
&\iff 2J_{\alpha \nabla f}R_{\alpha \nabla g}z - R_{\alpha \nabla g}z = z \\
&\iff R_{\alpha \nabla f}R_{\alpha \nabla g}z = z
\end{aligned} \tag{1.10}$$

- (Douglas-Rachford Splitting)

$$\begin{aligned}
0 = \nabla f(x) + \nabla g(x) &\iff \left(\frac{R_{\alpha \nabla f}R_{\alpha \nabla g}}{2} + \frac{1}{2}\right)z = z \\
&\iff z = J_{\alpha \nabla f}(2J_{\alpha \nabla g} - I)z + (I - J_{\alpha \nabla g})z
\end{aligned} \tag{1.11}$$

1.10.2 能量函数类

首先还是考虑单个函数的优化问题

$$\min f(x) \tag{1.12}$$

假设 x^* 是最优解，为了证明算法迭代的收敛性，我们需要找到一个能量函数（一般称为Lyapunov函数），它刻画了当前迭代点和最优解之间的距离，我们将能量函数记作 $E(x^*, x^k)$ 并且要求

- $E(x^*, x^k) \geq 0$;
- $E(x^*, x^{k+1}) < E(x^*, x^k)$.

那么我们的算法就是在收敛的。

常见的能量函数有

- (1) $E(x^*, x^k) = \|x^k - x^*\|^2$;
- (2) $E(x^*, x^k) = f(x^k) - f(x^*)$;
- (3) $E(x^*, x^k) = f(x^*) - f(x^k) - \langle \nabla f(x^k), x^* - x^k \rangle$, Bregman 距离.

备注 1.10.2. 能量函数的选择考验直觉和经验。

1.11 Douglas-Rachford 分裂算法的等价形式

$$\begin{aligned}
& 0 = \nabla f(x) + \nabla g(x) \\
\iff & 0 = (I + \alpha \nabla f)x - (I - \alpha \nabla g)x \\
\iff & 0 = (I + \alpha \nabla g)x - R_{\alpha \nabla g}(I + \alpha \nabla f)x, \quad R_{\alpha \nabla g} = 2(I + \alpha \nabla g)^{-1} - I \\
\iff & 0 = (I + \alpha \nabla f)x - R_{\alpha \nabla g}z, \quad z = (I + \alpha \nabla g)x \\
\iff & R_{\alpha \nabla g}z = (I + \alpha \nabla f)J_{\alpha \nabla g}z, \quad x = J_{\alpha \nabla g}z \text{ 可以称为解映射} \\
\iff & J_{\alpha \nabla f}R_{\alpha \nabla g}z = J_{\alpha \nabla g}z \\
\iff & J_{\alpha \nabla f}R_{\alpha \nabla g}z = \frac{R_{\alpha \nabla g} + I}{2}z \\
\iff & 2J_{\alpha \nabla f}R_{\alpha \nabla g}z - R_{\alpha \nabla g}z = z \\
\iff & R_{\alpha \nabla f}R_{\alpha \nabla g}z = z \\
\iff & \left(\frac{R_{\alpha \nabla f}R_{\alpha \nabla g} + I}{2} \right) z = z \\
\iff & z = J_{\alpha \nabla f}(2J_{\alpha \nabla g} - I)z + (I - J_{\alpha \nabla g})z \tag{1.13}
\end{aligned}$$

1. $z_{k+1} = T(z_k)$, 其中 $T = \frac{R_{\alpha \nabla f}R_{\alpha \nabla g} + I}{2} + \frac{1}{2}$
2. $z_{k+1} = J_{\alpha \nabla f}(2J_{\alpha \nabla g} - I)z + (I - J_{\alpha \nabla g})z_k$
3. $z_{k+1} = z_k + \mathbf{prox}_{\alpha f}(2\mathbf{prox}_{\alpha g}(z_k) - z_k) - \mathbf{prox}_{\alpha g}(z_k)$
4. $z_{k+1} = (I + \alpha \partial f)^{-1}[(I - \alpha \partial f)(I + \alpha \partial g)^{-1} + \alpha \partial f]z_k$
5. $x_{k+1} = J_{\alpha \partial f}[J_{\alpha \partial g}(I - \alpha \partial f) + \alpha \partial f]x_k$

以上等价形式从不同角度揭示了 Douglas-Rachford 分裂方法的结构本质：既可以理解为两次反射的组合，也可以视为近端算子的协调操作，进而统一了最优化中的投影法、变分不等式法与不动点迭代法。

1.11.1 收敛性分析

DR 方法可以看成是一个不动点迭代，因此要证明收敛性，我们需要证明以下两个结论：

1. y_k 收敛到 $F(y)$ 的不动点 y^*
2. $x_{k+1} = \mathbf{prox}_f(z_k)$ 收敛到 $x^* = \mathbf{prox}_f(z^*)$

在证明收敛性之前，需要先定义两个映射：

$$\begin{aligned} F(z) &= z + \mathbf{prox}_g(2\mathbf{prox}_f(z) - z) - \mathbf{prox}_f(z), \\ G(z) &= z - F(z) = \mathbf{prox}_f(z) - \mathbf{prox}_g(2\mathbf{prox}_f(z) - z). \end{aligned}$$

我们要用到这两个函数的 **firmly nonexpansive** (co-coercive with parameter 1) 性质：

$$\begin{aligned} (F(z) - F(\hat{z}))^T(z - \hat{z}) &\geq \|F(z) - F(\hat{z})\|_2^2, \quad \forall z, \hat{z}, \\ (G(z) - G(\hat{z}))^T(z - \hat{z}) &\geq \|G(z) - G(\hat{z})\|_2^2, \quad \forall z, \hat{z}. \end{aligned}$$

证明. 令 $x = \mathbf{prox}_f(z)$, $\hat{x} = \mathbf{prox}_f(\hat{z})$,

$$\nu = \mathbf{prox}_g(2x - z), \quad \hat{\nu} = \mathbf{prox}_g(2\hat{x} - \hat{z}).$$

则根据

$$F(z) = z + \nu - x, \quad F(\hat{z}) = \hat{z} + \hat{\nu} - \hat{x}$$

有

$$\begin{aligned} (F(z) - F(\hat{z}))^T(z - \hat{z}) &\leq (z + \nu - x - \hat{z} - \hat{\nu} + \hat{x})^T(z - \hat{z}) - (x - \hat{x})^T(z - \hat{z}) + \|x - \hat{x}\|^2 \\ &= (\nu - \hat{\nu})^T(z - \hat{z}) + \|z - x - (\hat{z} - \hat{x})\|_2^2 \\ &= (\nu - \hat{\nu})^T(2x - z - 2\hat{x} + \hat{z}) - \|\nu - \hat{\nu}\|^2 + \|F(z) - F(\hat{z})\|^2 \\ &\geq \|F(z) - F(\hat{z})\|_2^2, \end{aligned}$$

其中最后一步用到了 $\mathbf{prox}_f, \mathbf{prox}_g$ 算子的 firm nonexpansiveness 性质：

$$(x - \hat{x})^T(z - \hat{z}) \geq \|x - \hat{x}\|_2^2, \quad (2x - z - 2\hat{x} + \hat{z})^T(\nu - \hat{\nu}) \geq \|\nu - \hat{\nu}\|_2^2.$$

同理可证 G 的 firm nonexpansiveness 性质。证毕。 \square

然后我们可以根据以下的不动点迭代公式证明前面提到的收敛性：

$$z_{k+1} = (1 - \rho_k) z_k + \rho_k F(z_k) = z_k - \rho_k G(z_k),$$

其中需假设 F 的不动点存在，且满足 $0 \in \partial f(x^*) + \partial g(x^*)$ ，以及松弛参数

$$\rho_k \in [\rho_{\min}, \rho_{\max}], \quad 0 < \rho_{\min} < \rho_{\max} < 2.$$

证明. 设 z^* 为 $F(z)$ 的不动点（也即 $G(z)$ 的零点），考虑 $\{z_k\}$ 步进化：

$$\|z_{k+1} - z^*\|_2^2 - \|z_k - z^*\|_2^2 = 2(z_{k+1} - z_k)^T(z_k - z^*) + \|z_{k+1} - z_k\|_2^2.$$

带入 $z_{k+1} = z_k - \rho_k G(z_k)$, 并利用 G 的 firm nonexpansiveness, 可得

$$\|z_{k+1} - z^*\|_2^2 \leq -\rho_k(2 - \rho_k)\|G(z_k)\|_2^2 \leq -M\|G(z_k)\|_2^2,$$

其中 $M = \rho_{\min}(2 - \rho_{\max}) > 0$ 。上述不等式说明

$$M \sum_{k=0}^{\infty} \|G(z_k)\|_2^2 \leq \|z_0 - z^*\|_2^2, \quad \|G(z_k)\| \rightarrow 0.$$

还可以得到 $\|z_k - z^*\|_2$ 是单调不增的, 因此有界。再由 $\|z_k - z^*\|_2$ 单调不增、故极限 $\lim_{k \rightarrow \infty} \|z_k - z^*\|_2$ 存在, 又由于有界, 故存在收敛子序列。

记 \bar{z} 为一个收敛子序列收敛到的极限点, 根据 G 的连续性有

$$0 = \lim_{k \rightarrow \infty} G(z_{k_j}) = G(\bar{z}),$$

即 \bar{z} 是 G 的零点, 且极限 $\lim_{k \rightarrow \infty} \|z_{k_j} - z^*\|_2$ 存在。

接着需要证明唯一性。假设 \bar{z}, \hat{z} 是两个不同的极限点, 收敛极限

$$\lim_{k \rightarrow \infty} \|z_k - \bar{z}\|_2, \quad \lim_{k \rightarrow \infty} \|z_k - \hat{z}\|_2$$

都存在, 因此

$$\|\bar{z} - \hat{z}\|_2 = \lim_{k \rightarrow \infty} \|z_k - \hat{z}\|_2 = \lim_{k \rightarrow \infty} \|z_k - \bar{z}\|_2 = 0.$$

从而 $\bar{z} = \hat{z}$, 即极限唯一。 □

- Fejér 单调性蕴含 $\{x_k\}$ 有界, 且对于任意 $p \in C$, 距离序列 $\{\|x_k - p\|\}$ 收敛;
- 结合 Bolzano–Weierstrass 引理 (有界序列存在收敛子列) 与极限点唯一性, 可推出全序列收敛, 且其极限落在 C 中的某一点。

1.12 收敛性分析: 基于Mann迭代的Douglas–Rachford算法

我们考虑求解以下形式的包含问题:

$$\text{find } x \in \mathcal{H} \text{ such that } 0 \in A(x) + B(x),$$

其中 \mathcal{H} 是Hilbert空间, $A, B: \mathcal{H} \rightrightarrows \mathcal{H}$ 为最大单调算子。

1.12.1 DR算法的等价迭代形式

我们定义两个反射算子：

$$R_A := 2J_{\lambda A} - I, \quad R_B := 2J_{\lambda B} - I,$$

其中 $J_{\lambda A} := (I + \lambda A)^{-1}$ 是A的近端映射。

1.12.2 最大单调性假设与约束资格

为了确保 $A + B$ 是最大单调的，我们采用如下假设：

- $A = \partial f, B = \partial g$, 其中 $f, g \in \Gamma_0(\mathcal{H})$;
- 且满足 Rockafellar 约束资格条件：

$$\text{dom}(f) \cap \text{int dom}(g) \neq \emptyset.$$

根据 Rockafellar 的定理，在上述条件下有：

$$\partial f + \partial g = \partial(f + g), \quad \text{且为最大单调算子.}$$

因此， $A + B$ 最大单调性成立，从而问题具有良定义的解集。

构造映射：

$$T := \frac{1}{2}(I + R_B R_A),$$

则 Douglas–Rachford 算法可表示为如下固定点迭代：

$$x^{k+1} = T(x^k).$$

1.12.3 Opial引理

引理 1.12.1 (Opial引理). 设 \mathcal{H} 为 Hilbert 空间， $T : \mathcal{H} \rightarrow \mathcal{H}$ 为非扩张映射， $\{x^k\}$ 是 T 的不动点迭代，即 $x^{k+1} = T(x^k)$ 。若满足：

1. $\text{Fix}(T) \neq \emptyset$;
2. 任意 $z \in \text{Fix}(T)$ ，序列 $\|x^k - z\|$ 收敛；
3. 所有弱收敛子序列极限均属于 $\text{Fix}(T)$;

则 $\{x^k\}$ 弱收敛于某个 $x^* \in \text{Fix}(T)$ 。

1.12.4 Mann迭代与非扩张算子

引理 1.12.2. 若 $T : \mathcal{H} \rightarrow \mathcal{H}$ 为非扩张映射, $\{\alpha_k\} \subset (0, 1)$ 满足

$$\sum_{k=0}^{\infty} \alpha_k(1 - \alpha_k) = \infty,$$

则 Mann 迭代

$$x^{k+1} = (1 - \alpha_k)x^k + \alpha_k T(x^k)$$

弱收敛于某个 $x^* \in \text{Fix}(T)$, 若 $\text{Fix}(T) \neq \emptyset$ 。

1.12.5 DR收敛性定理

定理 1.12.1 (Douglas–Rachford算法的弱收敛性). 设 A, B 为最大单调算子, 且解集 $\text{zer}(A+B) \neq \emptyset$ 。令

$$T := \frac{1}{2}(I + R_B R_A),$$

则 T 为非扩张映射。设初始点 $x^0 \in \mathcal{H}$, 生成序列 $\{x^k\}$ 如下:

$$x^{k+1} = T(x^k),$$

则该序列弱收敛于某个 $x^* \in \text{Fix}(T)$, 且

$$z^* := J_{\lambda B}(x^*) \in \text{zer}(A + B).$$

证明. 我们依次验证 Opial 引理三点条件:

1. 由于 $\text{zer}(A + B) \neq \emptyset$, 从而 $\text{Fix}(T) \neq \emptyset$;
2. T 为非扩张映射, 故 $\|x^{k+1} - z\| \leq \|x^k - z\|$, 由 Fejér 单调性可知 $\|x^k - z\|$ 收敛;
3. 若 $x^{k_j} \rightharpoonup x^\infty$, 则由非扩张映射与 Demiclosedness 原理 (即 $T - I$ 为 demiclosed at 0), 可得 $x^\infty \in \text{Fix}(T)$ 。

由 Opial 引理可知, $\{x^k\}$ 弱收敛于某个不动点 $x^* \in \text{Fix}(T)$ 。

进一步, 由 $x^* \in \text{Fix}(T) \Rightarrow z^* := J_{\lambda B}(x^*) \in \text{zer}(A + B)$ (经典结果) 得证。 \square

1.13 全序列收敛的常用技巧

在优化算法收敛性证明中，除“极限点唯一 + 有界 + 收敛子列极限相同”之外，还有以下几种典型手段可直接或间接得到全序列收敛：

1. **Fejér (或 quasi-Fejér) 单调性** 证明迭代序列 $\{x_k\}$ 相对于某凸集 C 满足

$$\|x_{k+1} - p\| \leq \|x_k - p\|, \quad \forall p \in C,$$

或带有可控误差项的 quasi-Fejér 单调，从而推得有界性、距离收敛，再由极限点唯一性得出全序列收敛。

2. **Opial 引理 (Hilbert 空间)** 若 $\{x_k\}$ 满足

- 对任意 $z \in C$, $\lim_{k \rightarrow \infty} \|x_k - z\|$ 存在;
- 所有聚点均属于某非空闭凸集 C ,

则 x_k 强收敛到 C 中某一点。

3. **KL (Kurdyka–Łojasiewicz) 性质** 若目标函数 Φ 在邻域内满足 KL 不等式，并且迭代具有“足够下降”及“子梯度有界”性质，则序列在有限长度内收敛，进而强收敛。

4. **严格下降 + Cauchy 性** 当能证明

$$f(x_k) - f(x_{k+1}) \geq \gamma \|x_{k+1} - x_k\|^2, \quad \sum_k (f(x_k) - f(x_{k+1})) < \infty,$$

则 $\|x_{k+1} - x_k\| \rightarrow 0$ ，由有界性与完备性推出全序列收敛。

5. **PL 条件 / 强凸性** 若 f 满足 Polyak–Łojasiewicz 条件或强凸性质，可直接获得线性收敛率 $f(x_{k+1}) - f^* \leq (1 - \mu\alpha)(f(x_k) - f^*)$ 进而全序列收敛。
6. **Demiclosed-at-zero 原则** 对非扩张映射 T ，若 $x_k - T(x_k) \rightarrow 0$ 且 $x_{k_j} \rightharpoonup \bar{x}$ ，则 $\bar{x} = T(\bar{x})$ 。配合 Opial 条件或唯一极限点推弱/强收敛。
7. **投影不动点与距离单调** 将迭代写作 $x_{k+1} = T(x_k)$ ，并证明 $\|x_{k+1} - x^*\| \leq \|x_k - x^*\| - \delta \|x_{k+1} - x_k\|^2$ ，利用三项不等式或 Fejér 单调，得全序列收敛。

1.14 Scaling 缩放技巧

1.15 Kurdyka-Łojasiewicz 条件

KL 条件是用于分析非凸优化问题中算法收敛性的强有力工具，尤其在凸但不可微或非凸优化问题中起着核心作用。KL 条件本质上是函数在临界点附近的一种“渐进良性行为”的刻画。

定义 1.15.1 (KL 函数). 设 $\phi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ 是下半连续的，并且在某点 x^* 附近是有限值。若存在：

- 一个邻域 U 使得 $x^* \in U$;
- 一个常数 $\eta \in (0, +\infty]$;
- 一个函数 $\varphi \in \Phi_\eta$ ，其中

$$\Phi_\eta := \{\varphi \in C^0([0, \eta]) \cap C^1((0, \eta)) \mid \varphi(0) = 0, \varphi' > 0\};$$

使得对于所有 $x \in U$ 满足 $x \neq x^*$ 且 $\phi(x^*) < \phi(x) < \phi(x^*) + \eta$ ，都有

$$\varphi'(\phi(x) - \phi(x^*)) \cdot \|\partial\phi(x)\| \geq 1,$$

其中 $\partial\phi(x)$ 表示 *Clarke* 次微分或 *Fréchet* 次微分。

则称 ϕ 在 x^* 附近满足 **KL 条件**。

备注 1.15.1. KL 条件并不要求目标函数是凸的或光滑的，只要满足一定的“正则性”。很多常见的非凸函数，如半代数函数 (*polynomial, piecewise linear, ℓ_1 范数*等) 都满足 KL 条件。

示例 1.15.1 (KL 函数的形式). 常用的 KL 函数 $\varphi(s)$ 形式如下：

$$\varphi(s) = cs^{1-\theta}, \quad \theta \in [0, 1), c > 0.$$

这种形式能反映不同类型的收敛速度：

- $\theta = 0$ 时：有限步收敛；
- $\theta \in (0, \frac{1}{2}]$ ：线性收敛；
- $\theta \in (\frac{1}{2}, 1)$ ：亚线性收敛。

备注 1.15.2. 如果一个优化问题的目标函数满足 KL 条件，并且算法具有适当的下降性质，则可以几乎自动推导出全局收敛性和速率。

在基于 KL 性质证明迭代算法收敛性的过程中，通常需要借助以下三个关键不等式：

1. KL 不等式（梯度模长 \geq 目标函数值下降量）：

$$\|\nabla f(x_k)\| \geq \varphi'(f(x_k) - f^*),$$

其中 φ 是 KL 函数， f^* 表示目标函数的极小值。该不等式表明，当函数值靠近极小值时，梯度趋于零。

2. 充分下降：

$$|f(x_k) - f(x_{k+1})| \geq \alpha \|x_k - x_{k+1}\|^2,$$

其中 $\alpha > 0$ 是某个固定常数。此不等式说明函数值在每次迭代中具有充分下降性，是算法收敛的重要依据。

3. 梯度模长估计：

$$\|x_k - x_{k+1}\| \geq \beta \|\nabla f(x_k)\|,$$

其中 $\beta > 0$ 是某个固定常数。该不等式用于连接变量变化幅度与梯度大小，从而与 KL 不等式结合，形成收敛性的闭环链式估计。

1.16 Farkas 引理

Farkas 引理是凸分析、线性规划和优化理论中的基础结果，描述了线性系统可行性的一种二择一关系。

引理 1.16.1 (Farkas 引理, 标准版). 设 $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. 恰好有且只有以下两种情况之一成立:

1. 存在 $x \in \mathbb{R}^n$, 使得 $Ax = b$ 且 $x \geq 0$;
2. 存在 $y \in \mathbb{R}^m$, 使得 $A^T y \geq 0$ 且 $b^T y < 0$.

直观理解:

- 要么系统 $Ax = b, x \geq 0$ 有解;
- 要么存在一个向量 y , 作为“证伪者”, 证明无解。

本质上反映了凸集的“要么有交集, 要么可以被超平面严格分开”的原理。

1.17 动量法与外推步

在连续优化中, 标准梯度下降法 (GD) 可看作无质量粒子在负梯度力场中移动, 由于对当前梯度的反应是即时的, 在条件数较差的地形中易产生剧烈震荡。

引入动量 (Momentum) 或惯性 (Inertial Acceleration), 相当于给粒子赋予质量。粒子不仅受当前梯度影响, 还会保持之前运动的速度。这使得粒子能冲过平坦的局部极小值, 抵消震荡, 沿着真正下降方向加速。

1.17.1 多项式重球法 (Polyak's Heavy Ball)[16]

- 更新公式:

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$$

- 机制: 盲目冲锋。先评估当前点 x_k 的梯度, 再结合上一时刻的惯性 ($x_k - x_{k-1}$) 决定下一步。缺点是容易在极小值附近冲过头。

1.17.2 Nesterov 加速梯度法 (NAG)

- 更新公式:

$$y_k = x_k + \beta_k(x_k - x_{k-1})$$

$$x_{k+1} = y_k - \alpha \nabla f(y_k)$$

- **机制：** 预判修正。先按历史惯性推算出一个预测点 y_k ，然后在 y_k 处计算梯度。这种“向前看”的能力自带“刹车”机制，将凸优化的收敛率提升至理论下界 $O(1/k^2)$ 。

1.17.3 外推法 (Extrapolation Method)

在优化语境下，狭义的“外推法”通常指 **外推梯度法 (Extragradiant Method)**，由 Korpelevich (1976) 提出 [11]。它与 Nesterov 都有“探路”的思想，但机制完全不同。

- **更新公式：**

$$y_k = x_k - \alpha \nabla f(x_k) \quad (\text{外推试探步})$$

$$x_{k+1} = x_k - \alpha \nabla f(y_k) \quad (\text{真实更新步})$$

- **核心区别：** Nesterov 的外推是基于**历史惯性**（不用算额外梯度）；而真正的外推法是基于**当前梯度**先试走一步，在试探点再算一次梯度。
- **应用场景：** 虽然单步计算量翻倍（算两次梯度），但在处理极小极大问题（Min-Max Problems，如生成对抗网络 GANs 的鞍点优化）时，外推法能有效抑制旋转向量场中的发散现象，是不可或缺的利器。

1.17.4 连续时间视角：ODE 推导与渐隐阻尼

将离散算法取连续时间极限，是理解加速现象的本质方法。通过有限差分近似，并设定参数缩放 $\beta = 1 - \gamma \Delta t$ 且 $\alpha = \Delta t^2$ ，当 $\Delta t \rightarrow 0$ 时，重球法收敛于：

$$\ddot{x}(t) + \gamma \dot{x}(t) + \nabla f(x(t)) = 0$$

这对应了经典力学：**加速度 + 恒定摩擦阻尼 = 负梯度动力**。

Nesterov 方法 (Su, Boyd, Candes, 2014) 的精髓在于其动量参数 β_k 不是常数，而是随迭代次数动态变化的，通常取 $\beta_k = \frac{k-1}{k+2}$ [22]。将其进行类似的有限差分连续化推导，会得到一个极具启发性的常微分方程：

$$\ddot{x}(t) + \frac{3}{t} \dot{x}(t) + \nabla f(x(t)) = 0$$

物理意义解析：

- 这里的阻尼系数是 $\frac{3}{t}$ ，被称为**渐隐阻尼 (Vanishing Damping)**。
- **早期 (t 较小)：** 阻尼非常大，相当于把重球放在粘稠的蜂蜜中，防止算法在初期由于梯度过大而乱跑或剧烈震荡。
- **后期 (t 变大)：** 阻尼趋近于 0，重球获得了极大的动能（惯性），能够在平坦区域快速滑行，逼近最优解。这就完美解释了为什么 NAG 具有极致的加速效果。

1.17.5 总结对比

算法	探路依据	梯度计算点	核心优势
动量法 (Momentum)	无探路, 看脚下	当前点 x_k	翻越局部极小, 加速下坡
外推梯度法 (Extragradient)	用 当前梯度 试探	试探点 y_k	解决鞍点/极小极大问题的发散
Nesterov (NAG)	用 历史惯性 预判	预判点 y_k	最优的凸优化收敛率, 自带刹车

Chapter 2

Diffusion

2.1 from DDPM to DDIM

命题 2.1.1. 已知在去噪扩散隐式模型 (DDIM)[19]中, 前向过程[9]的边缘分布为 $q(x_t | x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I)$ 。反向去噪的条件分布定义为高斯分布:

$$q(x_{t-1} | x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \sigma_t^2 I) \quad (2.1)$$

且其均值 $\tilde{\mu}_t(x_t, x_0)$ 满足:

$$\tilde{\mu}_t(x_t, x_0) = \sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \frac{x_t - \sqrt{\bar{\alpha}_t}x_0}{\sqrt{1 - \bar{\alpha}_t}} \quad (2.2)$$

证明. Step 1: 基于高斯分布的性质, 我们假设 x_{t-1} 可以由 x_0, x_t 以及一个独立的标准高斯噪声 $\epsilon \sim \mathcal{N}(0, I)$ 线性组合而成:

$$x_{t-1} = ax_0 + bx_t + c\epsilon$$

由于已知给定 x_t, x_0 时分布的方差为 $\sigma_t^2 I$, 而上述假设中唯一的随机项是 $c\epsilon$, 因此条件方差为 $c^2 I$ 。故可直接得出:

$$c = \sigma_t$$

Step 2: 利用重参数化展开 x_t 根据边缘分布已知条件 1, 我们可以用重参数化技巧将 x_t 表示为:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_t$$

(其中 $\epsilon_t \sim \mathcal{N}(0, I)$ 且与 ϵ 相互独立) 将 x_t 代入 Step 1 的线性假设中, 得到仅包含 x_0 和噪声项的表达式:

$$x_{t-1} = ax_0 + b(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_t) + \sigma_t\epsilon$$

合并同类项:

$$x_{t-1} = (a + b\sqrt{\bar{\alpha}_t})x_0 + b\sqrt{1 - \bar{\alpha}_t}\epsilon_t + \sigma_t\epsilon$$

Step 3: 匹配边缘分布 $q(x_{t-1}|x_0)$ 的均值与方差为了使推导成立, 上式必须与边缘分布已知条件 2 的均值和方差完全一致。匹配方差求 b : 上式右侧随机项为 $b\sqrt{1-\bar{\alpha}_t}\epsilon_t + \sigma_t\epsilon$ 。因噪声独立, 方差直接相加, 并使其等于目标分布方差 $(1-\bar{\alpha}_{t-1})$:

$$b^2(1-\bar{\alpha}_t) + \sigma_t^2 = 1-\bar{\alpha}_{t-1}$$

解得:

$$b = \frac{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}}$$

匹配均值求 a : 上式右侧 x_0 的系数必须等于目标分布的均值系数 $\sqrt{\bar{\alpha}_{t-1}}$:

$$a + b\sqrt{\bar{\alpha}_t} = \sqrt{\bar{\alpha}_{t-1}}$$

解得:

$$a = \sqrt{\bar{\alpha}_{t-1}} - b\sqrt{\bar{\alpha}_t}$$

Step 4: 代入系数, 得出最终均值回到 Step 1 的线性假设, 当给定 x_t 和 x_0 时, 噪声项期望为 0。因此条件分布的均值 $\tilde{\mu}_t$ 为:

$$\tilde{\mu}_t(x_t, x_0) = ax_0 + bx_t$$

将 Step 3 得到的 a 代入:

$$\tilde{\mu}_t(x_t, x_0) = (\sqrt{\bar{\alpha}_{t-1}} - b\sqrt{\bar{\alpha}_t})x_0 + bx_t$$

提取公因式 b :

$$\tilde{\mu}_t(x_t, x_0) = \sqrt{\bar{\alpha}_{t-1}}x_0 + b(x_t - \sqrt{\bar{\alpha}_t}x_0)$$

最后, 将 b 的完整表达式代入:

$$\tilde{\mu}_t(x_t, x_0) = \sqrt{\bar{\alpha}_{t-1}}x_0 + \frac{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}}(x_t - \sqrt{\bar{\alpha}_t}x_0)$$

证毕 □

命题 2.1.2 (DDIM 采样公式的推导). *DDIM*在反向采样过程中, 若由神经网络 $\epsilon_\theta(x_t)$ 预测给定 x_t 下的噪声, 从 x_t 生成 x_{t-1} 的迭代公式为:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} \cdot \epsilon_\theta(x_t) + \sigma_t\epsilon \quad (2.3)$$

其中 $\epsilon_t \sim \mathcal{N}(0, I)$ 为标准高斯噪声。

证明. 由前向过程的边缘分布 $q(x_t | x_0)$, 应用重参数化技巧, 可将 x_t 表达为:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \quad (2.4)$$

其中 $\epsilon \sim \mathcal{N}(0, I)$ 为前向过程加入的真实噪声。

在反向生成过程中, x_0 且真实噪声 ϵ 均未知。以训练所得的神经网络 $\epsilon_\theta(x_t)$ 作为对真实噪声 ϵ 的近似估计。对上式进行代数变换, 可得原始图像 x_0 的估计值 \hat{x}_0 :

$$\hat{x}_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t)}{\sqrt{\bar{\alpha}_t}} \quad (2.5)$$

根据已知的反向条件分布 $q(x_{t-1} | x_t, x_0)$, 引入独立的高斯随机变量 $\epsilon_t \sim \mathcal{N}(0, I)$, 则随机变量 x_{t-1} 的一次采样可写为:

$$x_{t-1} = \tilde{\mu}_t(x_t, \hat{x}_0) + \sigma_t \epsilon_t \quad (2.6)$$

将预测的原图 \hat{x}_0 代入给定的均值等式 $\tilde{\mu}_t(x_t, \hat{x}_0)$ 中, 得:

$$\tilde{\mu}_t(x_t, \hat{x}_0) = \sqrt{\bar{\alpha}_{t-1}} \hat{x}_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \frac{x_t - \sqrt{\bar{\alpha}_t} \hat{x}_0}{\sqrt{1 - \bar{\alpha}_t}} \quad (2.7)$$

考察上式第二项中的分子部分。由 \hat{x}_0 的定义式反推可得:

$$x_t - \sqrt{\bar{\alpha}_t} \hat{x}_0 = \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t) \quad (2.8)$$

将该等式代入 $\tilde{\mu}_t(x_t, \hat{x}_0)$ 的表达式中, 消去分母 $\sqrt{1 - \bar{\alpha}_t}$, 得化简后的均值:

$$\tilde{\mu}_t(x_t, \hat{x}_0) = \sqrt{\bar{\alpha}_{t-1}} \hat{x}_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(x_t) \quad (2.9)$$

最后, 将化简后的均值代入 x_{t-1} 的采样方程, 并将 \hat{x}_0 的完整定义式展开代入第一项, 整理即得最终的采样公式:

$$x_{t-1} = \underbrace{\sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \right)}_{\text{预测的 } x_0 \text{ 及其缩放}} + \underbrace{\sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(x_t, t)}_{\text{指向 } x_t \text{ 的方向 (预测的噪声)}} + \underbrace{\sigma_t \epsilon_t}_{\text{随机噪声}} \quad (2.10)$$

□

2.2 Flow Matching 中边缘速度场公式的推导与理解

在 Flow Matching (FM) 中, 边缘速度场 (Marginal Velocity Field) $u_t(x)$ 的公式如下:

$$u_t(x) = \int u_t(x|x_1) p_{1|t}(x_1|x) dx_1 \quad (2.11)$$

该公式的核心来源是概率流的连续性方程 (Continuity Equation)。为了让边缘分布 $p_t(x)$ 能够随着时间正确演化, 边缘速度场 $u_t(x)$ 必须是所有可能的“条件速度场” $u_t(x|x_1)$ 在给定当前位置 x 时的后验期望。

2.2.1 数学推导：基于连续性方程

在流模型中，概率密度的演化由连续性方程控制。对于给定目标样本 x_1 时的条件分布 $p_{t|1}(x|x_1)$ ，其对应的条件速度场 $u_t(x|x_1)$ 满足以下连续性方程：

$$\frac{\partial}{\partial t} p_{t|1}(x|x_1) + \nabla \cdot (u_t(x|x_1) p_{t|1}(x|x_1)) = 0 \quad (2.12)$$

我们的目标是找到一个边缘速度场 $u_t(x)$ ，使其能驱动边缘分布 $p_t(x)$ ，即满足边缘连续性方程：

$$\frac{\partial}{\partial t} p_t(x) + \nabla \cdot (u_t(x) p_t(x)) = 0 \quad (2.13)$$

推导过程如下：

首先，边缘分布 $p_t(x)$ 是条件分布在所有可能的目标数据 $x_1 \sim p_1(x_1)$ 上的积分：

$$p_t(x) = \int p_{t|1}(x|x_1) p_1(x_1) dx_1 \quad (2.14)$$

对等式两边同时对时间 t 求偏导：

$$\frac{\partial}{\partial t} p_t(x) = \int \frac{\partial}{\partial t} p_{t|1}(x|x_1) p_1(x_1) dx_1 \quad (2.15)$$

将条件连续性方程代入（即替换 $\frac{\partial}{\partial t} p_{t|1}$ ）：

$$\frac{\partial}{\partial t} p_t(x) = - \int \nabla \cdot (u_t(x|x_1) p_{t|1}(x|x_1)) p_1(x_1) dx_1 \quad (2.16)$$

由于散度算子 $\nabla \cdot$ 是对 x 求导，而积分是对 x_1 进行的，我们可以将散度算子提取到积分号外：

$$\frac{\partial}{\partial t} p_t(x) = - \nabla \cdot \left(\int u_t(x|x_1) p_{t|1}(x|x_1) p_1(x_1) dx_1 \right) \quad (2.17)$$

根据概率的乘法公式（或贝叶斯公式的分子部分），联合概率密度满足 $p_{t|1}(x|x_1) p_1(x_1) = p(x, x_1) = p_{1|t}(x_1|x) p_t(x)$ 。将其代入积分中：

$$\frac{\partial}{\partial t} p_t(x) = - \nabla \cdot \left(\int u_t(x|x_1) p_{1|t}(x_1|x) p_t(x) dx_1 \right) \quad (2.18)$$

由于边缘概率 $p_t(x)$ 与积分变量 x_1 无关，可以将其提取到积分号外：

$$\frac{\partial}{\partial t} p_t(x) = - \nabla \cdot \left(\left[\int u_t(x|x_1) p_{1|t}(x_1|x) dx_1 \right] p_t(x) \right) \quad (2.19)$$

将上式与目标边缘连续性方程 $\frac{\partial}{\partial t} p_t(x) = - \nabla \cdot (u_t(x) p_t(x))$ 对比。要使等式成立，方括号内的部分必然就是边缘速度场 $u_t(x)$ ：

$$u_t(x) = \int u_t(x|x_1) p_{1|t}(x_1|x) dx_1 \quad (2.20)$$

2.2.2 直观物理解

抛开繁琐的推导，等式右侧实际上是一个期望值：

$$u_t(x) = \mathbb{E}_{x_1 \sim p_{1|t}(\cdot|x)}[u_t(x|x_1)] \quad (2.21)$$

从“粒子运动”的视角可以直观地理解这个期望：

- **缺乏全局信息：**假设你是一个在时间 t 走到位置 x 的粒子，你需要决定下一步的总体速度 $u_t(x)$ ，但你不知道你最终确切要走到终点样本 x_1 是谁。
- **条件速度 $u_t(x|x_1)$ ：**如果明确告诉你终点是某个特定的 x_1 ，你就有明确的运动方向和速度 $u_t(x|x_1)$ 。
- **后验概率猜想 $p_{1|t}(x_1|x)$ ：**既然不知道确切终点，你只能根据当前所在的位置 x ，去“反推”或“猜测”终点 x_1 可能是谁。这个概率分布就是后验概率 $p_{1|t}(x_1|x)$ 。
- **求加权平均：**最终你的决定，就是把“去往所有可能终点的速度向量”，按照“那个终点有多大可能是真的终点”来进行加权平均（求期望）。

综上，边缘速度场本质上就是在给定当前状态 x 的情况下，对所有可能条件速度场的加权期望。

2.3 从sde到ode

2.3.1 方法一

设定前向加噪过程为 Ito SDE: $dx = f(x, t)dt + g(t)dw$ ，其中 $f(x, t)$ 为漂移项， $g(t)$ 为扩散项。该过程诱导的边缘概率密度 $p_t(x)$ 随时间的演化由 Fokker-Planck (F-P) 方程精确描述：

$$\frac{\partial p_t(x)}{\partial t} = -\nabla \cdot [f(x, t)p_t(x)] + \frac{1}{2}g(t)^2\Delta p_t(x)$$

为了寻找等效的确定性轨迹，我们将上式改写为流体力学中的连续性方程形式：

$$\frac{\partial p_t}{\partial t} = -\nabla \cdot (p_t \mathbf{v})$$

。利用恒等式 $\Delta p_t = \nabla \cdot (\nabla p_t)$ 变换 F-P 方程的二阶项，并结合得分函数（Score Function）定义 $\nabla \log p_t(x) = \frac{\nabla p_t}{p_t}$ ，可将方程右侧合并为：

$$\frac{\partial p_t(x)}{\partial t} = -\nabla \cdot \left[p_t(x) \left(f(x, t) - \frac{1}{2}g(t)^2 \nabla \log p_t(x) \right) \right]$$

由此定义出确定性速度场 $\mathbf{v}(x, t) = f(x, t) - \frac{1}{2}g(t)^2 \nabla \log p_t(x)$ 。最终得到与原 SDE 边缘分布完全一致的速率流 ODE[21]：

$$dx = \left[f(x, t) - \frac{1}{2}g(t)^2 \nabla \log p_t(x) \right] dt$$

2.3.2 方法二

1. 概率密度的期望表示

利用狄拉克函数的性质，将边缘分布表示为期望：

$$p_t(x) = \mathbb{E}_{x_t}[\delta(x - x_t)]$$

2. 泰勒展开与期望消去

考虑微小时间步 Δt ，前向过程为：

$$x_{t+\Delta t} = x_t + f_t(x_t)\Delta t + g_t\sqrt{\Delta t}\epsilon$$

对 $\delta(x - x_{t+\Delta t})$ 做泰勒展开（保留到 $O(\Delta t)$ ）：

$$\begin{aligned} \delta(x - x_{t+\Delta t}) &\approx \delta(x - x_t) \\ &\quad - (f_t\Delta t + g_t\sqrt{\Delta t}\epsilon) \cdot \nabla_x \delta(x - x_t) \\ &\quad + \frac{1}{2}(g_t\sqrt{\Delta t}\epsilon \cdot \nabla_x)^2 \delta(x - x_t) \end{aligned}$$

对两边取期望，并利用：

$$\mathbb{E}[\epsilon] = 0, \quad \mathbb{E}[\epsilon\epsilon^T] = I$$

得到：

$$p_{t+\Delta t}(x) = p_t(x) - \nabla_x \cdot [f_t p_t(x)\Delta t] + \frac{1}{2}g_t^2\Delta t \nabla_x \cdot \nabla_x p_t(x)$$

3. 引入 σ_t 的等效族

推广得到一般形式（Fokker-Planck 方程）：

$$\frac{\partial p_t}{\partial t} = -\nabla_x \cdot \left[\left(f_t - \frac{1}{2}(g_t^2 - \sigma_t^2)\nabla_x \log p_t \right) p_t \right] + \frac{1}{2}\sigma_t^2 \nabla_x \cdot \nabla_x p_t$$

4. 概率流 ODE

取 $\sigma_t = 0$ ，扩散项消失，得到确定性 ODE：

$$dx = \left[f_t(x) - \frac{1}{2}g_t^2 \nabla_x \log p_t(x) \right] dt$$

2.4 扩散模型核心架构

2.4.1 U-net

U-net 作为去噪网络骨干在扩散模型中被广泛使用 [18, 9]。

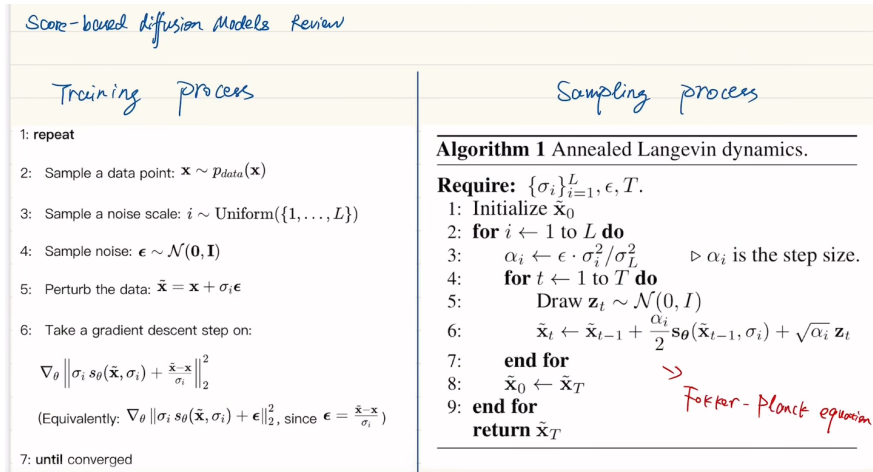


图 2.1: score matching [10, 21]

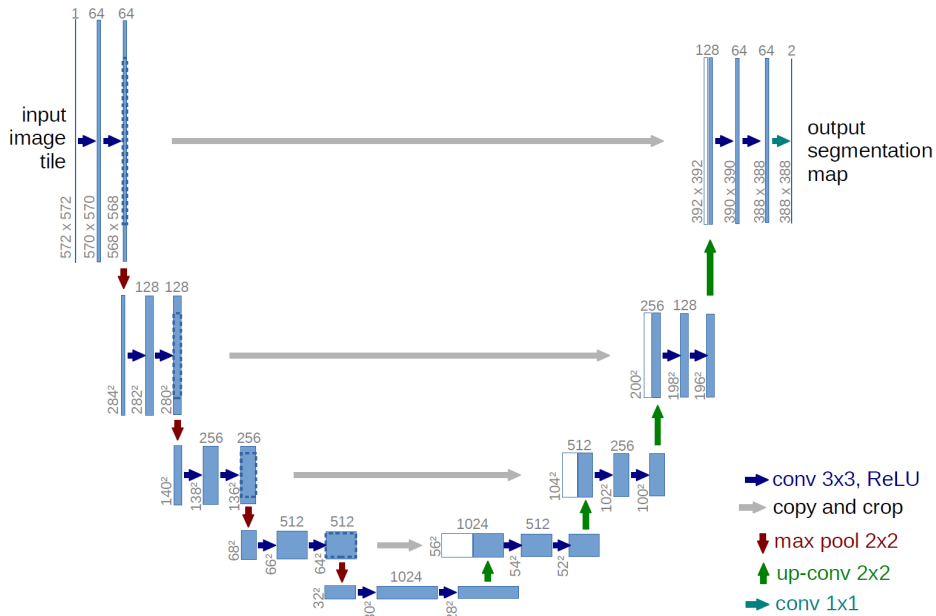


图 2.2: U-net

2.5 Diffusion-Distillation DPO

这里的 DPO 指 Direct Preference Optimization [17]。

2.5.1 Diffusion-Distillation Progressive

1. 假设教师模型为 f_{ϕ} ，学生模型为 f_{θ} 。在时间步 t ， z_t 为带噪图像。蒸馏的目标是让 学生模型通过一步映射，达到教师模型两步采样的效果。定义时间步序列为：

$$t_n > t_{n-0.5} > t_{n-1}$$

2. 教师模型的轨迹计算 (Teacher Trajectory) 教师模型从 z_{t_n} 出发, 经过两次确定性采样 (如 DDIM) 得到目标点 $z_{t_{n-1}}^T$: 第一步 ($t_n \rightarrow t_{n-0.5}$):

$$z_{t_{n-0.5}} = \text{Sample}_\phi(z_{t_n}, t_n, t_{n-0.5})$$

第二步 ($t_{n-0.5} \rightarrow t_{n-1}$):

$$z_{t_{n-1}}^T = \text{Sample}_\phi(z_{t_{n-0.5}}, t_{n-0.5}, t_{n-1})$$

其中, Sample_ϕ 通常定义为:

$$z_s = \sqrt{\frac{\alpha_s}{\alpha_t}} z_t + \left(\sqrt{1 - \alpha_s} - \sqrt{\frac{\alpha_s(1 - \alpha_t)}{\alpha_t}} \right) f_\phi(z_t, t)$$

3. 学生模型的一步预测 (Student Mapping) 学生模型 f_θ 接收相同的输入 z_{t_n} , 直接预测目标状态:

$$\hat{z}_{t_{n-1}}^S = f_\theta(z_{t_n}, t_n)$$

4. 蒸馏损失函数 (Distillation Loss) 在训练过程中, 我们最小化学生输出与教师两步采样结果之间的平方误差:

$$\mathcal{L}_{distill} = \mathbb{E}_{z \sim q, t \in \{2, 4, \dots, N\}} \left[w(t) \left\| f_\theta(z_{t_n}, t_n) - \text{stop_grad}(z_{t_{n-1}}^T) \right\|_2^2 \right]$$

stop_grad: 教师模型的输出不参与梯度更新。 $w(t)$: 权重系数, 通常设为 1 或基于信号对比度 (SNR) 的加权。

5. 递归更新逻辑 (Recursive Update) 在完成一轮蒸馏 (如步数从 $N \rightarrow N/2$) 后, 将当前的学生模型作为下一轮的教师模型:

$$\phi \leftarrow \theta$$

并重复上述过程, 直到步数缩减至目标范围 (如 4 步或 1 步)。

2.5.2 Distillation-DPO改进方向

1. 从“加速”迈向“实时”: 引入非扩散或流匹配架构目前 Distillation-DPO 虽然将推理时间从 17.87s 压缩到了 3.38s (8 步采样), 但这对于高速行驶的自动驾驶车辆 (10Hz 实时性要求) 仍显不足。改进方向: 可以考虑将**蒸馏目标从扩散模型转向流匹配 (Flow Matching) 框架**。流匹配学习的是从噪声到数据的线性轨迹, 天然比扩散模型的弯曲路径更容易实现单步 (1-step) 推理。借鉴方法: 参考 LiNeXt, 这种非扩散架构通过单次前向传递 (N2C 模块) 即可实现 0.09s 的推理速度, 远超目前的 8 步蒸馏方案。

2. 引入保障机制: 防止胜者分支的绝对质量退化, DPO 通过拉大胜者和败者的对数似然差值来优化。但最新研究 Diffusion-SDPO 指出, 盲目扩大偏好边际 (Margin) 可能导

致胜者分支的重构误差反而上升。改进方向：引入保障性更新规则（Safeguarded Update Rule），根据胜者和败者梯度的几何对齐情况动态调整缩放系数，确保在优化偏好的同时，高质量样本的绝对物理精度（CD 指标）非增。

3. 提高轨迹恢复精度：由前向近似转向逆转恢复在推导时，为了计算方便，用前向扩散过程 $q(x_{1:T}|x_0)$ 近似了逆向推理轨迹 $p_\eta(x_{1:T}|x_0)$ 。这种近似会导致计算精度损失，尤其是在步数极少的情况下。改进方向：采用 Inversion-DPO 的思想，利用**确定的 DDIM 逆转**（Inversion）轨迹来精确找回从胜者/败者样本到初始噪声的对应路径。这可以将原本的四个 KL 散度项简化为两个精确匹配项，使训练收敛速度提升 2 倍以上。

5. 增强物理连贯性：引入“二阶惯性”轨迹当前的蒸馏模型主要关注一阶位置更新。在 3D 场景中，点云的补全路径如果能符合物理规律，生成的结构会更稳健。改进方向：借鉴 OAT-FM（最优加速度传输）理论，在生成路径中引入加速度（Acceleration）或惯性（Inertia）项。通过最小化物理作用量，使点云的移动轨迹更加笔直且符合机械动力学，从而缓解 3D 补全中常见的细节模糊问题。

2.6 基于强化学习的扩散模型

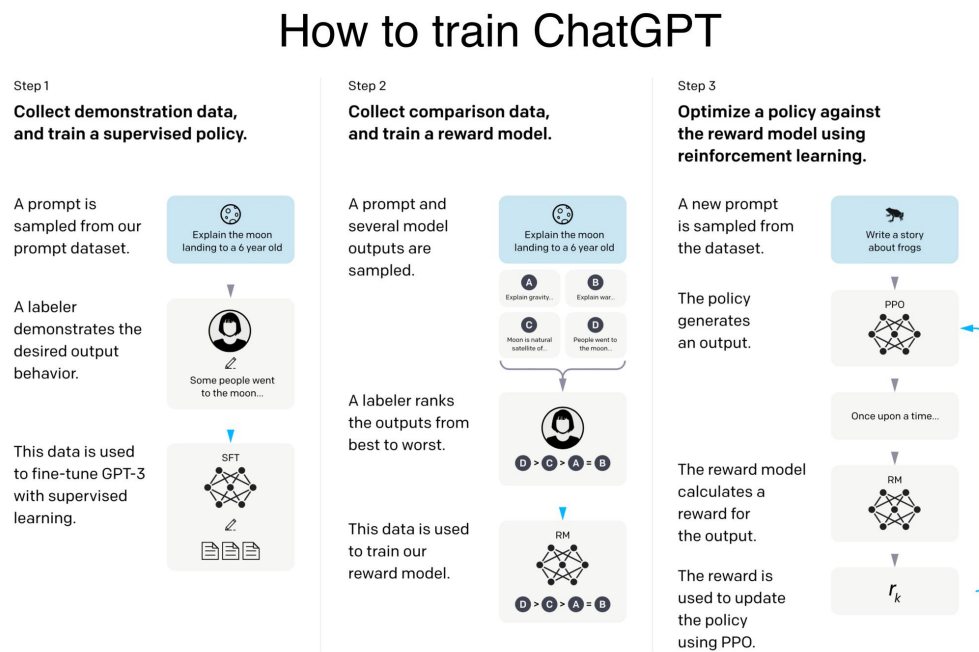


图 2.3: RLHF

RLHF（Reinforcement Learning from Human Feedback）参考自人类反馈训练范式 [15]。

2.7 flow mathching + GRPO

2.7.1 GRPO的缺陷

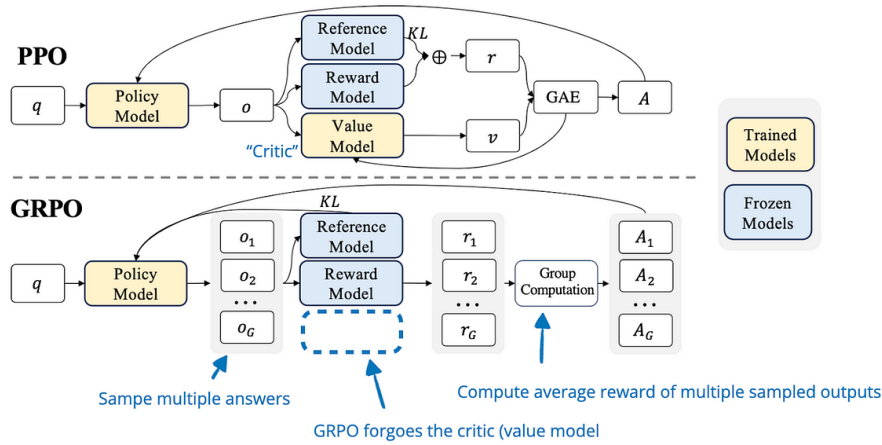


图 2.4: grpo

随着模型推理深度的增加和生成序列的急剧延长，GRPO以及整个RLVR范式暴露出一个致命的理论与工程瓶颈：信用分配（Credit Assignment）的极度稀疏性。在典型的RLVR设定中，模型在生成一条长达数千个Token的推理轨迹后，环境仅仅在序列末尾返回一个标量结果（如1代表通过，0代表失败或者编译报错返回负分）。这种序列级别的标量反馈彻底掩盖了底层环境状态的丰富细节。当模型在一个长序列的代码中仅仅因为拼错了一个变量名而导致整体失败时，标量反馈无法告知模型具体的错误位置。更为严重的是，在极其困难的推理任务中，探索初期的采样组往往会全军覆没，导致所有输出的奖励均相等。此时，GRPO计算出的相对优势会直接坍缩为零，导致梯度消失和学习过程的完全停滞。

2.7.2 SDPO

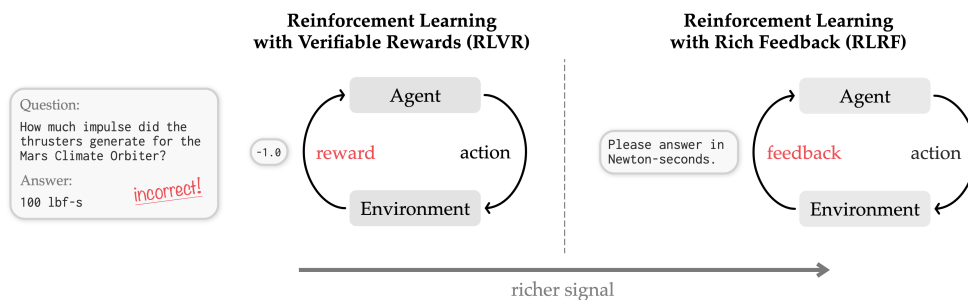


图 2.5: SDPO

Algorithm 1 SDPO

Input: Language model π_θ ; dataset with questions x ; number of rollouts G per question; environment to obtain feedback for attempts.

- 1: **repeat**
- 2: Sample question x from dataset.
- 3: Sample responses: $\{y_i\}_{i=1}^G \sim \pi_\theta(\cdot | x)$.
- 4: Evaluate responses to obtain feedback f_i .
- ▷ **Self-distillation:**
- 5: Compute log-probs of self-teacher
 $\log \pi_\theta(y_{i,t} | x, f_i, y_{i,<t})$.
- 6: Update θ with gradient descent on $\mathcal{L}_{\text{SDPO}}(\theta)$.
- 7: **until** converged

图 2.6: pseudocode of SDPO

基于自蒸馏的策略优化（Self-Distillation Policy Optimization, SDPO）。该研究不仅在算法层面提出了创新的损失函数设计，更在概念层面上将LLM的强化学习从RLVR推向了**基于丰富反馈的强化学习（Reinforcement Learning with Rich Feedback, RLRF）**的新纪元。SDPO算法的核心哲学在于：不依赖任何外部的强大教师模型或显式的奖励模型，而是让当前策略模型在接收到环境的富文本反馈后，化身为自身的“导师”，通过上下文回溯找出错误，并将这种修正后的认知以密集信号（Dense Signal）的形式蒸馏回自身的参数中。这一机制完美契合了人类学习的认知过程——不仅要知其然（成功或失败），更要知其所以然（为什么失败以及如何改进），从而实现了样本效率和最终性能的指数级跃升。

2.7.3 Diffusion + SDPO

扩散模型中融合自蒸馏与策略优化（SDPO）理念已经催生了多个前沿的数学模型和研究框架：

1. 奖励加权教师与前向 KL 蒸馏 (Reward-weighted Teacher & Forward-KL) [28]

数学形式：为了在扩散模型中实现自蒸馏并避免模式崩溃，模型会在离线状态下采用前向 KL 散度（Forward-KL）作为损失函数：

$$J(\theta) = \sum_{t=1}^T \mathbb{E}_{x_t \sim u_t} [\text{KL}(p_{t-1}^*(\cdot | x_t) \| \pi_\theta(\cdot | x_t))]$$

其中， p_{t-1}^* 是带有“特权信息”（即生成后计算出的高奖励）的“自我老师”分布，它由基础分布加上奖励加权构成：

$$p_{t-1}^*(x_{t-1} | x_t) \propto p_{\text{pre}}^{t-1}(x_{t-1} | x_t) e^{v_{t-1}(x_{t-1})/\tau}$$

（在此公式中， v_{t-1} 为时间步的价值/奖励估计， τ 为调节平滑度的温度系数）。

2. 逐步扩散策略优化 (Stepwise Diffusion Policy Optimization, 简称也是 SDPO) [25]

数学形式：该方法打破了只在生成末尾给奖励的局限，专门为少步数（Few-step）扩散模型设计了包含双状态轨迹采样（跟踪噪声状态与预测干净状态）的密集奖励策略更新目标。其核心策略梯度优化目标可以表示为：

$$\nabla_{\theta} \mathcal{J}(\theta) = \mathbb{E}_{\{a_t, s_t\} \sim f_{\theta}} - \lambda D_{\text{KL}}(f_{\theta} \| f_{\text{SFT}})$$

其中 $A_{\phi}(s_t)$ 代表分配到每个去噪时间步 t 上的密集优势估计（Stepwise advantage estimates）。

3. 受保护的偏好自蒸馏 (Diffusion-SDPO: Safeguarded DPO) [6]

数学形式与机制：在基于偏好的扩散模型自蒸馏（Diffusion-DPO）中，盲目扩大高分样本（胜者）和低分样本（败者）的概率差会导致胜者的重建误差也变大，从而降低图像质量。Diffusion-SDPO 从数学上改进了更新规则：通过一阶分析推导出了一个严格的闭式缩放系数（Closed-form scaling coefficient）。该系数根据“败者梯度”与“胜者梯度”的几何对齐程度，自适应地缩放针对败者的惩罚梯度，从而在数学上严格保证在每一步优化中，胜者（Preferred output）的重建误差呈单调非递增状态（Non-increasing）。

2.8 Toy idea: Diffusion + 连续优化（ADMM/DRSM）

1. 核心痛点对齐纯连续优化（ADMM）：擅长处理复杂的物理硬约束，但极度依赖初值，大范围场景易陷局部最优。纯扩散模型（Diffusion）：全局先验好、路径平滑，但概率生成机制无法严格保证无碰撞、动力学等“硬约束”（易穿障）。直接做后处理又会破坏扩散的全局连贯性。

2. 核心思路：“训练推理解耦，扩散管全局，ADMM 管局部约束”训练阶段（纯扩散）：零优化介入，完全保留标准扩散训练，模型专注学习优质轨迹的全局分布先验。推理阶段（去噪+投影交替）：在反向去噪环节，将模型预测的粗路径 $\hat{\tau}_{t-1}$ 作为锚点，构建带有硬约束的局部微调优化（本质是一个近端算子投影）：

$$\min_{\tau} \frac{1}{2} \|\tau - \hat{\tau}_{t-1}\|_2^2 + \lambda \cdot \mathcal{L}_{\text{constraint}}(\tau)$$

利用 ADMM/PDHG 进行轻量求解，将粗路径“拉回”到合适的可行域内，再送入下一步去噪。

3. 预期价值该方案在不增加训练复杂度、不破坏原生扩散链路的前提下，实现了“扩散降维提供全局最优初值 + ADMM 提供绝对安全可行性保障”的优势互补。如果验证可行，这种“生成先验 + 分裂投影”的范式不仅能解决复杂环境轨迹规划，还能自然迁移到需要严格满足几何约束的 3D 点云补全等生成任务中？

2.9 Optimal transport in flow matching

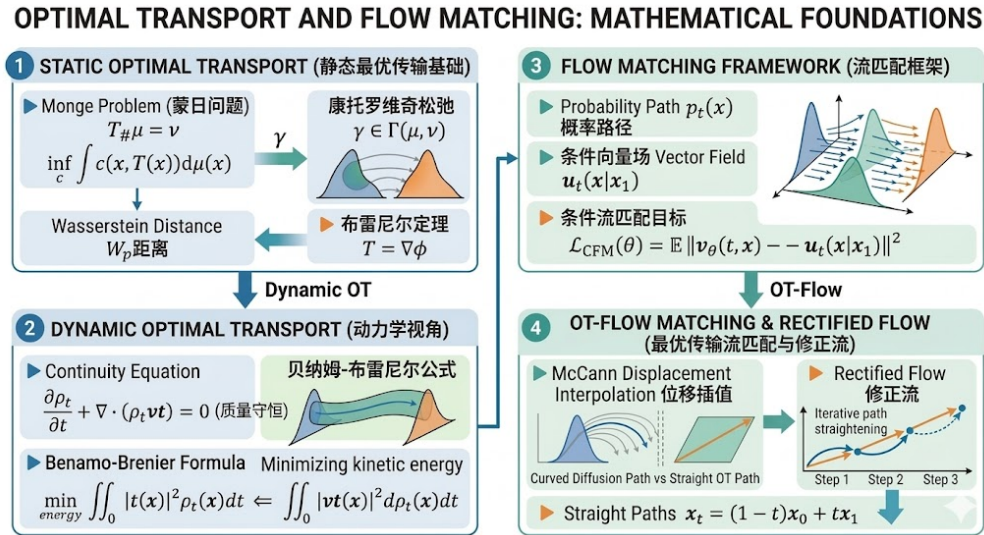


图 2.7: OT in flow matching

引言

连续标准化流、概率流常微分方程与最优传输理论在近几年备受关注，生成式人工智能（Generative AI）领域经历了从生成对抗网络（GANs）到去噪扩散概率模型（DDPMs）[9]的范式变迁。然而，尽管基于 SDE 的扩散模型[21]在图像、音频及科学数据生成上取得了瞩目成就，其底层的马尔可夫链或 SDE 数值积分机制却导致了极高的采样延迟与计算成本。为了克服这一固有缺陷，出现了基于常微分方程（ODEs）的确定性演化模型，即连续标准化流（Continuous Normalizing Flows, CNFs）与概率流常微分方程（Probability Flow ODEs）[21]。连续标准化流通过积分一个时变向量场来实现从简单先验分布（如标准高斯分布）到复杂高维数据分布的确定性映射。尽管这种确定性映射能够实现精确的似然估计并支持更高效的常微分方程求解器，但传统 CNFs 依赖于昂贵的最大似然估计进行训练，在训练期间需要反复进行前向与反向的 ODE 模拟，导致其在超高维数据（如百万像素图像）上几乎无法扩展。正是在这一背景下，流匹配（Flow Matching, FM）作为一种全新的“无模拟（Simulation-free）”训练范式应运而生[12]。它通过直接回归条件概率路径（Conditional Probability Paths）所诱导的向量场，彻底解耦了训练过程与 ODE 求解过程。更为关键的是，流匹配框架揭示了一个生成模型的核心自由度：我们不仅可以学习向量场，还可以自由设计数据分布在时间上的演化路径。如何设计一条“最优”的演化路径？这正是最优传输（Optimal Transport, OT）理论[23]发挥决定性作用的领域。通过引入最优传输中的 McCann 位移插值（Displacement Interpolation）[14]，流匹配与同期提出的修正流（Rectified Flow）[13]能够构造出在概率测度空间中呈恒定速度的“直线（Straight

Paths) 演化轨迹。直线轨迹彻底最小化了传输动能，不仅从根本上消除了 ODE 数值积分的截断误差，还使得模型在单步或极少步数 (Few-step) 内生成高保真数据成为可能。

2.9.1 静态最优传输的数学底座

最优传输理论[23]的起源可以追溯到 1781 年法国数学家 Gaspard Monge 提出的土木工程问题：如何以最小的代价将一堆泥土（分布 μ ）移动到一个指定形状的坑（分布 ν ）中。进入 20 世纪 40 年代，苏联数学家 Leonid Kantorovich 对其进行了线性松弛。

Monge 问题与前推测度约束

最优传输问题建立在概率测度空间之上。设 (X, d_X) 和 (Y, d_Y) 为两个可分完备的度量空间，在深度学习中，通常假设 $X = Y = \mathbb{R}^d$ 。令 $\mathcal{P}(X)$ 表示 X 上的全体概率测度构成的空间。设 $\mu \in \mathcal{P}(X)$ 为源测度（代表生成模型中的先验噪声）， $\nu \in \mathcal{P}(Y)$ 为目标测度（代表真实数据分布）。我们需要寻找一个传输映射（Transport Map） $T: X \rightarrow Y$ 。

为了保证在传输过程中质量守恒，映射 T 必须满足前推测度（Push-forward measure）的严格约束，记作 $T_{\#}\mu = \nu$ ，即：对于目标空间 Y 中的任意波莱尔可测集（Borel set） $B \subset Y$ ，均有：

$$\nu(B) = \mu(T^{-1}(B))$$

该等式意味着，所有被映射 T 投射到区域 B 中的源空间点的质量总和，必须精确等于目标分布在区域 B 中的质量。进一步地，如果在 \mathbb{R}^d 空间中，测度 μ 和 ν 均具有关于勒贝格测度绝对连续的概率密度函数 $p(x)$ 和 $q(y)$ ，且假设传输映射 T 是一个光滑的微分同胚（Diffeomorphism），那么根据多元微积分的变量代换公式，前推测度约束可以被写为如下形式：

$$p(x) = q(T(x)) |\det(\nabla T(x))|$$

定义代价函数 $c: X \times Y \rightarrow [0, +\infty]$ ，它量化了将单位质量从坐标 x 移动到坐标 y 的“成本”。在给定代价函数后，Monge 问题旨在寻找一个满足前推测度约束的传输映射 T ，使得总传输代价最小化：

$$\inf_{T: T_{\#}\mu = \nu} \int_X c(x, T(x)) d\mu(x)$$

Monge 问题的拓扑障碍与非存在性

从优化理论的视角来看，Monge 问题存在极其严重的解析困难。

- 约束条件 $T_{\#}\mu = \nu$ 高度非线性且非凸。
- 在许多病态或离散的测度配置下，Monge 问题的可行域可能是空集

考虑一个极端的反例：若 μ 是一个聚集在单点的狄拉克测度（Dirac delta），而 ν 是两个分离的、权重各为 0.5 的点测度的组合。由于函数 T 是单值的，它无法将一个输入点映射到两个不同的输出点（即数学上禁止质量分裂，Mass splitting），因此此时不存在任何满足前推约束的映射 T 。

Kantorovich 松弛与无穷维线性规划

为了克服 Monge 映射不能分裂质量的缺陷，Leonid Kantorovich 于 1942 年提出，不再寻找一个确定性的、从源到目标的单值映射 $x \mapsto T(x)$ ，而是寻找一个在乘积空间 $X \times Y$ 上的传输计划或联合耦合（Joint Coupling）。定义 $\Pi(\mu, \nu)$ 为乘积空间 $X \times Y$ 上所有以 μ 和 ν 为边缘分布（Marginals）的联合概率测度的集合。用积分的形式严格表达，即对于任意 $\gamma \in \Pi(\mu, \nu)$ ，对于所有连续有界函数 $f \in C_b(X)$ 和 $g \in C_b(Y)$ ，下述边缘约束均成立：

$$\int_{X \times Y} f(x) d\gamma(x, y) = \int_X f(x) d\mu(x)$$

$$\int_{X \times Y} g(y) d\gamma(x, y) = \int_Y g(y) d\nu(y)$$

在这个框架下， $d\gamma(x, y)$ 的物理意义是：从源分布的 x 位置抽取多少质量，并将其传输到目标分布的 y 位置。这种松弛机制天然地允许了一个位置的质量被分散到多个目标位置。Kantorovich 问题因此被定义为在容许耦合集合中最小化期望代价：

$$\min_{\gamma \in \Pi(\mu, \nu)} \int_{X \times Y} c(x, y) d\gamma(x, y)$$

与 Monge 问题不同，集合 $\Pi(\mu, \nu)$ 总是非空的（因为独立联合分布 $\mu \otimes \nu$ 永远是其中的一个元素）。

更重要的是，在弱拓扑下， $\Pi(\mu, \nu)$ 是一个紧致凸集（Compact convex set），且目标泛函 $\gamma \mapsto \int c(x, y) d\gamma$ 是关于测度 γ 的线性泛函。这使得 Kantorovich 问题在本质上成为了一个无穷维的线性规划问题。根据 Prokhorov 定理与下半连续性，可以严格证明在代价函数 $c(x, y)$ 连续且下有界的条件下，Kantorovich 问题的全局最优解必定存在。

Wasserstein 度量与机器学习

定义 2.9.1 (p -Wasserstein 距离). 基于 Kantorovich 问题，当底层代价函数为度量空间距离的 p 次方 ($c(x, y) = d(x, y)^p$, $p \geq 1$) 时，Wasserstein 距离定义为

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu, \nu)} \int_{X \times Y} d(x, y)^p d\gamma(x, y) \right)^{1/p}$$

其中 $\Pi(\mu, \nu)$ 表示所有以 μ 和 ν 为边缘分布的联合概率测度的集合， $d(x, y)$ 是底层空间的度量。

对于流匹配和概率流常微分方程的研究，最具核心意义的是 W_2 距离（基于二次代价）。

定义 2.9.2 (p -Wasserstein 空间). 设 (X, d) 为度量空间， $p \geq 1$ 为实数。定义 p -阶 Wasserstein 空间为：

$$\mathcal{P}_p(X) = \left\{ \mu \in \mathcal{P}(X) \mid \int_X d(x, x_0)^p d\mu(x) < \infty, \text{ 对某个 } x_0 \in X \right\}$$

在 $\mathcal{P}_p(X)$ 上, p -Wasserstein 距离定义为:

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu, \nu)} \int_{X \times X} d(x, y)^p d\gamma(x, y) \right)^{1/p}$$

其中 $\Pi(\mu, \nu)$ 表示所有以 μ 与 ν 为边缘的联合概率测度的集合。于是 $(\mathcal{P}_p(X), W_p)$ 构成度量空间。

备注 2.9.1. x_0 是参考点, 不妨取原点。如果一个分布的尾部太厚 (比如某些重尾分布), 把它的质量搬运到另一个分布可能需要无穷大的代价。限制 p 阶矩有限, 是为了确保在这个空间里, 任意两个分布 $\mu, \nu \in \mathcal{P}_p(X)$ 之间的 p -Wasserstein 距离 $W_p(\mu, \nu)$ 都是有限的实数。

离散情况 (Discrete Cases)——矩阵视角

实际计算中常见分布往往是离散的 (例如采样点分布)。此时, Wasserstein 距离最优传输问题可通过矩阵方法求解。

假设源分布由离散位置 x_1, \dots, x_m 及权重向量 $a \in \mathbb{R}^m$ ($\sum a_i = 1$) 给出, 目标分布由 y_1, \dots, y_n 及权重 $b \in \mathbb{R}^n$ ($\sum b_j = 1$) 给出。代价矩阵 $C \in \mathbb{R}^{m \times n}$: $C_{ij} = d(x_i, y_j)^p$, 表示将单元质量从 x_i 运至 y_j 的代价。传输矩阵 $P \in \mathbb{R}^{m \times n}$: P_{ij} 表示从 x_i 运送到 y_j 的质量。离散耦合集定义:

$$U(a, b) = \left\{ P \in \mathbb{R}^{m \times n} \mid P_{ij} \geq 0; \sum_{j=1}^n P_{ij} = a_i \forall i; \sum_{i=1}^m P_{ij} = b_j \forall j \right\}$$

常记作 $P \mathbb{K}_n = a$, $P^\top \mathbb{K}_m = b$ (分别代表行和、列的约束), 这样共有 $m + n$ 个一阶线性约束。

最优离散传输总代价为

$$L_C(a, b) = \min_{P \in U(a, b)} \langle C, P \rangle = \min_{P \in U(a, b)} \sum_{i=1}^m \sum_{j=1}^n C_{ij} P_{ij}$$

基于上述定义, 离散 Wasserstein 距离为

$$W_p(a, b) = (L_C(a, b))^{1/p} = \left(\min_{P \in U(a, b)} \sum_{i=1}^m \sum_{j=1}^n d(x_i, y_j)^p P_{ij} \right)^{1/p}$$

Wasserstein 重心 (Wasserstein Barycenter)

最优传输的一个重要应用是“分布重心”的计算, 即寻找一组分布的“几何平均”分布, 使得它到所有给定分布的加权 Wasserstein 距离之和最小。

给定分布集合 $\{p_1, p_2, \dots, p_K\}$ 以及权重 $\{\lambda_1, \lambda_2, \dots, \lambda_K\}$ ($\lambda_k \geq 0, \sum_k \lambda_k = 1$), 其 Wasserstein 重心 q^* 定义为

$$q^* = \arg \min_{q \in Q} \sum_{k=1}^K \lambda_k W(q, p_k)$$

这里 Q 为候选分布的空间, W 通常取 W_2 距离。

备注 2.9.2. *Wasserstein* 重心可看作经典均值在概率分布空间中的几何推广。它可用于图像融合、插值、聚类任务，在图形学、机器学习、生成模型等领域均有广泛应用，能够生成一组分布之间平滑、中庸的“中间态”分布。

比较维度	Monge 优化问题	Kantorovich 松弛问题
优化变量	确定性单值映射函数 $T: X \rightarrow Y$	联合概率测度（耦合） $\gamma \in \Pi(\mu, \nu)$
质量分裂	绝对禁止	允许
约束条件	前推测度约束 $T_{\#}\mu = \nu$	边缘分布约束（线性算子，凸集合）
解的存在性	在离散测度或维度不匹配时可能无解	在任意概率测度下均存在最优解
生成模型对应模型	在推理时的确定性 ODE 路径	训练 Flow Matching 时的联合采样分布

Kantorovich 对偶理论

Kantorovich 对偶问题定义为：寻找函数 $\phi \in L^1(\mu)$ 和 $\psi \in L^1(\nu)$ ，在满足约束 $\phi(x) + \psi(y) \leq c(x, y)$ 对于所有 $(x, y) \in X \times Y$ 成立的前提下，最大化如下对偶泛函（推导见附录.0.2）：

$$\sup_{\phi, \psi} \left(\int_X \phi(x) d\mu(x) + \int_Y \psi(y) d\nu(y) \right)$$

其中， ϕ, ψ 也被称为势能函数。在对偶问题中，可以用“包工头最大利润问题”来理解其经济学意义：包工头（对应于势函数 ϕ, ψ 的选择）想让自己的总代办费 $\int \phi(x) d\mu + \int \psi(y) d\nu$ 尽可能大。但他受到一个铁律约束：无论选哪条运货路线 (x, y) ，他收的总费 $\phi(x) + \psi(y)$ 不能高于客户自己运输花的钱 $c(x, y)$ 。现在假定包工头已经定好了每个起点 x 的收费标准 $\phi(x)$ ，他接下来就会把每个终点 y 的收费标准 $\psi(y)$ 尽量抬到最大。但由于约束始终要对每个 x 成立，他最多只能收

$$\psi(y) \leq c(x, y) - \phi(x) \quad \forall x$$

因此， $\psi(y)$ 能取的最大值就是所有 $c(x, y) - \phi(x)$ 的最小值，也就是

$$\psi(y) = \inf_{x \in X} \{c(x, y) - \phi(x)\}$$

这个 ψ 叫做 ϕ 的 c -变换，记作 $\phi^c(y)$ 。于是，理论上只需琢磨如何取好 ϕ ， ψ 自然随之确定，最终对偶问题转化为在所有 c -凹函数（即形如 ϕ^c 的函数）中寻找最好的一种 ϕ 。

进一步，若代价函数采用二次距离 $c(x, y) = \frac{1}{2}\|x - y\|^2$ ，我们可以将 c 展开为

$$c(x, y) = \frac{1}{2}\|x\|^2 - x \cdot y + \frac{1}{2}\|y\|^2$$

代入 c -变换公式，便有

$$\phi^c(y) = \inf_x \left\{ \frac{1}{2} \|x\|^2 - x \cdot y + \frac{1}{2} \|y\|^2 - \phi(x) \right\}$$

将与 x 无关的项 $\frac{1}{2} \|y\|^2$ 移到 \inf 外，

$$\phi^c(y) = \frac{1}{2} \|y\|^2 - \sup_x \left(x \cdot y - \left[\frac{1}{2} \|x\|^2 - \phi(x) \right] \right)$$

令 $\hat{\phi}(\cdot) = \frac{1}{2} \|\cdot\|^2 - \phi(\cdot)$ ，容易验证 $\hat{\phi}(\cdot)$ 是凸的（共轭函数），也被称为 **Brenier 势能函数**。

2.9.2 Brenier 定理与最优传输映射的还原

在实际的深度生成模型（例如流匹配或连续标准化流）中，我们最终需要构建的是一个确定性的函数。也就是说，我们必须从松弛的 Kantorovich 耦合 γ 中，找回确定性的 Monge 映射 T 。法国数学家 Yann Brenier 将最优传输映射的存在性、唯一性与凸势函数（Convex Potentials）的梯度联系在了一起[3, 23]。

Knott-Smith 最优性准则与次微分图

由上面的讨论，当考虑二次代价 $c(x, y) = \frac{1}{2} \|x - y\|^2$ 时，对偶问题等同于寻找一个凸函数 $\hat{\phi}$ 及其凸共轭 $\hat{\phi}^*(y) = \sup_x \{ \langle x, y \rangle - \hat{\phi}(x) \}$ ，满足 Fenchel-Young 不等式：

$$\hat{\phi}(x) + \hat{\phi}^*(y) \geq \langle x, y \rangle$$

Knott-Smith 最优性准则由此得出：一个容许耦合 $\pi^\dagger \in \Pi(\mu, \nu)$ 是 Kantorovich 最优传输计划的充要条件是，存在一个下半连续的凸函数 $\phi: \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ ，使得 π^\dagger 的支撑集严格包含在 ϕ 的次微分图中，即对于 π^\dagger -几乎处处的点对 (x, y) ，必定满足边界情况：

$$\phi(x) + \phi^*(y) = \langle x, y \rangle \iff y \in \partial\phi(x)$$

这一准则将寻找最优耦合转化为寻找一个其导数图能够容纳概率测度质量的凸函数。

基于上述凸分析，Brenier 定理给出了 Monge 映射存在且唯一的核心定理。

定理 2.9.1 (Brenier 定理). 设 $X = Y = \mathbb{R}^d$ ，源测度 μ 和目标测度 ν 均具有有限的二阶矩。进一步假设 μ 绝对连续于勒贝格测度（或更一般的正测度）。则对于二次代价 $c(x, y) = \frac{1}{2} \|x - y\|^2$ ，有：

- **唯一性与存在性：** Kantorovich 松弛问题存在唯一的全局最优耦合 γ^\dagger ；
- **确定性映射诱导：** 由于 μ 的绝对连续性，存在下半连续凸函数 ϕ ，使得 ϕ 在 μ -几乎处处可微，并且最优耦合 γ^\dagger 集中于图像 $(x, \nabla\phi(x))$ 上。即，有唯一的确定性映射 $T^\dagger(x) = \nabla\phi(x)$ 使

$$d\gamma^\dagger(x, y) = d\mu(x) \delta_{y=T^\dagger(x)}$$

该映射 T^\dagger 也是 Monge 问题的唯一最优解；

- **最优传输的梯度场表达**：该唯一最优传输映射 T^\dagger 可由下半连续凸函数 ϕ （即 Brenier 势函数）的梯度给出：

$$T^\dagger(x) = \nabla\phi(x), \quad \text{对 } \mu\text{-几乎处处的 } x \text{ 成立。}$$

备注 2.9.3. 在最优传输框架下，无论是多么复杂的两个高维分布，它们之间始终存在一个唯一的、由凸势场梯度驱动的最优映射，即

$$v(x_t) = \nabla \left(\phi(x_0) - \frac{1}{2} \|x_0\|^2 \right)$$

网络在学习过程中，权重结构能够限制或逼近某个凸函数的梯度（正如凸势流 *Convex Potential Flows* 所尝试的那样），那么该生成模型就在渐近地逼近真正的最优传输映射，也就能学到“不交叉、耗能最小”的完美直线路径。

备注 2.9.4. 从偏微分方程的视角来看，若 μ 和 ν 拥有光滑的密度 p 和 q ，将 $T = \nabla\phi$ 代入变量代换公式，我们会得到著名的完全非线性偏微分方程——*Monge-Ampère* 方程：

$$p(x) = q(\nabla\phi(x)) \det(D^2\phi(x))$$

其中 $D^2\phi(x)$ 是凸势函数的海森矩阵，其正定性保证了传输映射的单调性与非交叉性。

2.9.3 从几何测地线到流体力学视角

Monge 映射 $T(x)$ 和 Kantorovich 耦合 $\gamma(x, y)$ 均是静态的描述。然而，连续标准化流（CNF）与概率流常微分方程本质上是动态的系统。要将静态的最优传输引入生成模型，需要经历一次从“纯几何”到“流体力学”的视角落阶。

1. McCann 位移插值：纯几何视角的拉格朗日直线

在得知了每个粒子 $x_0 \sim \mu_0$ 的最优目标地点是 Brenier 映射 $T(x_0) = \nabla\phi(x_0)$ 后，中间的过渡形态应该是什么样？1997 年，加拿大数学家 Robert McCann 从纯欧氏几何的直觉出发，提出了位移插值（Displacement Interpolation）[14]。既然底层空间两点间直线最短，他便让每一个概率粒子以恒定的速度沿直线行进至 $T(x_0)$ ，这定义了一个随时间演化的插值推前测度：

$$\mu_t = ((1-t)\text{Id} + tT)_\# \mu_0$$

备注 2.9.5. 相比于简单且物理上不合理的密度线性插值 $\mu_t = (1-t)\mu_0 + t\mu_1$ （这会导致质量在出发点原地隐身，并在终点凭空出现），McCann 插值是质量沿空间位移的真实流动。这在宏观上表现为整个概率云的平滑“推移与扭曲”变形。命题 .0.1 证明了这条轨迹构成了 Wasserstein 空间 $(P_2(\mathbb{R}^d), W_2)$ 中的匀速测地线（*Constant-speed Geodesic*）。

McCann 的公式提供的是拉格朗日视角的向量场：粒子的速度 $v = T(x_0) - x_0$ 强依赖于它最初始的纯噪声状态 x_0 。然而，在生成模型的推理阶段，神经网络（如 U-Net 或

DiT) 处于“失忆”状态。当网络面对一个中间时刻 t 的模糊图像 x_t 时, 它根本不知道这个图像最初是由哪个噪声 x_0 演化而来的。神经网络和 ODE 求解器真正需要的, 是欧拉视角的全局向量场 $v_t(x)$ 。

在 McCann 匀速直线下

$$\|T(x) - x\|^2 = \int_0^1 \|v(x)\|^2 dt, \quad W_2^2 = \int_0^1 \int_{\mathbb{R}^d} \|v(x)\|^2 p_0(x) dx dt.$$

对固定 t 令 $y = X_t(x)$, 根据推前测度 (质量守恒) $p_0(x) dx = p_t(y) dy$, 即可把内层积分改写为

$$\int_{\mathbb{R}^d} \|v_t(y)\|^2 p_t(y) dy$$

(变量名再记作 x)。于是“找最优映射 T ”被改写为“找最优时空场 (p_t, v_t) ”的动能极小化。但动能泛函本身不足以刻画真实输运, 必须附加质量守恒约束。欧拉坐标下该约束正是连续性方程。

定理 2.9.2 (连续性方程 Continuity Equation).

$$\frac{\partial p_t}{\partial t} + \nabla \cdot (v_t p_t) = 0$$

其中边界条件为 p_0 与 p_1 。

证明. 设拉格朗日流 $\frac{dX_t}{dt} = v_t(X_t)$ 、 $J_t = \det \nabla X_t$, 由换元 $p_0(x) = p_t(X_t(x))J_t(x)$ 对 t 求导, 结合链式法则与 Liouville 公式 $\partial_t J_t = J_t \nabla \cdot v_t$, 约去 $J_t > 0$ 后得到

$$\partial_t p_t + \nabla \cdot (p_t v_t) = 0.$$

定理得证。□

备注 2.9.6. 方程意味着在空间任一点 x , 密度的随时间变化率严格等于概率通量 $(p_t v_t)$ 流出该点的散度的负值。

于是, 得到动态的 **Benamou-Brenier 公式**[2]:

$$W_2^2(p_0, p_1) = \inf_{(p_t, v_t)} \left\{ \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 p_t(x) dx dt \right\} \quad (2.22)$$

其中 \inf 取遍满足 $p_t|_{t=0} = p_0$ 、 $p_t|_{t=1} = p_1$ 及 $\partial_t p_t + \nabla \cdot (p_t v_t) = 0$ 的 (p_t, v_t) 。

这一公式将最优传输从一个几何映射问题, 转变为了寻找**最优控制速度场**的泛函变分问题。根据变分法与微积分基本引理, 动能的极小化要求粒子的加速度为零 (无受力状态), 这在物理上同样印证了 McCann 的匀速直线运动, 并且将粒子的轨迹转化为了满足连续性方程的欧拉速度场 $v_t(x)$ 。

2.9.4 Flow Matching框架

基于静态 Monge 映射、连续性方程与 McCann 位移插值，可以得到 Flow Matching 框架的数学推导[12]。

CNF 瓶颈与边缘向量场回归

流匹配的直接优化对象是连续标准化流（CNFs）中的时变向量场 $v_\theta(t, x)$ 。设 $u_t(x)$ 为生成该边缘概率路径 $p_t(x)$ 的真实边缘向量场（Marginal Vector Field），则理想 Flow Matching (FM) 损失函数为：

$$\mathcal{L}_{\text{FM}}(\theta) = \int_0^1 \mathbb{E}_{x \sim p_t(x)} [\|v_\theta(t, x) - u_t(x)\|^2] dt$$

然而对于极其复杂的高维自然图像或分子结构分布，真实的演化路径 $p_t(x)$ 及其驱动向量场 $u_t(x)$ 是在所有目标数据点上积分形成的边缘混合分布，其数学解析式是绝对不可解的，无法直接用于神经网络的反向传播。

因此，对于每一个特定的真实数据点 x_1 ，我们可以人为设计一条极其简单的条件概率路径 $p_t(x|x_1)$ ，也就是：在知道粒子最终要抵达终点 x_1 的前提下，粒子在 t 时刻应该处于的分布。由命题 2.2 边缘向量场 $u_t(x)$ 实际上就是条件向量场的后验期望，即：

$$u_t(x) = \int u_t(x|x_1) \frac{p_t(x|x_1) q_{\text{data}}(x_1)}{p_t(x)} dx_1 = \mathbb{E}_{x_1 \sim p_{1|t}(\cdot|x_t)} [u_t(x|x_1)]$$

定理 2.9.3 (Lipman [12]). 流匹配损失与条件流匹配损失的梯度是一致的：

$$\nabla_\theta \mathcal{L}_{\text{FM}}(\theta) = \nabla_\theta \mathcal{L}_{\text{CFM}}(\theta).$$

特别地，流匹配损失的极小值点即为边缘速度场 $u_t(x)$ 。

证明. 记

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, Z \sim q, X_t \sim p_{t|Z}(\cdot|Z)} [D(u_t(X_t|Z), u_t^\theta(X_t))].$$

则有

$$\begin{aligned} \nabla_\theta \mathcal{L}_{\text{FM}}(\theta) &= \nabla_\theta \mathbb{E}_{t, X_t \sim p_t} D(u_t(X_t), u_t^\theta(X_t)) \\ &= \mathbb{E}_{t, X_t \sim p_t} \nabla_\theta D(u_t(X_t), u_t^\theta(X_t)) \\ &\stackrel{(i)}{=} \mathbb{E}_{t, X_t \sim p_t} \nabla_v D(u_t(X_t), u_t^\theta(X_t)) \nabla_\theta u_t^\theta(X_t) \\ &= \mathbb{E}_{t, X_t \sim p_t} \nabla_v D\left(\mathbb{E}_{Z \sim p_{Z|t}(\cdot|X_t)} [u_t(X_t|Z)], u_t^\theta(X_t)\right) \nabla_\theta u_t^\theta(X_t) \\ &\stackrel{(ii)}{=} \mathbb{E}_{t, X_t \sim p_t} \mathbb{E}_{Z \sim p_{Z|t}(\cdot|X_t)} [\nabla_v D(u_t(X_t|Z), u_t^\theta(X_t)) \nabla_\theta u_t^\theta(X_t)] \\ &\stackrel{(iii)}{=} \mathbb{E}_{t, X_t \sim p_t} \mathbb{E}_{Z \sim p_{Z|t}(\cdot|X_t)} [\nabla_\theta D(u_t(X_t|Z), u_t^\theta(X_t))] \\ &\stackrel{(iv)}{=} \nabla_\theta \mathbb{E}_{t, Z \sim q, X_t \sim p_{t|Z}(\cdot|Z)} [D(u_t(X_t|Z), u_t^\theta(X_t))] \\ &= \nabla_\theta \mathcal{L}_{\text{CFM}}(\theta). \end{aligned}$$

其中，(i)、(iii)：应用链式法则；(ii)：对给定 X_t 使用 $u_t(X_t) = \mathbb{E}_{Z|X_t}[u_t(X_t|Z)]$ ；(iv)：利用贝叶斯定理（或塔式法则，将条件期望转回联合期望）。结论得证。 \square

由此，难以评估的 \mathcal{L}_{FM} 被等价替换为完全无模拟、可扩展的条件流匹配（CFM）损失：

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}(0,1), x_1 \sim q_{\text{data}}, x \sim p_t(\cdot|x_1)} [\|v_\theta(t, x) - u_t(x|x_1)\|^2]$$

于是，我们只需要随机从数据集中抽取一个样本 x_1 ，再依据我们自己设计的简单条件路径采出一个加噪的 x_t ，然后让网络去拟合那个具有解析形式的简单向量场 $u_t(x|x_1)$ 即可。这彻底打破了传统 CNF 需要通过 ODE Solver 解常微分方程来获得梯度的瓶颈。

高斯条件概率路径的严格求导

CFM 框架留下的最后一个问题是：该如何设计 $p_t(x|x_1)$ 及其向量场 $u_t(x|x_1)$ ？在绝大多数主流模型（包含 Stable Diffusion 3 等最新架构[4]）中，条件路径被建模为各向同性的高斯概率路径：

$$p_t(x|x_1) = \mathcal{N}(x; \mu_t(x_1), \sigma_t(x_1)^2 I)$$

其中 $\mu_t(x_1)$ 控制均值从 0 向 x_1 演化， $\sigma_t(x_1)$ 控制方差从 ≈ 1 向 ≈ 0 退火。我们可以用重参数化技巧采样出时刻 t 的状态：

$$x = \mu_t(x_1) + \sigma_t(x_1)\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

对 x 关于时间 t 求导：

$$\frac{dx}{dt} = \dot{\mu}_t(x_1) + \dot{\sigma}_t(x_1)\epsilon$$

由于 $\epsilon = \frac{x - \mu_t(x_1)}{\sigma_t(x_1)}$ ，将其代入上式，我们得到了高斯路径的解析条件向量场公式：

$$u_t(x|x_1) = \dot{\mu}_t(x_1) + \frac{\dot{\sigma}_t(x_1)}{\sigma_t(x_1)}(x - \mu_t(x_1))$$

这表明向量场 u_t 是关于空间坐标 x 的仿射线性函数，其可以极度轻量化地嵌入到回归目标中。

OT-Flow Matching: 当流匹配遇见 McCann 位移插值

流匹配框架最具魅力的部分，在于它能够直接具象化前文所述的最优传输 McCann 位移插值。基于高斯条件概率路径，Lipman 等提出了 OT-Flow Matching[12]。为了逼近 McCann 插值，我们在构造均值和方差序列时，采用最为简单暴力的线性插值：

$$\mu_t(x_1) = tx_1, \quad \sigma_t(x_1) = 1 - t + \sigma_{\text{min}} \approx 1 - t$$

当 $t = 0$ 时， $\mu_0 = 0, \sigma_0 = 1$ ；当 $t = 1$ 时， $\mu_1 = x_1, \sigma_1 \approx 0$ （完美塌缩到数据点 x_1 ）。此时重参数化后的采样点为： $x_t = tx_1 + (1 - t)x_0$ 。我们将其对时间求导数：

$$\dot{\mu}_t(x_1) = x_1, \quad \dot{\sigma}_t(x_1) = -1$$

将这两个导数代入前文推导的条件向量场公式中：

$$u_t(x_t|x_1) = x_1 + \frac{-1}{1-t}(x_t - tx_1)$$

注意到 $x_t - tx_1$ 恰好等于 $(1-t)x_0$ ，代入并消去分数项的分子分母：

$$u_t(x_t|x_1) = x_1 - \frac{1-t}{1-t}x_0 = x_1 - x_0$$

备注 2.9.7. 基于 OT 构造的流匹配中，条件向量场 $u_t(x_t|x_1) = x_1 - x_0$ 是一个与时间 t 完全无关、仅仅取决于端点空间差异的常数向量。这意味着网络在学习时具有极小的预测方差。

2.9.5 矫正流（Rectified Flow）与深度拉直机制

Rectified Flow[13] 的核心思路同样是用常微分方程 $dZ_t = v(Z_t, t)dt$ 连接经验观测分布 π_0 与 π_1 。其第一步（1-Rectification）建立在随机独立抽取的数据对 $(X_0, X_1) \sim \pi_0 \otimes \pi_1$ 基础之上，强迫模型去拟合线性插值路径 $X_t = tX_1 + (1-t)X_0$ ：

$$\min_v \int_0^1 \mathbb{E} \|(X_1 - X_0) - v(X_t, t)\|^2 dt$$

在随机抽取的情况下，不同粒子的线性轨迹 $(X_0 \rightarrow X_1)$ 必定会在多维空间中发生交叉。因此，在拟合 MSE 损失时，模型学到的宏观向量场 $v(X_t, t)$ 会在交叉点处取这些独立轨迹的“平均值”，宏观的 1-Rectified Flow 仍然呈现出一定的弯曲性。

为了彻底实现完美的最优传输，Rectified Flow 引入了递归重耦合机制，称为Reflow。使用训练好的 1-RF 模型，从 π_0 采样 X_0 ，通过求解 ODE 生成目标点 X'_1 。这一步建立了具有关联性的确定性耦合偶对 (X_0, X'_1) 。用这对新的、被模型自然“解缠（Un-tangled）”的数据对作为新的目标边界，重新训练第二个向量场模型（2-Rectification）。**凸运输成本非增定理**严格证明了，单次矫正流学到的确定性耦合方案，相对于其初始的任何耦合，都会在任意凸函数 c 作为代价函数的情境下，严格降低期望运输成本 $\mathbb{E}[c(X_1 - X_0)]$ 。

随着 K -Rectification 的递归进行，模型轨迹在测度空间内的“拉直程度”以 $\mathcal{O}(1/K)$ 的速率收敛于绝对零曲率[13]。此时，常微分方程的积分步长对结果精度毫无影响。在图像生成的实践中，这意味着在两次 Reflow 之后，生成网络在单步内的积分质量，就可比拟传统 SDE 扩散模型经过 1000 步采样的质量，同时将计算延迟压缩了两个数量级（参见 Rectified Flow 与大规模文生图系统中的实证[13, 4]）。

表 2.1: 少步生成相关方法对比（代表性文献：Flow Matching [12]，一致性模型 [20]，MeanFlow [8]，Shortcut [5]）

模型与方法	解决的核心痛点	核心创新点
一致性模型 (CM)	只要模型还在拟合速度场，就无法摆脱对数值求解器和离散时间步长的依赖，必然带来计算延迟与累积近似误差。	彻底抛弃速度场演化逻辑，要求神经网络直接学习全局解映射。强行约束同一生成轨迹上的任意阶段噪声点，都必须一次性预测出相同的终点干净数据。
MeanFlow	拟合瞬时切线速度进行大步长外推会产生巨大几何偏离；同时一致性模型极度依赖指数移动平均网络和繁琐的离散步长设计，训练容易崩溃。	引入微积分恒等式，将建模目标从微观的“瞬时速度”升维至宏观的“区间平均速度”，完美吸收了路径上的曲率信息，实现了无误差的单步直接跨越，且完全无需教师模型或蒸馏即可从零训练。
Shortcut	一致性模型强制映射到终点剥夺了推理的灵活性和纠错能力；大幅度单步生成会引发严重的分类器引导复合误差与低频频域偏差。	在网络中显式引入步长距离参数，让模型学习在轨迹上按需“抄捷径”。推理时可自由决定是一步到位还是多步分段，并通过多级小波损失和缩放最优传输进一步稳定生成轨迹。

延伸阅读：科学空间（苏剑林）<https://kexue.fm/archives/10958>）

- 一致性模型 (CM)[20]:

理论基础建立在严格的边界条件 $f_\theta(x_0, 0) = x_0$ 以及一致性约束 $f_\theta(x_t, t) = f_\theta(x_{t'}, t')$ 之上。在没有任何解析解的情况下，它通过优化相邻离散时间步的连锁递推误差来实现全局映射，其目标函数为：

$$\mathbb{E} [d(f_\theta(x_{t_k}, t_k), f_{\bar{\theta}}(x_{t_{k-1}}, t_{k-1}))]$$

- MeanFlow [8]:

依据积分定理将特定时间区间内的平均速度严谨定义为 $u_\theta(x_t, r, t) = \frac{1}{t-r} \int_r^t v_\theta(x_\tau, \tau) d\tau$ 。核心突破在于推导出了 MeanFlow 恒等式，解析地连接了平均速度与瞬时速度：

$$v_\theta(x_t, t) = u_\theta(x_t, r, t) + (t - r) \frac{d}{dt} u_\theta(x_t, r, t)$$

通过引入停止梯度截断二阶导数运算，最终使得模型可以直接基于该恒等式高效学习平均位移。

- **Shortcut [5]:**

理论基础是相对距离上的自洽性公理：一个大跨度跃迁的结果必须等价于细分小步生成的累加。它优化的是一个基于相对距离缩放的自洽性损失，要求大跨度（ $2d$ ）预测结果与连续两次小跨度（ d ）生成的聚合靶点 $s_{\text{consistency}}$ 完全重合，公式表达为：

$$\mathbb{E} [\|s_{\theta}(x_t, t, c, 2d, w) - s_{\text{consistency}}\|^2]$$

2.10 Score Distillation

原始 Score Distillation Sampling (SDS) 的梯度形式为：

$$\nabla_{\theta} \mathcal{L}_{\text{SDS}} \propto (\epsilon_{\text{pretrain}}(x_t, t, y) - \epsilon) \frac{\partial x_t}{\partial \theta}$$

变分的 Variational Score Distillation (VSD) 梯度公式为：

$$\nabla_{\theta} \mathcal{L}_{\text{VSD}} \propto (\epsilon_{\text{pretrain}}(x_t, t, y) - \epsilon_{\phi}(x_t, t, c, y)) \frac{\partial x_t}{\partial \theta}$$

两者的唯一区别是在减号右侧这项——SDS 减去的是简单的高斯噪声 ϵ ，而 VSD 减去的是由学生网络（如 LoRA）预测的噪声 ϵ_{ϕ} 。

SDS 的直观隐喻： SDS 假定你生成的 3D 模型其实只是一个无意义的点，优化过程等价于：老师（大模型， $\epsilon_{\text{pretrain}}$ ）告诉你“往东走才像沙堡”，而你自己完全是杂乱的噪声（ ϵ ），于是你就“用老师说的方向，减去自己的随机起点方向”。

问题与后果： 由于大模型见过成千上万种“沙堡”，而你这个点对模型而言只是随机噪声，最终老师只能“平均”所有沙堡的形态教给你，导致结果极为平滑、缺乏细节——Mode Collapse（模式坍塌）现象。

而 VSD 的本质改动，让学生网络主动生成自己对噪声的估计，再和老师的指导对齐，极大丰富了可表达性，也避免了 SDS 那种“所有样本都被平均拉向一处”的模糊化本质问题。

2.10.1 KL 梯度流与 Score Distillation 的统一

核心目标： 证明当学生分布 $q(x)$ 逼近教师分布 $p(x)$ （即最小化 $D_{\text{KL}}(q||p)$ ）时，空间中每个粒子的最速下降速度场（在 Wasserstein 意义下）满足

$$v(x) = S_p(x) - S_q(x).$$

定义教师与学生的 Score：

$$S_p(x) = \nabla_x \log p(x), \quad S_q(x) = \nabla_x \log q(x).$$

有

$$S_q(x) - S_p(x) = \nabla_x \log q(x) - \nabla_x \log p(x) = \nabla_x (\log q(x) - \log p(x)) = \nabla_x \log \frac{q(x)}{p(x)}.$$

考虑泛函

$$\mathcal{F}[q] = D_{\text{KL}}(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx.$$

对被积项按 q 求导（等价于求第一变分）：

$$\frac{d}{dq} \left(q \log \frac{q}{p} \right) = \log \frac{q}{p} + q \cdot \frac{d}{dq} \left(\log \frac{q}{p} \right) = \log \frac{q}{p} + q \cdot \frac{1}{q} = \log \frac{q}{p} + 1.$$

故

$$\frac{\delta \mathcal{F}}{\delta q(x)} = \log \frac{q(x)}{p(x)} + 1 =: E(x).$$

这里 $E(x)$ 可解释为位置 x 处粒子的相对信息势能。

在 Wasserstein 梯度流框架下，最速下降速度场为

$$v(x) = -\nabla_x \left(\frac{\delta \mathcal{F}}{\delta q(x)} \right) = -\nabla_x E(x).$$

代入上式得到

$$v(x) = -\nabla_x \left(\log \frac{q(x)}{p(x)} + 1 \right) = -\nabla_x \log \frac{q(x)}{p(x)} = -(\nabla_x \log q(x) - \nabla_x \log p(x)) = S_p(x) - S_q(x).$$

备注 2.10.1. 设概率密度演化为 $\rho_t(x)$ ，速度场为 $v_t(x)$ 。质量守恒给出连续性方程

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

给定能量泛函 $\mathcal{F}[\rho]$ ，其沿时间的变化率为

$$\frac{d}{dt} \mathcal{F}[\rho_t] = \int \frac{\delta \mathcal{F}}{\delta \rho_t}(x) \partial_t \rho_t(x) dx.$$

代入连续性方程并做分部积分（边界项取零）得到

$$\frac{d}{dt} \mathcal{F}[\rho_t] = - \int \frac{\delta \mathcal{F}}{\delta \rho_t} \nabla \cdot (\rho_t v_t) dx = \int \rho_t \nabla \left(\frac{\delta \mathcal{F}}{\delta \rho_t} \right) \cdot v_t dx.$$

在 Wasserstein-2 几何中，切向量范数定义为

$$\|v_t\|_{T_{\rho_t} \mathcal{P}_2}^2 = \int \rho_t(x) \|v_t(x)\|^2 dx.$$

因此，“最速下降”就是在给定该范数尺度下，让 $\frac{d}{dt} \mathcal{F}[\rho_t]$ 最小。由 Cauchy-Schwarz 不等式，最优方向必须为

$$\nabla \left(\frac{\delta \mathcal{F}}{\delta \rho_t} \right)$$

2.11 Flow Distillation

在连续流（Continuous-in-time Flow）模型中，数据的生成过程被描述为一个ODE：

$$dx_t = v(x_t, t)dt$$

其中 $t \in [0, 1]$, $x_0 \sim p_0$, $x_1 \sim p_1$

教师模型已经训练好了一个速度场 $v_p(x, t)$ ，它能完美地把噪声推向目标分布。流蒸馏的目标，是训练一个学生模型 $v_q(x, t)$ ，让它在任意时间 t 和任意空间位置 x 上，都能模仿老师的速度。因此，标准的流蒸馏损失函数定义为两者速度场的均方误差积分：

$$\mathcal{L}_{distill} = \int_0^1 \mathbb{E}_{x \sim p_t} \left[\|v_q(x, t) - v_p(x, t)\|^2 \right] dt$$

直接在上面的积分里处理 p_t 是非常困难的，因为边缘分布 p_t 随着时间变化。为了简化，在 Rectified Flow / OT-Flow 的设定下，粒子走的是匀速直线运动。对于从 x_0 出发的粒子，它在 t 时刻的位置 x_t 是：

$$x_t = (1 - t)x_0 + tx_1 = x_0 + t(x_1 - x_0)$$

因为是匀速直线运动，粒子在其整条轨迹上的速度是一个常数（仅与起点和终点有关）：

$$v(x_t, t) = \frac{dx_t}{dt} = x_1 - x_0$$

此时，我们根据 Brenier 极分解定理2.9.1，代入最优传输的终点坐标 $x_1 = \nabla\Phi(x_0)$ ，得到每个粒子从起点 x_0 触发时的速度场：

$$V(x_0) = \nabla\Phi(x_0) - x_0$$

（注：这里大写的 $V(x_0)$ 表示粒子在整个直线上恒定的速度向量，它仅依赖于初始位置 x_0 ）

我们把 OT-Flow 的直线特性代回第一步的蒸馏目标公式中。因为整个轨迹上速度是恒定的 $v(x_t, t) \equiv V(x_0)$ ，并且我们把关于 $x \sim p_t$ 的期望通过路径映射拉回到底层的高斯噪声 $x_0 \sim p_0$ ，时间积分 $\int_0^1 dt$ 直接被消去了。流蒸馏的优化目标被极大地简化为：

$$\mathcal{L}_{distill} = \mathbb{E}_{x_0 \sim p_0} \left[\|V_q(x_0) - V_p(x_0)\|^2 \right]$$

老师给出的最优路径速度： $V_p(x_0) = \nabla\Phi_p(x_0) - x_0$ 学生当前的预测路径速度： $V_q(x_0) = \nabla\Phi_q(x_0) - x_0$ 代入损失函数：

$$\mathcal{L}_{distill} = \mathbb{E}_{x_0 \sim p_0} \left[\|(\nabla\Phi_q(x_0) - x_0) - (\nabla\Phi_p(x_0) - x_0)\|^2 \right]$$

$$\mathcal{L}_{distill} = \mathbb{E}_{x_0 \sim p_0} \left[\|\nabla\Phi_q(x_0) - \nabla\Phi_p(x_0)\|^2 \right] = \mathbb{E}_{x_0 \sim p_0} \left[\|T_q(x_0) - T_p(x_0)\|^2 \right]$$

在最优传输流的框架下，匹配两个网络的速度场等价于最小化它们所对应的 Brenier 凸势能梯度场之间的 L^2 距离。

Di-Bregman 是一种紧凑的统一框架，旨在通过基于 Bregman 散度的密度比匹配来重新定义扩散模型蒸馏[27]。

2.11.1 flow distillation、Bregman散度、Wasserstein空间的统一

MSE Loss是 Wasserstein 距离的上界

已知 p_1 与 q_1 之间的 2-Wasserstein 距离平方为:

$$W_2^2(p_1, q_1) = \inf_{\pi \in \Pi(p_1, q_1)} \int_{\mathcal{X} \times \mathcal{X}} \|y - z\|^2 d\pi(y, z)$$

其中 $\Pi(p_1, q_1)$ 表示所有以 p_1 和 q_1 为边缘分布的联合概率测度的集合。

为了联系映射函数 T_p, T_q 与联合分布 π ，通过基础测度 p_0 构造一个特定的联合测度 $\hat{\pi}$ 。定义映射 $F: \mathcal{X} \rightarrow \mathcal{X} \times \mathcal{X}$ ，使得 $F(x) = (T_p(x), T_q(x))$ 。令 $\hat{\pi}$ 为 p_0 经过 F 的推前测度，即:

$$\hat{\pi} = F_{\#}p_0 = (T_p, T_q)_{\#}p_0$$

由于 W_2^2 是在所有可能的耦合集合 Π 中取下确界，而我们构造的 $\hat{\pi}$ 只是其中一个合法的耦合，因此必然满足:

$$W_2^2(p_1, q_1) \leq \int_{\mathcal{X} \times \mathcal{X}} \|y - z\|^2 d\hat{\pi}(y, z)$$

利用推前测度积分公式:

$$\int_{\mathcal{X} \times \mathcal{X}} \|y - z\|^2 d\hat{\pi}(y, z) = \int_{\mathcal{X}} \|T_p(x) - T_q(x)\|^2 dp_0(x)$$

结论 1:

$$W_2^2(p_1, q_1) \leq \mathbb{E}_{x \sim p_0} [\|T_p(x) - T_q(x)\|^2] = \mathbb{E}_{x \sim p_0} [\|V_p(x) - V_q(x)\|^2]$$

这说明最小化速度场之间的 L^2 距离，等价于最小化生成分布与目标分布之间 W_2 距离的上界。

MSE Loss 等价于泛函空间上的 Bregman 散度

定义平方可积函数空间 $\mathcal{H} = L^2(\mathcal{X}, p_0; \mathbb{R}^d)$ 。对于任意两个向量场 $F, G \in \mathcal{H}$ ，其内积定义为:

$$\langle F, G \rangle_{\mathcal{H}} = \int_{\mathcal{X}} \langle F(x), G(x) \rangle dp_0(x)$$

相应的范数为 $\|F\|_{\mathcal{H}}^2 = \langle F, F \rangle_{\mathcal{H}}$ 。

定义严格凸泛函与 Fréchet 导数在空间 \mathcal{H} 上，定义“动能”泛函 $\Omega: \mathcal{H} \rightarrow \mathbb{R}$ 为:

$$\Omega(T) = \frac{1}{2} \|T\|_{\mathcal{H}}^2 = \frac{1}{2} \int_{\mathcal{X}} \|T(x)\|^2 dp_0(x)$$

为了定义 Bregman 散度，需要求 Ω 关于映射 T 的 Fréchet 导数（变分导数）。对于任意扰动场 $H \in \mathcal{H}$:

$$\lim_{\epsilon \rightarrow 0} \frac{\Omega(T + \epsilon H) - \Omega(T)}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{\frac{1}{2} \langle T + \epsilon H, T + \epsilon H \rangle_{\mathcal{H}} - \frac{1}{2} \langle T, T \rangle_{\mathcal{H}}}{\epsilon}$$

展开内积得到：

$$= \lim_{\epsilon \rightarrow 0} \frac{\epsilon \langle T, H \rangle_{\mathcal{H}} + \frac{1}{2} \epsilon^2 \|H\|_{\mathcal{H}}^2}{\epsilon} = \langle T, H \rangle_{\mathcal{H}}$$

由此可知，泛函 Ω 在 T 处的 Fréchet 梯度为恒等映射：

$$\delta\Omega(T) = T$$

Bregman 散度在泛函空间中的一般定义：

$$D_{\Omega}(F, G) = \Omega(F) - \Omega(G) - \langle \delta\Omega(G), F - G \rangle_{\mathcal{H}}$$

将 $F = T_p$ 且 $G = T_q$ ，并代入 Ω 的具体形式和 Fréchet 导数 $\delta\Omega(T_q) = T_q$ ：

$$D_{\Omega}(T_p, T_q) = \frac{1}{2} \|T_p\|_{\mathcal{H}}^2 - \frac{1}{2} \|T_q\|_{\mathcal{H}}^2 - \langle T_q, T_p - T_q \rangle_{\mathcal{H}}$$

展开内积：

$$D_{\Omega}(T_p, T_q) = \frac{1}{2} \|T_p\|_{\mathcal{H}}^2 - \frac{1}{2} \|T_q\|_{\mathcal{H}}^2 - \langle T_q, T_p \rangle_{\mathcal{H}} + \|T_q\|_{\mathcal{H}}^2$$

合并关于 $\|T_q\|_{\mathcal{H}}^2$ 的项：

$$= \frac{1}{2} \|T_p\|_{\mathcal{H}}^2 - \langle T_p, T_q \rangle_{\mathcal{H}} + \frac{1}{2} \|T_q\|_{\mathcal{H}}^2$$

根据内积空间的完全平方展开式，此式精确等价于：

$$= \frac{1}{2} \|T_p - T_q\|_{\mathcal{H}}^2$$

展开为积分形式：

$$D_{\Omega}(T_p, T_q) = \frac{1}{2} \int_{\mathcal{X}} \|T_p(x) - T_q(x)\|^2 dp_0(x)$$

引入 Flow Matching 速度场 $V(x) = T(x) - x$ 。因此，师生速度场的差值为：

$$V_p(x) - V_q(x) = (T_p(x) - x) - (T_q(x) - x) = T_p(x) - T_q(x)$$

代入上述公式，最终得到：

$$D_{\Omega}(T_p, T_q) = \frac{1}{2} \int_{\mathcal{X}} \|V_p(x) - V_q(x)\|^2 dp_0(x) = \frac{1}{2} \mathcal{L}_{\text{MSE}}$$

结论 2：速度场的均方误差积分等于函数空间 $L^2(p_0)$ 中，以动能泛函 $\Omega[T] = \frac{1}{2} \int \|T(x)\|^2 p_0(x) dx$ 所诱导的 Bregman 散度。

令优化的目标泛函为 $\mathcal{F}[T_q] = D_{\Omega}(T_p, T_q)$ 。对该泛函按学生映射 T_q 求 Fréchet 导数：

$$\frac{\delta\mathcal{F}}{\delta T_q(x)} = T_q(x) - T_p(x) = V_q(x) - V_p(x)$$

在 $L^2(p_0)$ 泛函梯度流框架下，映射 T_q 的最速下降演化方向为：

$$\frac{\partial T_q(x)}{\partial \tau} = -\frac{\delta\mathcal{F}}{\delta T_q(x)} = V_p(x) - V_q(x)$$

这说明：Flow Distillation 中目标速度场与当前速度场的差值 $(V_p - V_q)$ ，正是 Bregman 散度在泛函空间中的最速下降梯度。

在实际中，映射函数由参数化神经网络 $V_q(x; \theta)$ 实现。为了超越单一参数点态估计导致的模式坍塌与多样性丧失，我们将神经网络参数 θ 建模为参数空间 Θ 上的概率分布 $\mu(\theta) \in \mathbb{W}_2(\Theta)$ 。在变分推断框架下，我们的目标在参数测度空间中极小化期望蒸馏泛函：

$$\min_{\mu \in \mathbb{W}_2(\Theta)} \mathcal{E}[\mu] := \mathbb{E}_{\theta \sim \mu} [\mathcal{L}_{distill}(\theta)]$$

基于前面的分析，参数分布 μ_τ 将沿着该泛函在 $\mathbb{W}_2(\Theta)$ 空间中的最速下降方向演化。对应到微观层面，依据链式法则，每个参数粒子 $\theta_\tau \sim \mu_\tau$ 的更新法则为：

$$\frac{d\theta_\tau}{d\tau} = -\mathbb{E}_{x_0 \sim p_0} \left[\left(\frac{\partial V_q(x_0; \theta_\tau)}{\partial \theta_\tau} \right)^\top (V_q(x_0; \theta_\tau) - V_p(x_0)) \right]$$

驱动网络演化的本质，正是参数流形在试图逼近泛函空间中 Bregman 散度的最速下降路径。

Chapter 3

Agent

3.1 面向 Lean 4 的语义检索与混合排序

3.1.1 问题背景

Lean 4 生态中，mathlib4 等库体量巨大；官方文档按**形式名称**检索、GitHub 上按**关键词**搜源码与 docstring，对不熟悉命名惯例的用户都不友好，且许多定理缺少 docstring，纯词面匹配也难以覆盖**语义相近但表述不同**的查询。因此需要能接受**非形式或自然语言**输入、返回相关形式条目的搜索引擎 [7, 1]。

3.1.2 LeanSearch: 非形式化语料 + 稠密向量检索 + 查询增广 [7]

数据侧（离线）：将 Mathlib4 中每条定理的**形式陈述**，经大语言模型翻译成**非形式名称与陈述**，与原名、docstring、以及通过文档超链接抽取的**相关定义**一并作为上下文喂给 LLM，以减轻“只看形式式子”带来的歧义。条目在库中存为**形式-非形式**配对；非形式侧整理为 **定理名：非形式陈述** 的文本，供向量化。

索引与检索：对非形式语料用**稠密嵌入模型**编码，写入向量库（文中采用 Chroma；近邻搜索使用 HNSW）。对**文档与查询**使用带任务说明的 instruction 模板（要求检索与给定条目**数学等价**的双语 Lean4+自然语言定理），以贴合检索任务、提升零样本表现。

查询侧（在线）：用户短查询先经 LLM **增广**为更完整、含 LaTeX 的说明，并尽量给出与原文等价的**非形式 + 形式**表述，再按与文档相同的结构套上 instruction 后嵌入，在嵌入空间用余弦距离取近邻。

评价：作者另建 Mathlib4 语义检索基准（多领域查询、分组标注、Precision@ k / Recall@ k 等），用于对比不同引擎与消融。

要点（检索视角）：LeanSearch 的检索主干是**单信号语义检索**（双塔式：查询与文档分别嵌入 + ANN），强项在于**数据构造**（上下文丰富的非形式化）与**查询改写**对齐文档格式；未在文中强调与 BM25 或图结构的显式融合。

3.1.3 LeanExplore: 多面嵌入 + BM25+ + PageRank 的混合排序 [1]

StatementGroup: Lean 中一条用户写的 `structure / inductive / theorem` 等往往会展开为多条底层声明。LeanExplore 将同一源码连续块生成的声明归为一条 **StatementGroup**，以主声明的 Lean 名与 docstring 作为组的代表，避免结果碎片化。

数据增强: 在 StatementGroup 级构建依赖有向图；按拓扑序用 LLM 为每组生成英文非形式描述，并在 prompt 中注入已翻译好的直接依赖组的文本，使翻译连贯。同一图上的对数 PageRank 作为库内重要性 / 中心性先验。

检索与排序 (核心):

- **语义通道:** 对每组维护多路文本——名称、docstring、由“非形式翻译 + 源文件路径抽取的关键词”拼成的复合非形式句等——分别用 `bge-base-en-v1.5` 嵌入；FAISS (倒排量化单元 + 单元内精确距离) 做 kNN；距离映射到 $[0, 1]$ 相似度后，用可配置阈值 (默认 0.525) 过滤候选。
- **词法通道:** 对通过语义阈值的候选，在拼接了名、docstring、非形式、形式展示文本、路径的文本上与用户查询算 BM25+。
- **结构通道:** 每条候选带上预计算的 PageRank。
- **融合:** 三类分数在当前候选集内分别 min-max 归一化到 $[0, 1]$ ，再按权重线性组合 (默认: 语义 1.0, BM25+ 1.0, PageRank 0.2)，降序输出。

与 Agent 的衔接: 提供网站、Python API、本地/远程模式，以及 MCP (Model Context Protocol) 服务，便于 LLM 将“搜 Lean 声明”作为工具调用，支撑对话式学习与自动证明中的前提查找 [1]。

3.1.4 两条路线对照 (检索方法)

维度	LeanSearch [7]	LeanExplore [1]
索引单元	定理级；非形式化文本 + 形式陈述配对	StatementGroup (用户级代码块)；多面文本嵌入
主检索信号	稠密嵌入 + 余弦 (Chroma/HNSW)	FAISS kNN 语义相似度
辅助信号	(文中以语义为主)	BM25+ 词法相关 + 依赖图 PageRank
查询处理	LLM 增广，使查询结构与文档对齐	直接嵌入用户查询 (可与混合分融合)

个人小结：若要把检索接入 ATP agent，LeanSearch 范式突出“把库与用户都拉到同一非形式语义空间”（离线非形式化 + 在线查询增广）；LeanExplore 范式突出“语义召回 + 词法纠偏 + 库图重要性”的三路融合，并解决 Lean 展开带来的展示粒度问题。二者互补点在于：前者强在查询-文档表述对齐，后者强在多信号排序与工程化 Agent 接口 (MCP)。

3.2 终身学习 LLM Agent：路线图与 AutoSkill

3.2.1 《Lifelong Learning of Large Language Model based Agents: A Roadmap》[26]

这是一篇综述，系统整理面向终身学习（持续 / 增量学习）的 LLM Agent：与仅把 LLM 当作静态黑盒、在数据流上持续微调不同，这里强调 Agent 与动态环境交互、在任务流中持续适应。全文按 Agent 三大构件组织：感知（单模态 / 多模态输入融合）、记忆（存储与检索随时间演化的知识，并区分工作记忆、情景记忆、语义记忆、参数化记忆等）、行动（在环境中执行接地、检索、推理等）。核心困难是灾难性遗忘与可塑性丧失之间的稳定性-可塑性权衡；文内还覆盖评测指标、基准、应用场景与开放问题。

3.2.2 《AutoSkill: Experience-Driven Lifelong Learning via Skill Self-Evolution》[24]

面向个性化、跨会话场景：用户会反复表达稳定偏好（如减少幻觉、文体规范、 workflow 约束），但以往多停留在检索式记忆，难以把交互沉淀为可复用能力。AutoSkill 把重复交互抽象为显式技能 (skill)：标准化为可编辑、可版本化的 SKILL.md 工件，包含名称、描述、可执行指令、触发条件、标签、示例与版本等；不更新基座模型参数，作为与模型无关的插件层。

运行时双环：（1）技能增强生成：对用户查询做检索式改写（消解指代、继承任务锚点），再混合检索技能库——稠密向量相似度与 BM25 分数归一化后加权，将命中技能注入上下文后生成回复；（2）技能演化：从新对话中抽取候选技能，经管理模块决定新增 / 合并 / 丢弃，必要时合并与版本递增，使技能库随经验自演化。抽取阶段主要依据用户查询序列而非模型回复，以降低噪声。该框架与「路线图」中的记忆-行动设计相呼应，但把积累单元从原始文本记忆提升为可组合、可审计的行为知识。

Chapter 4

PDE

4.1 分部积分公式

定理 4.1.1 (带系数的分部积分 (加权 Green 公式)). 设 $\Omega \subset \mathbb{R}^2$ 为具有分段光滑边界 $\partial\Omega$ 的有界域, n 为 $\partial\Omega$ 的外法向. 若 $p \in C^1(\bar{\Omega})$ 且 $u, v \in C^2(\bar{\Omega})$, 则恒有

$$\iint_{\Omega} \nabla \cdot (p \nabla u) v \, dx dy = \oint_{\partial\Omega} p \frac{\partial u}{\partial n} v \, ds - \iint_{\Omega} (p \nabla u) \cdot \nabla v \, dx dy.$$

其中 $\frac{\partial u}{\partial n} = \nabla u \cdot n$ 为外法向导数。

证明. 由乘积法则可得

$$\nabla \cdot ((p \nabla u) v) = (\nabla \cdot (p \nabla u)) v + (p \nabla u) \cdot \nabla v.$$

将上式改写为

$$(\nabla \cdot (p \nabla u)) v = \nabla \cdot ((p \nabla u) v) - (p \nabla u) \cdot \nabla v.$$

对 Ω 积分并使用散度定理 (高斯公式):

$$\iint_{\Omega} \nabla \cdot ((p \nabla u) v) \, dx dy = \oint_{\partial\Omega} ((p \nabla u) v) \cdot n \, ds = \oint_{\partial\Omega} p (\nabla u \cdot n) v \, ds = \oint_{\partial\Omega} p \frac{\partial u}{\partial n} v \, ds.$$

于是得到所需恒等式:

$$\iint_{\Omega} \nabla \cdot (p \nabla u) v \, dx dy = \oint_{\partial\Omega} p \frac{\partial u}{\partial n} v \, ds - \iint_{\Omega} (p \nabla u) \cdot \nabla v \, dx dy.$$

证毕。 □

参考文献

- [1] Justin Asher. Leanexplore: A search engine for lean 4 declarations. <https://arxiv.org/abs/2506.11085>, 2025. arXiv:2506.11085; website <https://www.leanexplore.com/>; code <https://github.com/justincasher/lean-explore>.
- [2] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84:375–393, 2000.
- [3] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.
- [4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, et al. Scaling rectified flow transformers for high-resolution image synthesis. <https://arxiv.org/abs/2403.03206>, 2024. arXiv:2403.03206.
- [5] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. <https://arxiv.org/abs/2410.12557>, 2024. arXiv:2410.12557.
- [6] Minghao Fu, Guo-Hua Wang, Tianyu Cui, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Diffusion-sdpo: Safeguarded direct preference optimization for diffusion models. <https://arxiv.org/abs/2511.03317>, 2025. arXiv:2511.03317.
- [7] Guoxiong Gao, Haocheng Ju, Jiedong Jiang, Zihan Qin, and Bin Dong. A semantic search engine for mathlib4. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8001–8013, 2024. Preprint arXiv:2403.13310; demo <https://leansearch.net>; code <https://github.com/reaslab/LeanSearch>.
- [8] Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. <https://arxiv.org/abs/2505.13447>, 2025. arXiv:2505.13447.

- [9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.
- [10] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.
- [11] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *USSR Computational Mathematics and Mathematical Physics*, 1976.
- [12] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.
- [13] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.
- [14] Robert J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.
- [15] Long Ouyang, Jeffrey Wu, Xu Jiang, Dimitre Almeida, Caleb Wainwright, P Mishkin, Jack Zhang, and etc. Training language models to follow instructions with human feedback. <https://arxiv.org/abs/2203.02155>, 2022. arXiv:2203.02155.
- [16] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 1964.
- [17] Rafael Rafailov, Tong Lu, James Kim, et al. Direct preference optimization: Your language model is secretly a reward model. <https://arxiv.org/abs/2305.18290>, 2023. arXiv:2305.18290.
- [18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.
- [19] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. <https://arxiv.org/abs/2010.02502>, 2020. arXiv:2010.02502.
- [20] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

- [21] Yang (Jiaming) Song and Stefano Ermon. Score-based generative modeling through stochastic differential equations. <https://arxiv.org/abs/1907.05600>, 2019. arXiv:1907.05600.
- [22] Weijie Su, Stephen Boyd, and Emmanuel Candès. A differential equation for modeling nesterov’s accelerated gradient method: Theory and insights. *Journal of Machine Learning Theory*, 2014.
- [23] Cédric Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2008.
- [24] Yutao Yang, Junsong Li, Qianjun Pan, Bihao Zhan, Yuxuan Cai, Lin Du, Jie Zhou, Kai Chen, Qin Chen, Xin Li, Bo Zhang, and Liang He. AutoSkill: Experience-driven lifelong learning via skill self-evolution. <https://arxiv.org/abs/2603.01145>, 2026. arXiv:2603.01145v2; code <https://github.com/ECNU-ICALK/AutoSkill>.
- [25] Ziyi Zhang, Li Shen, Sen Zhang, Deheng Ye, Yong Luo, Miaojing Shi, Dongjing Shan, Bo Du, and Dacheng Tao. Aligning few-step diffusion models with dense reward difference learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. TPAMI 2026 accepted; see <https://ieeexplore.ieee.org/document/11408100>.
- [26] Junhao Zheng, Chengming Shi, Xidi Cai, Qiuke Li, Duzhen Zhang, Chenxing Li, Dong Yu, and Qianli Ma. Lifelong learning of large language model based agents: A roadmap. <https://arxiv.org/abs/2501.07278>, 2025. arXiv:2501.07278v2; survey resources <https://github.com/qianlima-lab/awesome-lifelong-llm-agent>.
- [27] Yuanzhi Zhu, Eleftherios Tsonis, Lucas Degeorge, and Vicky Kalogeiton. One-step diffusion models with bregman density ratio matching. <https://arxiv.org/abs/2510.16983>, 2025. arXiv:2510.16983.
- [28] Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. Diffusion reinforcement learning via centered reward distillation. <https://arxiv.org/abs/2603.14128>, 2026. arXiv:2603.14128.

命题 .0.1 (McCann 位移插值的测地线性质). 设 $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ 为具有有限二阶矩的概率测度, 存在从 μ_0 到 μ_1 的最优传输映射 T (即 $T_{\#}\mu_0 = \mu_1$ 且 $W_2^2(\mu_0, \mu_1) = \int \|x - T(x)\|^2 d\mu_0(x)$). 定义映射路径

$$\pi_t = (1-t)\text{Id} + tT, \quad t \in [0, 1],$$

及相应的测度演化路径 $\mu_t = (\pi_t)_{\#}\mu_0$. 则测度路径 μ_t 是 Wasserstein 空间 $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ 中的一条匀速测地线, 即对于任意 $0 \leq s \leq t \leq 1$, 严格有

$$W_2(\mu_s, \mu_t) = (t-s)W_2(\mu_0, \mu_1).$$

证明. 当 $t=0$ 时, $\pi_0 = \text{Id}$, 故 $\mu_0 = \text{Id}_{\#}\mu_0$; 当 $t=1$ 时, $\pi_1 = T$, 且 $\mu_1 = T_{\#}\mu_0$, 分别与初始和目标分布一致。

为计算 μ_s 与 μ_t 间的 W_2 距离, 考虑对 μ_0 上单点 x , 令 $y = \pi_s(x), z = \pi_t(x)$, 构造联合概率分布

$$\gamma_{s,t} = (\pi_s, \pi_t)_{\#}\mu_0.$$

则 $\gamma_{s,t}$ 的边缘分布分别为 μ_s, μ_t , 故 $\gamma_{s,t} \in \Pi(\mu_s, \mu_t)$ 是一种合法耦合, 简单来说就是将同一原点映射得到的两个点配对。由定义有

$$W_2^2(\mu_s, \mu_t) \leq \int \|y - z\|^2 d\gamma_{s,t}(y, z).$$

再利用 $\gamma_{s,t}$ 的定义, 变量回代到 μ_0 空间上得

$$\int \|y - z\|^2 d\gamma_{s,t}(y, z) = \int \|\pi_s(x) - \pi_t(x)\|^2 d\mu_0(x).$$

且有

$$\pi_s(x) - \pi_t(x) = [(1-s)x + sT(x)] - [(1-t)x + tT(x)] = (t-s)[x - T(x)].$$

代入上式, 得

$$\int \|(t-s)[x - T(x)]\|^2 d\mu_0(x) = (t-s)^2 \int \|x - T(x)\|^2 d\mu_0(x).$$

而 $\int \|x - T(x)\|^2 d\mu_0(x) = W_2^2(\mu_0, \mu_1)$, 因此

$$W_2^2(\mu_s, \mu_t) \leq (t-s)^2 W_2^2(\mu_0, \mu_1).$$

两端开方, 得

$$W_2(\mu_s, \mu_t) \leq (t-s)W_2(\mu_0, \mu_1). \quad (1)$$

Wasserstein 空间为完备度量空间, 满足三角不等式。对路径上任意 $0 < s < 1$,

$$W_2(\mu_0, \mu_1) \leq W_2(\mu_0, \mu_s) + W_2(\mu_s, \mu_1).$$

由上述不等式，对区间两段分别有 $W_2(\mu_0, \mu_s) \leq sW_2(\mu_0, \mu_1)$ ， $W_2(\mu_s, \mu_1) \leq (1-s)W_2(\mu_0, \mu_1)$ ，相加得

$$W_2(\mu_0, \mu_s) + W_2(\mu_s, \mu_1) \leq (s + 1 - s)W_2(\mu_0, \mu_1) = W_2(\mu_0, \mu_1).$$

因此必须等号成立，即

$$W_2(\mu_0, \mu_s) = sW_2(\mu_0, \mu_1), \quad W_2(\mu_s, \mu_1) = (1 - s)W_2(\mu_0, \mu_1).$$

同理可证，对任意 $0 \leq s < t \leq 1$ ，

$$W_2(\mu_s, \mu_t) = (t - s)W_2(\mu_0, \mu_1).$$

定理得证。 □

命题 .0.2 (Kantorovich 对偶问题). *Kantorovich* 最优传输原问题

$$\inf_{\pi \in \Pi(\mu, \nu)} \int_{X \times Y} c(x, y) d\pi(x, y)$$

可等价转化为如下对偶问题：

$$\sup_{\substack{\phi, \psi \\ \phi(x) + \psi(y) \leq c(x, y)}} \left(\int_X \phi(x) d\mu(x) + \int_Y \psi(y) d\nu(y) \right),$$

其中约束 $\phi(x) + \psi(y) \leq c(x, y)$ 对所有 (x, y) 成立。

证明. 原问题是寻找联合分布（传输计划） $\pi(x, y)$ ，使总传输成本最小，即

$$\inf_{\pi \in \Pi(\mu, \nu)} \int_{X \times Y} c(x, y) d\pi(x, y),$$

其中 $\pi \in \Pi(\mu, \nu)$ 表示 π 的边缘分布为 μ 和 ν 。

在连续视角下，可通过测试函数 $\phi(x)$ 和 $\psi(y)$ 来构造对偶变量，等价表达边缘分布约束，即对任意连续有界函数 ϕ, ψ ，

$$\int_{X \times Y} (\phi(x) + \psi(y)) d\pi(x, y) = \int_X \phi(x) d\mu(x) + \int_Y \psi(y) d\nu(y).$$

据此构造“指示函数”型惩罚项：

$$\sup_{\phi, \psi} \left(\int_X \phi d\mu + \int_Y \psi d\nu - \int_{X \times Y} (\phi(x) + \psi(y)) d\pi(x, y) \right).$$

当 π 满足约束时，上确界为 0；否则趋于 $+\infty$ 。

将惩罚项加入原问题，得到

$$\inf_{\pi \geq 0} \left[\int_{X \times Y} c(x, y) d\pi(x, y) + \sup_{\phi, \psi} \left(\int_X \phi d\mu + \int_Y \psi d\nu - \int_{X \times Y} (\phi(x) + \psi(y)) d\pi(x, y) \right) \right].$$

可合并为拉格朗日函数

$$\inf_{\pi \geq 0} \sup_{\phi, \psi} \left[\int_X \phi d\mu + \int_Y \psi d\nu + \int_{X \times Y} (c(x, y) - \phi(x) - \psi(y)) d\pi(x, y) \right].$$

若满足 Fenchel-Rockafellar 或 Sion 极小极大定理条件, 可交换 \inf, \sup 顺序:

$$\sup_{\phi, \psi} \inf_{\pi \geq 0} \left[\int_X \phi d\mu + \int_Y \psi d\nu + \int_{X \times Y} (c(x, y) - \phi(x) - \psi(y)) d\pi(x, y) \right].$$

内层极小化关于 π 为

$$\inf_{\pi \geq 0} \int_{X \times Y} (c(x, y) - \phi(x) - \psi(y)) d\pi(x, y).$$

若某处有 $c(x, y) - \phi(x) - \psi(y) < 0$, 则 π 可取无穷大量, 使整体趋于 $-\infty$, 故必须满足

$$c(x, y) - \phi(x) - \psi(y) \geq 0, \quad \forall (x, y).$$

满足该约束时积分最小值为 0, 否则为 $-\infty$; 外层取 \sup 时自然避开 $-\infty$ 。

因此最终对偶问题为

$$\sup_{\substack{\phi, \psi \\ \phi(x) + \psi(y) \leq c(x, y)}} \left(\int_X \phi(x) d\mu(x) + \int_Y \psi(y) d\nu(y) \right).$$

这正是正文所述的 Kantorovich 对偶问题。 □