# Likun Wang research note@BlueWhaleLab

## Project Description:

> 基于能量模型的策略微调

### 课题核心：

在基础模型的基础上，通过能量模型进行二次优化/微调, 从而获得能量模型调制的目标分布：

$$p(y|x) \propto P_\theta(y|x) \cdot \exp\left(-E_\phi(x,y)\right)$$

其中，p(x)为目标分布， $P_\theta(x)$ 为基础模型， $\exp\left(-E_\phi(x)\right)$ 为能量模型

通过引入人工先验构造能量函数，使得 在给定初始解 $P_\theta(x)$ 上继续优化,从而获得更好的性能表现。

### 细化方向：

### 1.同目标优化：基于能量函数的RL策略优化算子：

令 $P_\theta(a|s) = \pi_\theta(a|s)$ ,其中 $\pi_\theta(a|s)$ 为高斯分布，通过最大熵RL进行优化：

$$J_{\text{MaxEnt}}(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t,a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t))\right)\right]$$

而 $E_\phi(x,y)$ 为能量函数，可定义为 $Q_\phi(x,y)$ 或 $\gamma^H Q_\phi(\mathbf{z}_H, \mathbf{a}_H) + \sum_{h=0}^{H-1} \gamma^h R_\phi(\mathbf{z}_h, \mathbf{a}_h)$

现状：**现有的最大熵RL能够通过显式计算熵的方式，使得高斯策略能够快速逼近与Q有关的Gibbs分布，然而Gibbs分布是一个多模态分布，复杂环境下高斯策略本身不可能贴近一个多模态分布，从而陷入局部最优，同时损失了策略的多模态性。**

核心claim：使用能量模型的梯度信息构建策略优化算子，在初始高斯分布的基础上进行调制，使得后验策略拥有多模态性和最优性。

**相对于高斯策略：**使用更多先验信息（价值函数梯度）通过MCMC/MPPI的方式进行优化，从而使得其获得更精确的优化信息引导和多模态性质。

**相对于朗之万策略：**高斯策略为能量函数引导的MCMC提供优质的初始解，使得优化速度更快，优化更稳定需要步数更少。

实现方式(以langevin采样为例)：

> $$a \sim \pi_\theta(a|s) \xrightarrow{\text{Langevin}} a' \sim p^*(a|s) \propto \exp(Q(s,a)/\alpha)$$
>
> $$a_{t+1} = a_t + \frac{\epsilon}{2}\nabla_a Q(s,a_t) + \sqrt{\epsilon}\cdot\mathcal{N}(0,I), a_0 \sim \pi_\theta(a|s)$$

Energy-matching:

S2AC(差异性):

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{s\sim D, a_\theta^L \sim \pi_\theta}\left[Q_\phi(s, a_\theta^L)\right] + \alpha \mathbb{E}_{s\sim D}\left[H(\pi_\theta(\cdot|s))\right]$$

$$\Delta f(a_i^l) = \mathbb{E}_{a_j^l \sim q_l}\left[k(a_i^l, a_j^l)\nabla_{a_j^l} \log p(a_j^l) + \nabla_{a_j^l} k(a_i^l, a_j^l)\right]$$

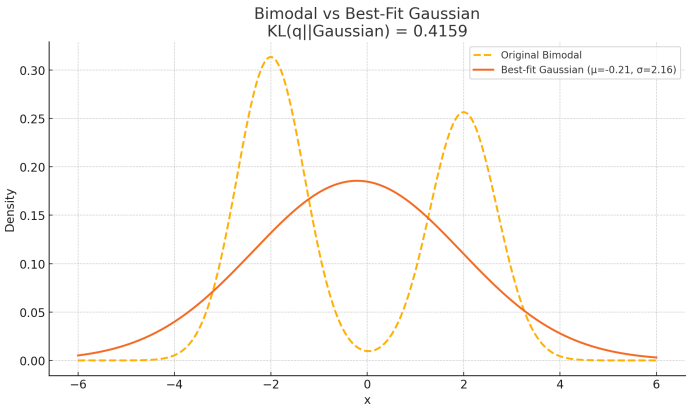
$$\rho_{t+\Delta t} = \arg\min_{\rho} \underbrace{\frac{W_2^2(\rho, \rho_t)}{2\Delta t}}_{\text{Transport Cost}} + \underbrace{\int V_\theta(x) \mathrm{d}\rho(x)}_{\text{Potential Energy}} + \underbrace{\varepsilon(t) \int \rho(x)\log\rho(x)\mathrm{d}x}_{\text{Internal Energy (-Entropy)}}.$$

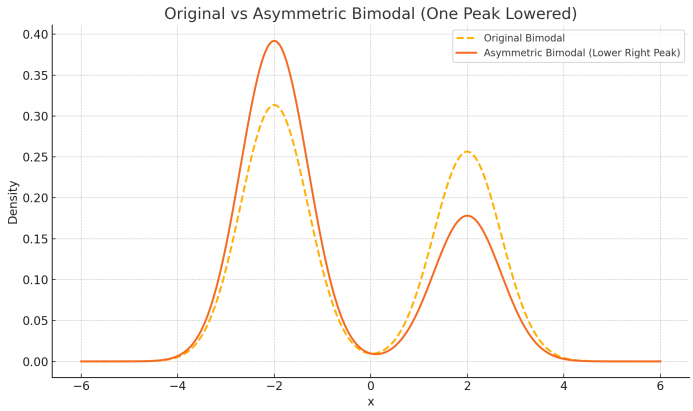
**MDP中，$V_\theta$ 如何学习?**

**二阶段优化法 vs Energy matching 策略**

由于此时高斯策略和策略优化算子的目标函数是同一个(Gibbs目标分布)，高斯策略为能量函数引导的MCMC提供优质的初始解和快速优化，而策略优化算子则针对当前策略进行单调二次优化，并同时获得多模态性。

应用场景：强化学习仿真环境，机器人locomotion任务等 DMC GYM




## 2.异目标优化：基于能量函数策略的调优：

令 $P_\theta(s) = P_{lm}(s)$ ,其中 $P_{lm}(s)$ 为大模型的预训练策略， $E_\phi(x)$ 是与输出有关的能量模型。用于进行大模型微调。

由于二者的优化方向可以不同，大模型为自回归训练，以精度导向为主，而能量模型则与不确定性或者其他先验指标为准，因此二者可以进行紧密结合。

**结合思路：**

1.无监督信号微调：以EMPO为例，通过新的方式估计不确定度，增加模型输出的置信能力（VNE）

2.插拔式微调器：通过独立训练一个重排序器，对大模型的输出内容进行提纯和优化(EORM)，使得其具有泛用性，能够和多种大模型结合

3.分子材料生成器：通过使用更好的模型和微调技术(RL),进行分子材料的生成和优化(CrystalFormer)

(生成能力增强)

## 3.策略结构优化：基于能量匹配的策略优化(Potential-Guided (JKO) AC)

> Claim核心:通过更复杂的策略结构代替高斯策略，将策略的优化表述为一个接近目标能量分布的过程（参考Energy Matching）

令 $P_\theta(a|s) = \pi_\theta(a|s)$ ,其中 $\pi_\theta(a|s)$ 为高斯分布，通过最大熵RL进行优化：

$$J_{\text{MaxEnt}}(\pi) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t \left( r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t)) \right) \right]$$

现状：**现有的最大熵RL能够通过显式计算熵的方式，使得高斯策略能够快速逼近与Q有关的Gibbs分布，然而Gibbs分布是一个多模态分布，复杂环境下高斯策略本身不可能贴近一个多模态分布，从而陷入局部最优，同时损失了策略的多模态性。**

方法：用初始高斯策略+energy/flow matching共同构建策略函数

### 1. JKO scheme:

考虑JKO泛函：s

$$\rho_{t+\Delta t} = \arg\min_{\rho} \underbrace{\frac{W_2^2(\rho, \rho_t)}{2\Delta t}}_{\text{Transport Cost}} + \underbrace{\int V_\theta(a) \mathrm{d}\rho(a)}_{\text{Potential Energy}} + \underbrace{\varepsilon(t) \int \rho(a) \log \rho(a) \mathrm{d}a}_{\text{Internal Energy (-Entropy)}}.$$

其含义是：下一个时刻的分布 $\rho_{t+\Delta t}$ 是在所有可能的概率分布中，能够使后面三项之和**最小化**的那一个。这三项代表了一个权衡：

- **传输成本**: 分布变化不能太剧烈（W距离尽可能小）。
- **势能**: 分布要向势能 $V_\theta(x)$ 更低的区域移动。
- **熵**: 分布要尽可能分散（熵要大，即负熵要小）。

因此，JKO方案不是简单地描述一个随意的演化，而是描述一个在每一步都遵循一个**局部最优**原则的演化过程。它在每一步都找到了一个在“传输成本”、“势能降低”和“熵最大化”之间达到最佳平衡的下一个状态。同时也被称为**Wasserstein空间中的梯度流 (gradient flow in Wasserstein space)**。

### 2. 与策略梯度的关系：

考虑策略分布 $\pi(a|s)$ 的演化过程，将势能函数设定为 $Q_\phi(s,a)$ ,推出 $\pi$ 在Wasserstein 上的最速下降梯度为：

$$\pi_{t+\Delta t}(\cdot \mid s) = \arg\min_{\pi(\cdot|s)} \left\{ \frac{1}{2\Delta t} W_2^2(\pi(\cdot \mid s), \pi_t(\cdot \mid s)) - \frac{1}{\alpha} \int_{\mathcal{A}} Q_\phi(s,a) \pi(a \mid s) da + \varepsilon(t) \int_{\mathcal{A}} \pi(a \mid s) \log \pi(a \mid s) da \right.$$

与策略梯度的关系：它们都遵循着一个共同的优化思想，可以概括为：**“在信任的范围内，最大化（或最小化）某个目标。”**把两者的优化问题抽象成以下形式：

**2.1 TRPO 形式（约束形式）：**

$$J_\pi = \max_{\pi_{\text{new}}}[G_t], \quad s.t. \quad D_{\text{kl}}(\pi_{\text{new}}, \pi_{\text{old}}) \leq \delta$$

其中，D 是一个距离或散度，用来衡量新旧策略的差异，$\delta$ 定义了“信任域”的大小。意思是：找到一个新策略，让收益尽可能提升，但前提是这个新策略不能和老策略‘离得太远’。

**2.2 JKO 形式（正则化/惩罚形式）：**

$$J_\pi = \min_{\pi_{\text{new}}}[\mathcal{F}_\pi + \lambda D_w(\pi_{\text{new}}, \pi_{\text{old}})]$$

其中，距离的惩罚项直接加入到了目标函数中。$\lambda$ 是一个正则化系数，权衡了优化目标和保持策略稳定之间的关系。意思是：找到一个新策略，它既要让成本(势能)尽可能降低，同时也要保证它和老策略‘离得不要太远’。在优化理论中，两种形式（**约束形式**和**惩罚项形式**）在很多情况下是紧密相关的，甚至可以认为是解决同一个问题的不同途径。

2.3 不同点：

**TRPO 使用的度量是 KL 散度：**衡量的是**信息上的差异**。表明两个策略分布在统计上的“距离”。

**JKO 使用的度量是 Wasserstein 距离：**衡量的是**物理移动的成本**。表明两个策略分布相互演化的一种成本，这是一种几何上的“距离”。

## 3. 与其他强化学习范式的联系：

考虑一阶最优性条件：

$$\frac{1}{\Delta t}(a - a') - \nabla_x Q_\theta(x) + \varepsilon(t)\nabla_x \log \rho_t(x) = 0, \quad (x, y) \in \text{supp}(\gamma_t).$$

- 当 $\epsilon(t) = 0$ 时，$\frac{1}{\Delta t}(a - a') - \nabla_x Q_\theta(x) = 0$ . 此时**退化为贪婪策略**，其表明了此时策略的上升方向刚好与价值函数的梯度方向完全一致。
- 当 $t \to \infty$ 时，此时可以认为策略已经到达稳态，$\frac{1}{\Delta t}(a - a') \approx 0$。令 $\alpha$ 此时认为 $-\nabla_x Q_\theta(x) + \varepsilon(t)\nabla_x \log \rho_t(x) = 0$, 重新整理成以下格式：

$$\pi_t(a) = \exp(\frac{Q_\theta(a)}{\varepsilon(t)} + C) \propto \exp(\frac{Q_\theta(a)}{\alpha})$$

此时，该方法演化为**最大熵强化学习**。

## 4. 与能量匹配的区别：

在能量匹配中，目标分布是一个稳态分布，这意味着分布的终点是确定的。所以只要服从JKO的框架，可以自行调制epsilon的值从而实现自由路径选择（即epsilon可以选择）。

分析这个条件可以发现：

**1.远场（off-manifold，**$\varepsilon(t)=0$，退化为速度场:

$$\frac{1}{\Delta t}(x-y)+\nabla_x V_\theta(x)=0.$$

**2.近场（on-manifold，**$\varepsilon(t)=\alpha$，此时认为粒子已经进入了势阱，此时分布趋于平稳，可以假设 $x\approx y$， $\varepsilon(t)=\varepsilon_{max}$ ,此时可以导出:

$$\nabla_x V_\theta(x)+\varepsilon_{max}\nabla_x \log \rho_t(x)=0.$$

$$\rho_\star(x)=\frac{1}{Z}\exp\left(-\frac{1}{\varepsilon_{\max}}V_\theta(x)\right),$$

而在RL中，由于Q值是随着策略的演化而演化的，同时价值函数也存在拟合误差的问题。这使得目标分布是一直变化的，我们拿不到目标地点的真实样本，但是我们能拿到这个能量函数的梯度，在局部指引我们的策略优化方向。此时，epsilon无法自由选择，如果令其等于0，则会陷入到局部最优中。

## 5. Method：

### 5.1 基于行为最优策略的势能景观构建与匹配

由于最优动作是未知的，所以需要通过别的方法构建速度场，采用buffer里面的行为最优策略作为参考构造速度：

$$\mathcal{L}_{\mathrm{OT}}=\mathbb{E}_{a\sim\pi_\theta,s\sim D}[Q^{\pi_\theta}(s,a)]-\mathbb{E}_{s,a\sim D}\left[\lambda(f(Q^{\pi_\beta^*}-Q^{\pi_\theta})\|v_\theta-(a-a^0)\|^2)\right]$$

$$\mathcal{L}_{\mathrm{CD}}=\mathbb{E}_{x\sim p_{\mathrm{data}}}\left[\frac{V_\theta(x)}{\varepsilon_{\max}}\right]-\mathbb{E}_{\tilde{x}\sim \mathrm{sg}(p_{\mathrm{eq}})}\left[\frac{V_\theta(\tilde{x})}{\varepsilon_{\max}}\right]$$

$$\mathcal{L}_\pi=\omega\mathcal{L}_{\mathrm{CD}}+\mathcal{L}_{\mathrm{OT}}$$

为了避免陷入局部最优，引入价值函数作为全局最优引导信号，JKO scheme重塑为：

$$\rho_{t+\Delta t}=\arg\min_\rho \underbrace{\frac{W_2^2(\rho,\rho_t)}{2\Delta t}}_{\text{Transport Cost}}+\underbrace{\int V_\theta(s,a)\mathrm{d}\rho(a)}_{\text{Potential Energy}}-\lambda\underbrace{\int Q_\phi(s,a)\mathrm{d}\rho(a)}_{\text{Q-Function}}+$$

$$\underbrace{\varepsilon(t)\int \rho(a)\log\rho(a)\mathrm{d}a}_{\text{Internal Energy (-Entropy)}}.$$

构建辅助势能场，势能函数定义为： $U_\theta(s,a)=V_\theta(s,a)-\lambda Q_\phi(s,a)$

此时一阶最优性条件满足如下公式：

$$\frac{1}{\Delta t}(a-a')-\nabla_a(V_\theta(s,a)+Q_\phi(s,a))+\varepsilon(t)\nabla_x\log\rho_t(x)=0,\quad (x,y)\in\mathrm{supp}(\gamma_t).$$

Gradient descend

采用最优行为最优策略的动作作为 $p_{\mathrm{eq}}$ ,定义为：

$$a\sim p_{\mathrm{eq}}$$

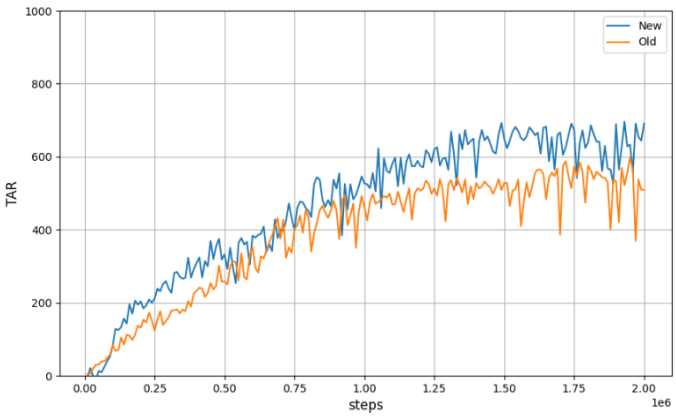
为了加速，直接选择top-k（20）个q值最高的a，作为正样本，使用langevin和随机动作构造负样本。

强化学习的本质：价值函数估计（更精确的估计方式，能量函数，转移模型构建）

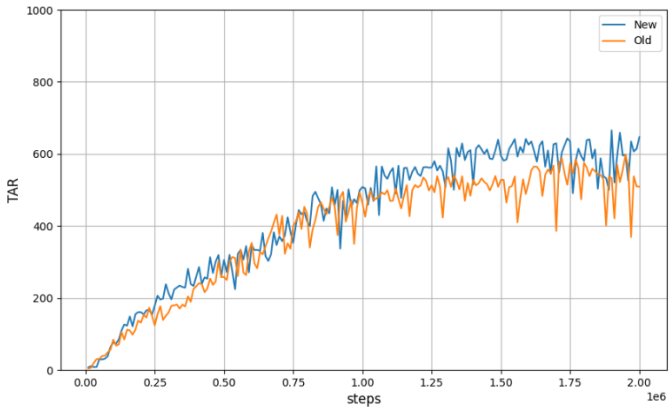
**General proof: 其他熵**

**Return Bound:**

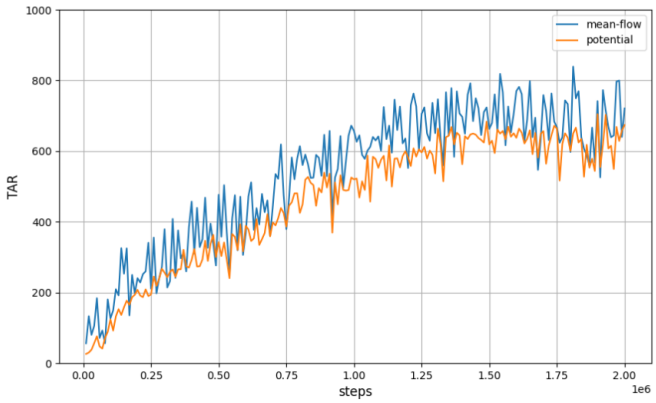
HMC dynamics 实验，LCD的补充 （Flow+短链langiven，先出效果）



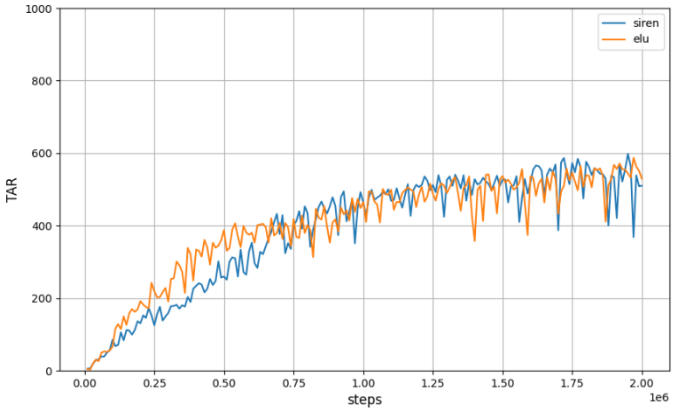

梯度裁剪改进+直接求导



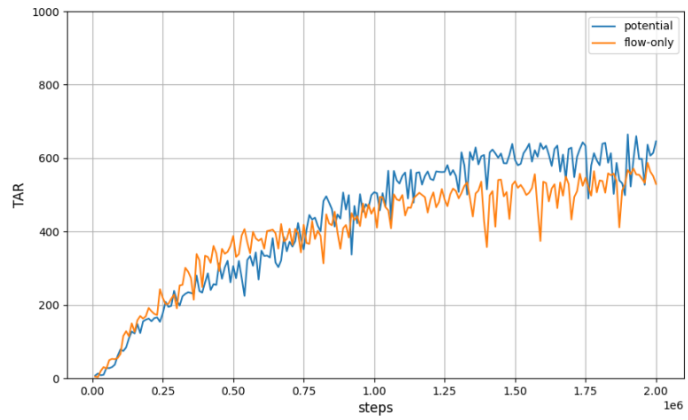

score function



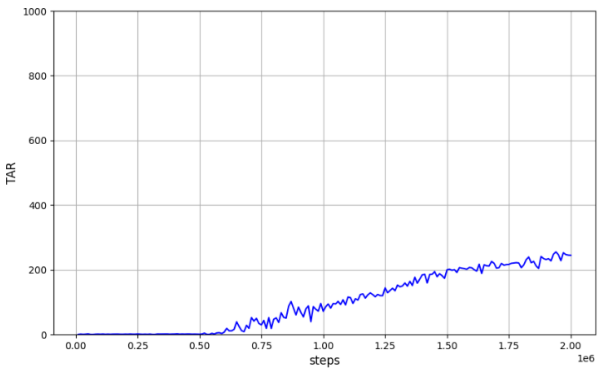

Mean flow



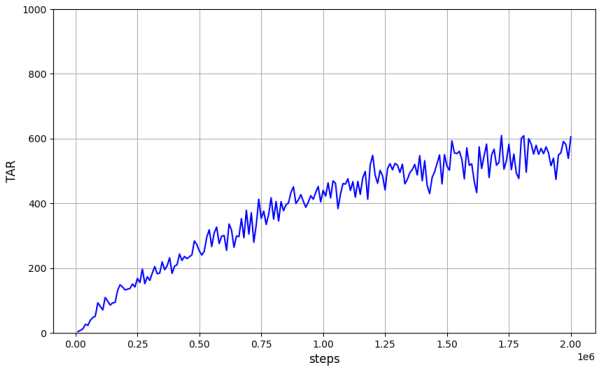

激活函数选择




组件选择



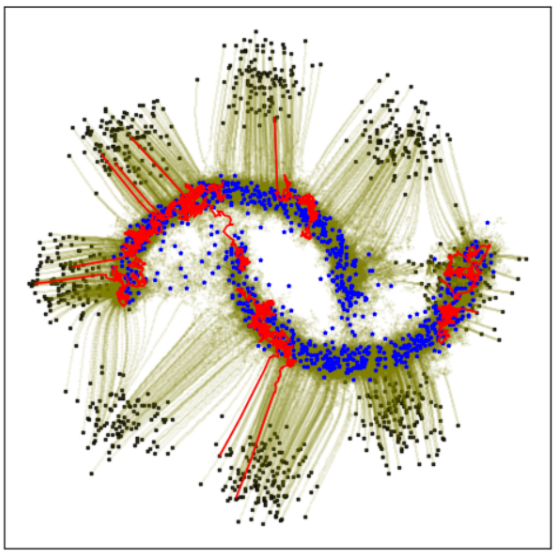

Without Q

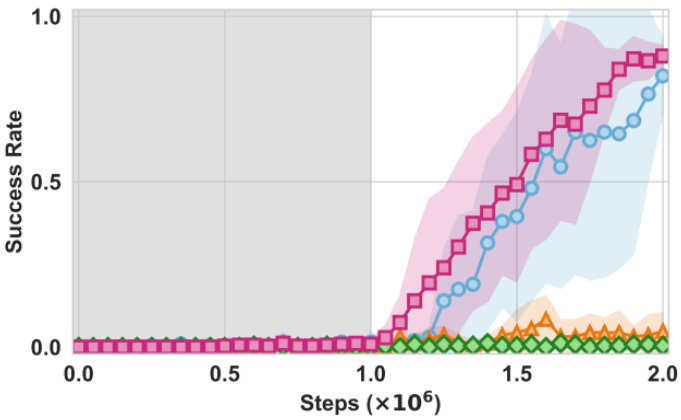



Baseline

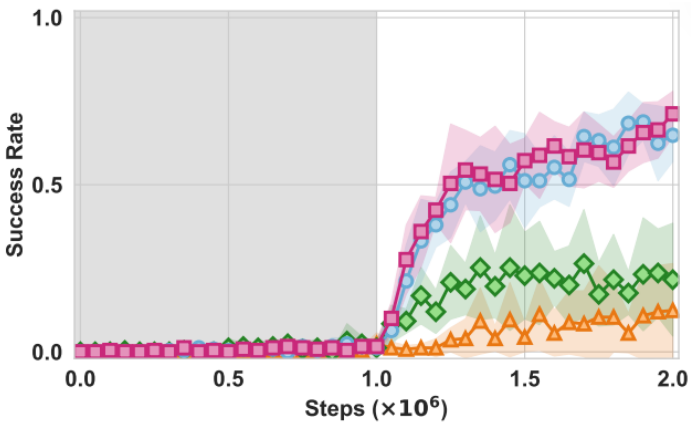
manipulation任务（专家轨迹克隆）

复杂分布建模能力（toy实验）

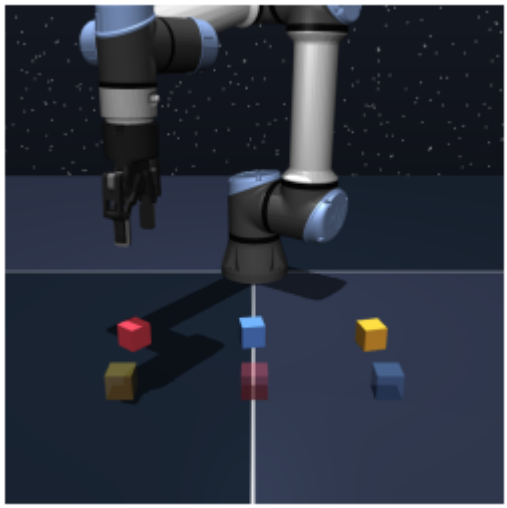
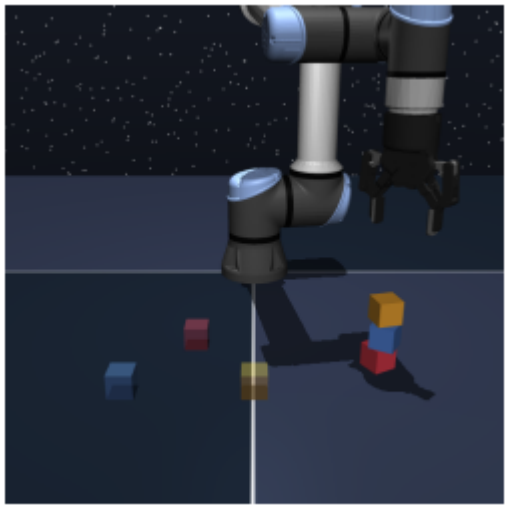





Cube-triple-task2




Cube-triple-task3

5.2 基于价值函数梯度的速度/梯度场构造

**真正理论完备的做法**

1.初始策略均匀分布化：令 $a_0 \sim U(0,1)$ 。第一步的时候熵为定值。梯度为零，此时有：

$$\frac{1}{\Delta t}(a - a') - \nabla_x Q_\theta(x) = 0.$$

此时可以认为，在初始分布为均匀分布的情况下，**此时策略的梯度刚好等于速度。在这种情况下，我们不用构造目标动作，也能构建此时的速度场。**

2.群体粒子优化：

针对同一个状态，批量生成batch化的粒子，在动作空间中优化

$$\mathcal{L}_{\text{vel}} = E_x[||\nabla V_\theta(a_t|s) - \nabla_a \text{min} Q_\phi(s,a)||_2^2].$$

以energy matching 为例：

$$a_1 = a_0 + \nabla_a Q_\phi(s, a_0)\Delta t, \quad a_0 \sim U(0,1)$$

$$a_{t+1} = a_t + \frac{\epsilon(t)}{2}\nabla_a Q_\phi(s, a_t) + \sqrt{\epsilon(t)} \cdot \mathcal{N}(0, I)$$

激活函数更换，平滑

## Schedule

Benchmark：DMC，GYM

Baseline：BRO, Simba，FlowRL，DIME

TODO: 端到端Qloss还是直接使用Qloss梯度？直接用这个sample的动作再加个maxQ的loss好，还是在采样过程中加个Q的梯度好

退化版本：纯模仿：

$$\mathcal{L}_{\text{OT}} = \mathbb{E}_{t\sim U(0,\tau^*)}^{x_{\text{data}}\in\mathcal{D}} \left[ \left\| \nabla_x V_\theta(x_t^{(i)}) + x_{\text{data}}^{(i)} - T(x_{\text{data}}^{(i)}) \right\|^2 \right].$$

$$\mathcal{L}_{\text{CD}} = \mathbb{E}_{x\sim p_{\text{data}}}\left[\frac{V_\theta(x)}{\varepsilon_{\max}}\right] - \mathbb{E}_{\tilde{x}\sim \text{sg}(p_{\text{eq}})}\left[\frac{V_\theta(\tilde{x})}{\varepsilon_{\max}}\right],$$

# Preliminaries

## 1.Energy Based Models (EBM)

### 定义：

EBM是一类通过标量能量函数 $E_\theta(x,y)$ 对样本分布建模的无监督模型，其基本思想是通过相容性对给定的 $(x,y)$ 进行评估，令低能量的 $(x,y)$ 对应高概率，从而学习数据的概率分布。定义两个变量x，y和能量函数 $E_\theta(x,y)$：当x和y相容时，能量函数取值较小；相反则取值较大：

$$p(y|x) \approx p_\theta(x,y) = \frac{\exp(-E_\theta(x,y))}{Z_\theta}, \quad Z_\theta = \int \exp(-E_\theta(x,y))dxdy$$

**与概率模型的关系：**概率模型是能量模型的一个特例。

> *能量可以看作没有正则化的负对数概率*

### 目标：

给定x，求得y的值使得能量函数尽可能小：

$$y = \text{argmin}E_\theta(x,y)$$

### 训练方式(以生成为例):

1. 对比散度（Contrastive Divergence, CD）

目标是最大化训练样本 $\{x_i\} \sim p_{\text{data}}$ 的 log-likelihood：

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log p_\theta(x_i) = \mathbb{E}_{x\sim p_{\text{data}}}[\log p_\theta(x)]$$

参数 $\theta$ 的梯度：

$$\nabla_\theta \log p_\theta(x) = \nabla_\theta\left(-E_\theta(x) - \log Z_\theta\right) = -\nabla_\theta E_\theta(x) - \nabla_\theta \log Z_\theta$$

$$\log Z_\theta = \log \int e^{-E_\theta(x')}dx' \quad \Rightarrow \quad \nabla_\theta \log Z_\theta = \frac{1}{Z_\theta}\nabla_\theta Z_\theta = \mathbb{E}_{x'\sim p_\theta}[\nabla_\theta E_\theta(x')]$$

最终梯度表达式为：

$$\nabla_\theta \log p_\theta(x) = -\nabla_\theta E_\theta(x) + \mathbb{E}_{x'\sim p_\theta}[\nabla_\theta E_\theta(x')]$$

$\mathbb{E}_{x'\sim p_\theta}[\nabla_\theta E_\theta(x')]$ 要求从模型分布 $p_\theta(x') \propto e^{-E_\theta(x')}$ 中采样 —— 但这个分布本身依赖 $E_\theta$，通常没有解析形式，必须用 **MCMC** 等方法近似采样。

由于 $x' \sim p_\theta$ 很难采样，CD 用从 $x \sim p_{\text{data}}$ 开始的 MCMC 链来近似这个模型分布采样。也就是：

- 用数据样本 x 生成一个伪样本 $\tilde{x} \approx x' \sim p_\theta$
- 用它来代替难算的模型期望：

$$\mathbb{E}_{x' \sim p_\theta}[\nabla_\theta E_\theta(x')] \approx \mathbb{E}_{x \sim p_{data}} \nabla_\theta E_\theta(\tilde{x})$$

因此，训练目标改为：

$$\nabla_\theta \mathcal{L}_{\mathrm{CD}}(\theta) = \mathbb{E}_{x \sim p_{\mathrm{data}}}[-\nabla_\theta E_\theta(x) + \nabla_\theta E_\theta(\tilde{x})], \quad x \sim T^k(x)$$

其中 $\tilde{x}$ 是从 x 通过运行 k 步 MCMC 得到的采样点，T为MCMC转移核，CD最终目标为

$$\mathrm{CD}(\theta) = \mathbb{E}_{x \sim p_{\mathrm{data}}}[E_\theta(x)] - \mathbb{E}_{\tilde{x} \sim T^k(x)}[E_\theta(\tilde{x})]$$

最小化这个差值，即希望真实样本能量低，模型样本能量高，形成对比。

### 2.评分函数对齐（Score Matching, SM）

**完全避免计算** $Z_\theta$ ，也不依赖采样。找到一个目标函数，通过真实数据训练模型。构造一个目标函数，使得模型的 score function 与数据的真实 score function **尽量接近：**

$$\mathcal{J}(\theta) = \mathbb{E}_{x \sim p_{\mathrm{data}}} \left[ \frac{1}{2} \| s_\theta(x) - s_{\mathrm{data}}(x) \|^2 \right]$$

不直接拟合 $p_\theta(x)$ ，而是拟合它的**分数函数，因为score function 就是能量函数关于 x 的负梯度**：

$$s_\theta := \nabla_x \log p_\theta(x) = -\nabla_x E_\theta(x)$$

SM的目标是构造一个目标函数，使得模型的 score function 与数据的真实 score function **尽量接近：**

$$\mathcal{J}(\theta) = \mathbb{E}_{x \sim p_{\mathrm{data}}} \left[ \frac{1}{2} \| s_\theta(x) - s_{\mathrm{data}}(x) \|^2 \right]$$

当 score function 来自某个密度函数的梯度时，可以构造一个完全依赖于模型和数据的目标函数，而不需要知道 $p_{\mathrm{data}}(x)$ ：

$$\mathcal{J}_{\mathrm{SM}}(\theta) = \mathbb{E}_{x \sim p_{\mathrm{data}}} \left[ \frac{1}{2} \| \nabla_x E_\theta(x) \|^2 - \nabla_x^2 E_\theta(x) \right]$$

局限性：

1. **需要二阶导数**：需计算 $\nabla_x^2 E_\theta(x)$ ，在高维空间和复杂模型（如CNN）中代价很高；
2. **仅适用于连续变量**：离散数据无法求导；
3. **对能量函数形式有要求**：需可导、最好是光滑函数。

### 3.噪声对比估计（Noise Contrastive Estimation，NCE）

NCE将密度估计问题转化为二分类问题 —— 让模型区分真实和噪声的数据，通过引入一个已知的噪声分布 $q(x)$ ，将密度估计转化为分类问题。

NCE通过构造一个**二分类任务实现密度估计。其中**正样本来自真实数据 $x \sim p_{\mathrm{data}}$ 。而负样本来自已知噪声分布 $x \sim q(x)$ 。对于每个正样本，配 k 个负样本；总样本数中，正样本比例为 $\frac{1}{1+k}$ ，负样

本比例为 $\frac{k}{1+k}$

学习一个二分类器 $D_\theta(x)$ 来估计给定 x 是来自真实分布的概率：

$$D_\theta(x) = \frac{p_\theta(x)}{p_\theta(x) + kq(x)} = \frac{\exp(-E_\theta(x))}{\exp(-E_\theta(x)) + kZ_\theta q(x)}$$

损失函数为：

$$\mathcal{L}_{\text{NCE}}(\theta) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D_\theta(x)] + k \cdot \mathbb{E}_{x\sim q(x)}[\log(1 - D_\theta(x))]$$

$$\mathcal{L}_{\text{NCE}}(\theta) = \sum_i \log \frac{\exp(-E_\theta(x_i))}{\exp(-E_\theta(x_i)) + kZ_\theta q(x_i)} + \sum_j \log \frac{kZ_\theta q(\tilde{x}_j)}{\exp(-E_\theta(x_j)) + kZ_\theta q(\tilde{x}_j)}$$

可将 $Z_\theta$ 视独立的变量进行优化，或者：

$$\log D_\theta(x) = -\log\left(1 + \frac{kZ_\theta q(x)}{f_\theta(x)}\right)$$

等价于logistic的二分类任务：

$$\text{logit}(x) = \log f_\theta(x) - \log q(x) - \log(kZ_\theta)$$

## 2.Reinforcement Learning with Max Entropy

考虑一个标准的**马尔可夫决策过程（Markov Decision Process, MDP）**，形式为五元组 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$，其中：

- $\mathcal{S}$ 为状态空间， $\mathcal{A}$ 为动作空间；
- $p(s'|s,a)$ 为环境的状态转移概率；
- $r(s,a)$ 为即时奖励函数；
- $\gamma \in (0,1)$ 为折扣因子，用于平衡长期与短期收益。

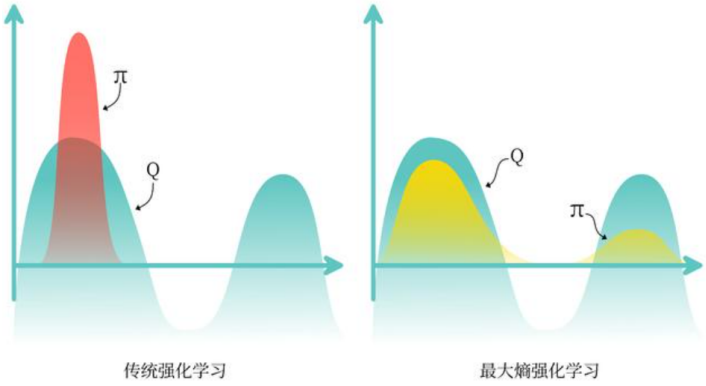
一个策略 $\pi(a|s)$ 定义了智能体在给定状态下采取动作的概率分布。经典强化学习的目标是寻找一个最优策略，最大化期望累积折扣回报：

$$J(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right].$$

在传统目标函数中引入策略熵项，用于增强策略的随机性以提高探索能力和鲁棒性。其优化目标变为：

$$J_{\text{MaxEnt}}(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t))\right)\right],$$

这一形式隐含地鼓励策略朝向一个与Q函数有关的**Boltzmann 分布逼近**，即高奖励动作概率更大，但整体仍保留一定的多样性和随机性。**最大熵强化学习的策略目标就是基于EBM的闭式解**




最大熵强化学习中的**策略最优解形式为 Q 函数诱导的 Boltzmann 分布**，与能量基模型（EBM）中的 Gibbs 分布具有一致的表达结构。虽然其训练目标并非直接基于能量模型的最大似然估计，但在已知 Q 函数的前提下，策略具有一个闭式的 EBM 形式解。因此，最大熵策略可以视为对能量模型的 KL 投影，其优化过程具备 EBM 解释。

**证明：**

定义目标分布（能量分布）为：

$$p^\star(a|s) = \frac{1}{Z(s)} \exp\left(\frac{Q(s,a)}{\alpha}\right),$$

最优策略 $\pi^\star(a|s)$ 是最小化 KL 散度的解：

$$\pi^\star(\cdot|s) = \arg\min_\pi \mathrm{KL}\left(\pi(\cdot|s) \parallel p^\star(\cdot|s)\right).$$

$$\begin{aligned}
\mathrm{KL}(\pi\|p^\star) &= \int \pi(a|s) \log \frac{\pi(a|s)}{p^\star(a|s)} da \\
&= \int \pi(a|s) \log \pi(a|s) da - \int \pi(a|s) \log p^\star(a|s) da \\
&= -\mathcal{H}(\pi(\cdot|s)) - \mathbb{E}_{a\sim\pi}\left[\log p^\star(a|s)\right] \\
&= -\mathcal{H}(\pi(\cdot|s)) - \mathbb{E}_{a\sim\pi}\left[\frac{1}{\alpha} Q(s,a) - \log Z(s)\right] \\
&= -\mathcal{H}(\pi(\cdot|s)) - \frac{1}{\alpha}\mathbb{E}_\pi[Q(s,a)] + \log Z(s).
\end{aligned}$$

由于 $\log Z(s)$ 与策略 $\pi$ 无关，优化问题等价于最大化下式：

$$\mathbb{E}_{a\sim\pi}\left[Q(s,a)\right] + \alpha \mathcal{H}(\pi(\cdot|s)),$$

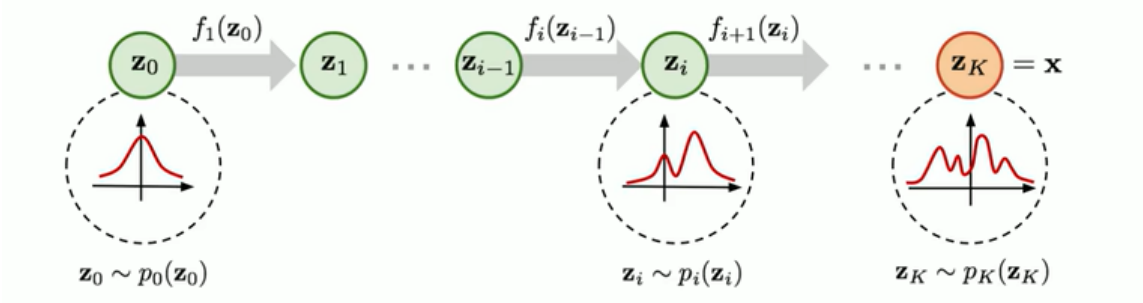
当 Q 函数已知，且动作空间 $\mathcal{A}$ 连续、策略无约束时，最大化上述目标函数的最优策略具有如下闭式解：

$$\pi^\star(a|s) = \frac{1}{Z(s)} \exp\left(\frac{Q(s,a)}{\alpha}\right)$$

这说明：**最大熵策略实质上是 Q 函数构造的 Boltzmann 分布**，温度系数 $\alpha$ 控制了策略的随机性，当 $\alpha \to 0$ 时策略趋于贪婪（argmax）。

## 3. Flow Matching

定义：一种 **基于常微分方程（ODE）的生成建模方法。**
核心思想：用一个**时间依赖的向量场** $v_t(x)$ 将一个容易采样的**源分布** $p_0(x_0)$ 逐渐变换成**目标分布** $p_1(x_1)$




## 1.基本定义

1. **源分布（base distribution）** $p_0(x_0)$

   通常选一个采样方便的分布，比如高斯分布、均匀分布等。

2. **目标分布（target distribution）** $p_1(x_1)$

   希望模型能生成的真实数据分布（例如图像、轨迹等）。

3. **概率路径（probability path）** $p_t(x_t)$

   定义了从 $t=0$ 到 $t=1$ 中粒子的演化过程，用于连接 $p_0$ 与 $p_1$。比如可以选择线性插值加噪声的路径，也可以设计更复杂的路径。在生成过程中， $p_t(x_t)$ 表示在 $t$ 时刻随机变量 $x_t$ **的分布**

4. **流函数 （flow）** $\phi_t(x_t)$

   是一个随时间变化的**确定性映射**，描述空间中**每个点**如何沿向量场进行移动，是一种描述从 $t=0$ 到 $t=1$ 粒子本身在空间中演化的方式（$[0,1]\times R_d \to R_d$）。给定一个起点 $x_0$，$\phi_t(x_0)$ 表示它在时间 t 的位置。

   $p_t(x_t)$ **可以由** $\phi_t(x_t)$ **推出，满足推前公式（Push-forward Equation）：**

$$p_t(x_t) = p_0(\phi_t^{-1}(x_t))\det|\frac{\delta\phi_t^{-1}}{\delta x_t}(x_t)|$$

5. **时间依赖向量场（time-dependent vector field）** $v_t(x_t)$

   表示在时刻 t 位置 $x_t$ 上的“流动方向”和“速度”。满足 ODE：

$$\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x)), \phi_0(x) \sim p_0\,,$$

   解这个方程得到 $\phi_1(x) \sim p_1$。

   **连续性方程（continuity equation）：并非每个** $v_t(x_t)$ **都能产生概率路径**。flow matching中，需要满足连续性方程：

$$\frac{d}{dt}p_t(x) + \nabla(p_t(x)v_t(x)) = 0$$

- 已知 $v_t$ → 如果它是一个光滑可逆的向量场，那么从任意初始分布 $p_0$ 出发，把它沿着 $v_t$ 推进，就一定能得到满足连续性方程的 $p_t$。

- 已知 $p_t$ → 如果 $p_t$ 真的来自某个确定性 ODE 流，那么一定存在一个 $v_t$ 让它们一起满足连续性方程。

**如果 $v_t$ 合法（光滑 + 无奇点 + 可逆），那么 $p_t$ 完全由 $p_0$ 和 $v_t$ 决定，并自动满足方程。**

## 2.训练目标

Flow matching的本质是**建模一个速度场** $v_t$ ，使得其能够生成 $p_1(x_1)$ 。因此初始训练目标为：

$$\mathbb{E}_{t\sim\mathcal{U}(0,1),x_t\sim p(x_t)}\left[\|u_t(x_t)-\underbrace{v_t(x_t)}_{\text{intractable}}\|_2^2\right],$$

直接拟合 $v_t(x_t)$ 很困难,于是 Flow Matching 用了一个关键技巧：**条件流匹配（CFM）**：

- 构造条件概率路径 $p_t(x_t|x_1)$
- 在已知 $x_1$ 的情况下，条件向量场 $v_t(x_t|x_1)$ 可以解析计算。
- 用这个条件向量场作为监督训练神经网络 $u_t(x_t|x_1)$ ，间接学到边缘向量场。

**条件：若边缘向量场存在，且由条件向量场生成的条件概率路径且满足连续性方程，则边缘向量场生成的边缘概率路径也满足连续性方程。**

$$\begin{aligned}\frac{d}{dt}p_t(x)&=\int(\frac{d}{dt}p_t(x|x_1))q(x_1)dx_1=-\int\text{div}(u_t(x|x_1)p_t(x|x_1))q(x_1)dx_1\\&=-\text{div}(\int(u_t(x|x_1)p_t(x|x_1))q(x_1)dx_1)=-\text{div}(p_t(x)u_t(x))\end{aligned}$$

由此，条件向量场训练目标变为：

$$\mathcal{L}_{\text{CFM}}(\theta)=\mathbb{E}_{t,q(x_1),p_t(x|x_1)}\left\|v_t(x)-u_t(x|x_1)\right\|^2$$

可以证明两个训练目标具有相同的梯度。

## 3.条件向量场的表达形式

前面的论述保证了条件向量场为训练目标下，能生成和边缘向量场一样的参数化神经网络。然而，条件向量场的形式却是未知的。因此，需要假定条件概率路径 $p_t$ 需满足的形式。

1.高斯条件概率路径

假设概率路径 $p_t$ 为高斯分布，则其满足：

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

此时对于随机变量 $x_t$ ,满足：

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

最终，条件向量场 $u_t(x_t|x_1)$ 的解析形式可以写为：

$$u_t(x|x_1)=\frac{\sigma_t'(x_1)}{\sigma_t(x_1)}(x-\mu_t(x_1))+\mu_t'(x_1).$$

2.方差爆炸形式（SM）

$$p_t(x|x_1) = \mathcal{N}(x|x_1, \sigma_{1-t}^2 I)$$
$$\mu_t(x_1) = x_1, \sigma_t(x_1) = \sigma_{1-t}$$
$$u_t(x|x_1) = -\frac{\sigma'_{1-t}}{\sigma_{1-t}}(x - x_1).$$

3.方差保留形式（DDPM）

$$p_t(x|x_1) = \mathcal{N}(x|\alpha_{1-t}x_1, (1 - \alpha_{1-t}^2)I)$$

### 4.生成过程（推断阶段

1. 从源分布 $p_0$ 采样 $x_0$ 。
2. 用数值 ODE 求解器（Euler、Heun、RK 等）沿着 $v_t$ 演化：

$$x_{t+\Delta t} = x_t + \Delta t \cdot v_t(x_t)$$

3. 得到 $x_1$ 作为生成样本。

# References:

## EBT：基于能量模型训练的大模型

> 核心观点：EBT 将预测任务视作输入-候选预测之间能量最小化的优化过程，借助梯度下降在推理阶段不断优化预测结果，具备动态计算、建模不确定性与预测验证等能力。

### 1.研究动机

当前“System 2 Thinking”方法存在以下局限：

- 模态/任务特定性强（只适用于数学、代码、语言等）
- 依赖监督或奖励信号（如RL中的verifier、reward model）
- 可扩展性差（如Diffusion模型需固定denoise步数）

是否可能通过纯无监督预训练方式，普适性地学习“系统2思维”？

### 2.方法概览

1. 能量函数建模（Verifier）

训练一个能量函数，评估“输入 `x` 与候选预测 `ŷ`”之间的匹配程度（能量越低，匹配越好，类比为“兼容度评估器”，而非直接预测器）。

2. 推理即优化

推理阶段从随机初始预测出发，不断进行梯度下降以最小化能量：

$$\hat{y}_{i+1} = \hat{y}_i - \alpha \nabla_{\hat{y}_i} E_\theta(x, \hat{y}_i)$$

将这个这个过程视为前向推理过程，每次前向传播都更新预测直到收敛。

**Algorithm 1:** Training

**Inputs:** Context $x$, Target $y$, EBM $E_\theta(x, \hat{y})$
**Hparams:** Steps $N$, Step Size $\alpha$, Loss $J(\cdot)$

Sample $\hat{y}_0 \sim \mathcal{N}(0, I)$;
**for** $i = 0, \ldots, N-1$ **do**
  $\hat{y}_{i+1} \leftarrow \hat{y}_i - \alpha \nabla_{\hat{y}_i} E_\theta(x, \hat{y}_i)$;
$\mathcal{L} \leftarrow J(\hat{y}_N, y)$;
**return** $\mathcal{L}$, *update* $E_\theta$;

**Algorithm 2:** Inference with Verification

**Inputs:** Context $x$, EBM $E_\theta(x, \hat{y})$
**Hparams:** Steps $N$, Step Size $\alpha$, Samples $M$

**for** $j = 1, \ldots, M$ **do**
  Sample $\hat{y}_{0,j} \sim \mathcal{N}(0, I)$;
  **for** $i = 0, \ldots, N-1$ **do**
    $\hat{y}_{i+1,j} \leftarrow \hat{y}_{i,j} - \alpha \nabla_{\hat{y}_{i,j}} E_\theta(x, \hat{y}_{i,j})$;
**return** $\hat{y}^* = \text{argmin}_j \, E_\theta(x, \hat{y}_{N,j})$;

| Facet | 说明 | 对应机制 |
| --- | --- | --- |
| 1. 动态计算资源分配 | 难题多思考，简单题少思考 | 步数可变的梯度下降 |
| 2. 不确定性建模 | 高能量 = 高不确定性 | 能量值本身反映置信度 |
| 3. 预测验证能力 | 判断当前预测是否可靠 | 能量函数即验证器 |

# EORM：基于能量模型训练的重排序器

> 核心观点：不再微调大语言模型，仅训练一个轻量级能量判别器 EORM，用“结果对/错”二值标签就能对同题多条 CoT 进行重排序，显著提升数学推理正确率，且计算、标注开销极低。

---

## 1 研究动机

CoT 采样昂贵:经典 *Self-Consistency* 依赖大量采样(≥128 条) 再投票，算力高昂。

过程监督成本高:Process-RM 需要逐步标注；RLHF 需偏好对，均耗时费力。

目标:只用“最后答案对/错”标签，在推断后重排序，高效挑出正确 CoT。

---

## 2 方法概览

a. 基座 LLM（冻结）一次性生成 k 条 CoT。

b. 轻量判别器 EORM 计算每条能量

c. 选能量最低者为最终答案，可早停。

d. 训练损失函数：

$$\mathcal{L}_G = \frac{1}{|P_G||N_G|} \sum_{i \in P_G} \sum_{j \in N_G} \log\big(1 + \exp(E_i - E_j)\big)$$

# EDLM：基于能量函数的扩散生成器

> 核心观点：EDLM 将能量模型引入离散扩散语言模型，有效缓解采样误差累积问题，实现了接近自回归模型的文本生成质量和更高的采样效率。

## 1 研究动机

核心问题：训练与采样分布不匹配。现有扩散模型在每个去噪步骤中尝试 并行预测所有被mask的token，但其建模方式将 联合分布简化为各token独立的条件分布，即：

$$p(x_0 \mid x_t) = \prod_i p(x_0^i \mid x_t)$$

这样忽略了 token 之间的序列依赖关系，导致以下两个直接后果：

1. 生成质量下降

- 并行采样时，无法捕捉上下文相关性，容易生成 语法错误或语义不连贯的句子；
- 误差在多步采样中逐步放大，尤其在 采样步数减少时表现更差。

2. 难以高效采样

- 为了弥补token独立建模的劣势，往往需要 更多的去噪步骤（>1000步） 才能得到可用文本；
- 导致采样速度慢、效率低，无法与自回归模型在实用场景中竞争。

3. 无法使用标准的似然评估

- 由于模型分布非归一化，很难精确计算其他标准指标；
- 导致评估结果不一致，难以与主流语言模型公平比较。

---

## 2.方法概览：

> 核心创新是 在每个扩散去噪步骤上引入序列级的能量模型（EBM） 来纠正采样误差

1. 残差能量建模（Residual Energy Form）：在每个步骤t的去噪分布建模为：

$$p_{\text{EDLM}}(x_{t-1} \mid x_t) = p_{\text{diffusion}}(x_{t-1} \mid x_t) \cdot \exp\left(-E(x_{t-1}, x_t)\right)$$

2. 两种能量函数实现方式：

- （a）基于预训练自回归语言模型（AR）构建能量函数（EDLM-AR）：

$$E(x_0, x_t) = -\log p_{\text{AR}}(x_0) + \log p_{\text{diff}}(x_0 \mid x_t) + C$$

- （b）通过噪声对比估计（Noise Contrastive Estimation, NCE）训练能量函数（EDLM-NCE）：

$$\mathcal{L}_{\mathrm{NCE}} = \mathbb{E}\left[\log \frac{1}{1+\exp(E(x^+, x_t))} + \log \frac{1}{1+\exp(-E(x^-, x_t))}\right]$$

# CrystalFormer：基于RL的分子材料生成器

> 核心观点：通过大语言模型的 RLHF 提升能力——将奖励驱动微调迁入材料设计

## 1.研究动机

元素组合和原子排布的可能性极大，传统 DFT 扫描耗时高。以往机器学习只能“评分”候选晶体，缺乏主动生成。

（RL）

## 2.方法概览

整体框架 = 预训练 CrystalFormer 生成模型 + 奖励模型（稳定性或性能） + PPO 强化微调。

目标函数：最大化 $r(x) - \tau \cdot KL(p_\theta \parallel p_{base})$ ，在追求奖励的同时保持与先验的一致性。

**模型结构**

· 输入编码：<空间群>-<Wyckoff>-<元素>-<分数坐标>-<晶格参数> 序列。

· Backbone：16-层自回归 Transformer，d_model=64，RoPE 位置编码。

· 输出头：离散 token 用 softmax；连续坐标/晶格参数用混合分布 (von-Mises / Gaussian)。

· 参数量：4.8 M

**奖励模型与任务设定**

任务 A – 稳定性： $r(x) = \text{-}E_{hull}(x)$ ，由 Orb-v2 ML 势能 + 凸包计算得到。

任务 B – 介电×带隙 FoM： $r(x) = \varepsilon_{tot} \cdot E_g$ ，带隙用 MEGNet，介电常数用 M3GNet。

**算法设计**

· 策略区分：p_base = 预训练权重 (冻结)；p_θ = 当前策略 (持续更新)。

· 使用 PPO：clip ε=0.2，batch 1000，内循环 3 步，梯度裁剪=1。

· 优势函数：A = [r(x) − τ·(log pθ − log p_base)] − EMA_baseline。

# KLE: 语义不确定性量化方法

> 核心观点：现有方法大多依赖于 token-level 概率（例如 token likelihood）或将语义聚类视为“硬等价类”，难以捕捉生成之间的“语义相似性”。该方法定义基于核函数的“语义核”（semantic kernel），并通过其 von Neumann 熵来刻画输出的不确定性。

## 1.研究动机

在 **Semantic Entropy（SE）** 中，两个回答要么“语义等价”被聚为同一簇，要么完全不相关，被认为是不同簇。这是**硬划分（hard clustering）** —— 信息损失严重，无法刻画“相近但不等价”的回答。

## 2.方法概览

1.使用DeBERTa-Large-MNLI模型构造语义对 $\{S_i, S_j\}$的边权 $W_{ij}$：

$$W_{ij} = f(\mathrm{NLI}_{ij}, \mathrm{NLI}_{ji}) = w^\top(\mathrm{NLI}_{ij} + \mathrm{NLI}_{ji})$$

2.构造图拉普拉斯矩阵 L = D - W编码图结构、**引入图结构上的几何信息**，进一步构造出核函数：

$$L = D - W$$

其中：

$$D_{ii} = \sum_j W_{ij}$$

3.使用核函数构造语义核矩阵K：

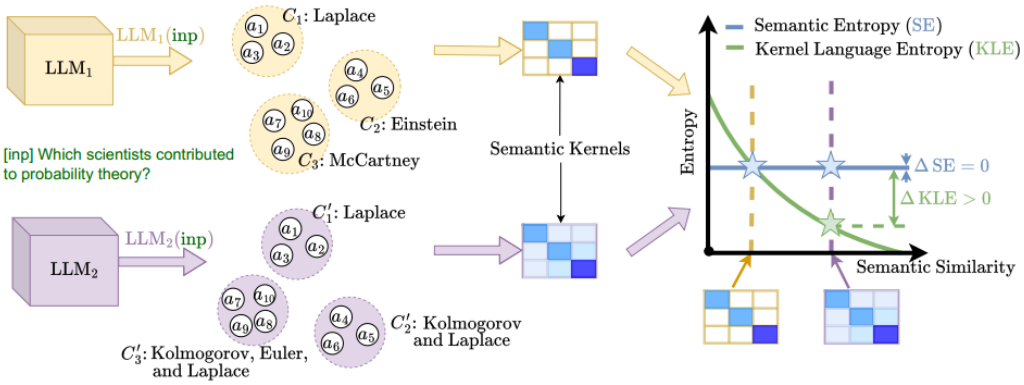
$$K_t = \exp(-tL)$$

$$K_{\nu,\kappa} = \left(\frac{2\nu}{\kappa^2} I + L\right)^{-\nu}$$

$$K_{\mathrm{sem}} = \alpha K_{\mathrm{heat}} + (1-\alpha) K_{\mathrm{SE}}$$

4.计算VNE：

$$\mathrm{KLE}(x) = \mathrm{VNE}(K_{\mathrm{sem}}) = -\operatorname{Tr}(K_{\mathrm{sem}} \log K_{\mathrm{sem}})$$




与聚类相比，该方法能够联系起不同cluster之间的语义关系，使得其能够随着总体语义关系的相近而降低熵

# Energy matching:

## 1. 核心思想

**Energy Matching** 将 **流匹配（Flow Matching/OT-CFM）** 和 **能量模型（EBM）** 统一到一个**时间无关的标量势函数** $V_\theta(x)$ 中：

- 在远离数据流形时 $\varepsilon \approx 0$，沿最优传输（OT）无旋速度场从噪声快速运输到数据。
- 在接近数据流形时 $\varepsilon \approx \varepsilon_{\max}$，切换为朗之万扩散，在势阱中达到玻尔兹曼分布。

同一个 $V_\theta$ 既是**速度场的势函数**，也是**EBM 能量函数**。

---

## 2. 理论推导与符号

方法基于 **Jordan–Kinderlehrer–Otto (JKO)** 变分形式：

$$\rho_{t+\Delta t} = \arg\min_{\rho} \frac{W_2^2(\rho, \rho_t)}{2\Delta t} + \int V_\theta(x) d\rho(x) + \varepsilon(t) \int \rho(x) \log \rho(x) dx$$

一阶条件：

$$\frac{1}{\Delta t}(x - y) + \nabla_x V_\theta(x) + \varepsilon(t) \nabla_x \log \rho_t(x) = 0, \quad (x, y) \in \mathrm{supp}(\gamma_t)$$

**解释**

- x：当前时刻的样本位置（沿 OT 测地线的插值点）。

- y：OT 配对的另一端点（起点或终点样本）。在从噪声到数据的运输中，$y$ 可以理解为起点噪声样本，而 $x$ 是运输过程中某一时刻的位置。

当 $\varepsilon(t) = 0$（运输阶段）：

$$\frac{1}{\Delta t}(x - y) + \nabla V_\theta(x) = 0$$

当 $t$ 很大（平衡阶段）：

$$\rho_{\mathrm{eq}}(x) \propto e^{-V_\theta(x)/\varepsilon_{\max}}$$

---

## 3. 双目标训练策略

### 3.1 流式目标（Flow-like Objective, $L_{\mathrm{OT}}$）

- 解最优传输（OT）映射 T，配对噪声样本和数据样本。
- 沿 Wasserstein 测地线插值 $x_t$ 。
- 用速度匹配：

$$L_{\mathrm{OT}} = \mathbb{E}\|\nabla V_\theta(x_t) + (x_{\mathrm{data}} - T(x_{\mathrm{data}}))\|^2$$

该目标学习到**时间无关**的速度场。

### 3.2 对比散度目标（Contrastive Divergence, $L_{\mathrm{CD}}$ ）

- 在数据流形附近用 EBM 方式微调：

$$L_{\mathrm{CD}} = \mathbb{E}_{x\sim p_{\mathrm{data}}} \frac{V_\theta(x)}{\varepsilon_{\max}} - \mathbb{E}_{\tilde{x}\sim p_{\mathrm{eq}}} \frac{V_\theta(\tilde{x})}{\varepsilon_{\max}}$$

- 负样本 $\tilde{x}$ 用朗之万动力学生成，初始化一半来自数据，一半来自噪声。

### 3.3 总损失

$$L = L_{\mathrm{OT}} + \lambda_{\mathrm{CD}} L_{\mathrm{CD}}$$

温度 $\varepsilon(t)$ 随时间分段线性从 0 升至 $\varepsilon_{\max}$ 。

## 4. 采样

- **无条件采样**：从高斯出发，按朗之万动力学推进到时间 $\tau_s$ 。

- **条件采样**：加入观测保真项，与 $V_\theta$ 相加形成后验势能，在其上运行朗之万。

# Meeting Record:

## 20250722

**key points of discussion:**

key point 1: xxx

key point 2: xxx

key point 3: xxx

**next step:**

1.RP方向讨论

a. 已经做的东西(概括)

b. Existing work

c. Future work（LLM+RL（reasoning enhancement: llm, ai for science））

2.PGAC讨论

3.语言确定