CEP

Title	Contrastive Energy Prediction for Exact Energy-Guided Diffusion Sampling in Offline Reinforcement Learning
日期	2023.05
发表单位	THU
github	thu-ml/CEP-energy-guided-diffusion: Official codebase for Exact Energy-Guided Diffusion Sampling via Contrastive Energy Prediction
上次编辑	@2025年4月19日 19:50
状态	Done
重要程度	⭐⭐

Exact Energy-Guided Sampling

目标是从下面的概率分布中采样

$$p_0(x_0) \propto q_0(x_0) e^{-\beta \mathcal{E}(x_0)}$$

最大的问题是中间阶段采样的Energy Guidance Term如何计算和估计

Exact Formulation of Intermediate Energy Guidance

Theorem 3.1 (Intermediate Energy Guidance). *Suppose* q_0 *and* p_0 *are defined as in Eq. (7). For* $t \in (0, T]$, *let* 表示加噪

$$p_{t0}(\boldsymbol{x}_t|\boldsymbol{x}_0) := q_{t0}(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t|\alpha_t \boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I}). \quad (8)$$

Denote $q_t(\boldsymbol{x}_t) := \int q_{t0}(\boldsymbol{x}_t|\boldsymbol{x}_0) q_0(\boldsymbol{x}_0) \mathrm{d}\boldsymbol{x}_0$ *and* $p_t(\boldsymbol{x}_t) := \int p_{t0}(\boldsymbol{x}_t|\boldsymbol{x}_0) p_0(\boldsymbol{x}_0) \mathrm{d}\boldsymbol{x}_0$ *as the marginal distributions at time* t, *and define*

$$\mathcal{E}_t(\boldsymbol{x}_t) := \begin{cases} \beta \mathcal{E}(\boldsymbol{x}_0), & t = 0, \\ -\log \mathbb{E}_{q_{0t}(\boldsymbol{x}_0|\boldsymbol{x}_t)} \left[e^{-\beta \mathcal{E}(\boldsymbol{x}_0)} \right], & t > 0. \end{cases} \quad (9)$$

Then q_t *and* p_t *satisfy*

$$p_t(\boldsymbol{x}_t) \propto q_t(\boldsymbol{x}_t) e^{-\mathcal{E}_t(\boldsymbol{x}_t)}, \quad (10)$$

and their score functions satisfy

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) = \underbrace{\nabla_{\boldsymbol{x}_t} \log q_t(\boldsymbol{x}_t)}_{\approx -\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)/\sigma_t} - \underbrace{\nabla_{\boldsymbol{x}_t} \mathcal{E}_t(\boldsymbol{x}_t)}_{\substack{\textit{energy guidance} \\ \textit{(intractable)}}}. \quad (11)$$

只要知道(11)式，就可以做到p_0中采样

(11)前一项已由训练好的DPM估计得到，只需要估计后一项，称为Intermediate Energy Guidance

Learning Energy Guidance by Contrastive Energy Prediction

分别从q_0和$\mathcal{N}(0, \mathbf{I})$采样$K$个独立样本，再从$[0, T]$均匀采样$t$，$f_\phi(\cdot, t)$为Intermediate Energy $\mathcal{E}_t$的估计网络

$$\min_{\phi} \mathbb{E}_{p(t)}\mathbb{E}_{q_0(\boldsymbol{x}_0^{(1:K)})}\mathbb{E}_{p(\boldsymbol{\epsilon}^{(1:K)})}\Bigg[-\sum_{i=1}^{K} \underbrace{e^{-\beta\mathcal{E}(\boldsymbol{x}_0^{(i)})}}_{\text{soft energy label}} \log \underbrace{\frac{e^{-f_\phi(\boldsymbol{x}_t^{(i)},t)}}{\sum_{j=1}^{K} e^{-f_\phi(\boldsymbol{x}_t^{(j)},t)}}}_{\text{predicted label}} \Bigg]. \tag{12}$$

Theorem 3.2. *Given unlimited model capacity and data samples, For all $K > 1$ and $t \in [0, T]$, the optimal f_{ϕ^*} in problem (12) satisfies $\nabla_{\boldsymbol{x}_t} f_{\phi^*}(\boldsymbol{x}_t, t) = \nabla_{\boldsymbol{x}_t}\mathcal{E}_t(\boldsymbol{x}_t)$.*

直观上，为了时得f_ϕ的梯度与$\mathcal{E}_t$相等，只需要两者为正比关系即可，因此可以通过学习K个样本的相对能量大小来实现

对Energy Label添加正则化来增加数值稳定性

$$\min_{\phi} \mathbb{E}_{p(t)}\mathbb{E}_{q_0(\boldsymbol{x}_0^{(1:K)})}\mathbb{E}_{p(\boldsymbol{\epsilon}^{(1:K)})}\Bigg[-\sum_{i=1}^{K} \underbrace{\frac{e^{-\beta\mathcal{E}(\boldsymbol{x}_0^{(i)})}}{\sum_{j=1}^{K} e^{-\beta\mathcal{E}(\boldsymbol{x}_0^{(j)})}}}_{\text{self-normalized energy label}} \log \underbrace{\frac{e^{-f_\phi(\boldsymbol{x}_t^{(i)},t)}}{\sum_{j=1}^{K} e^{-f_\phi(\boldsymbol{x}_t^{(j)},t)}}}_{\text{predicted label}} \Bigg]. \tag{13}$$

Comparison with Previous Methods for Guided Sampling

Table 1. Comparison between energy-guided sampling algorithms.

Method	Optimal Solution of Energy	Exact Guidance
CEP (ours)	$-\log \mathbb{E}_{q_{0t}(\boldsymbol{x}_0\|\boldsymbol{x}_t)}\left[e^{-\mathcal{E}_0(\boldsymbol{x}_0)}\right]$	✓
MSE	$\mathbb{E}_{q_{0t}(\boldsymbol{x}_0\|\boldsymbol{x}_t)}[\mathcal{E}_0(\boldsymbol{x}_0)]$	✗
DPS	$\mathcal{E}_0\left(\mathbb{E}_{q_{0t}(\boldsymbol{x}_0\|\boldsymbol{x}_t)}[\boldsymbol{x}_0]\right)$	✗

Previous Energy-Guided Samplers are Inexact

- **MSE for Predicting Energy**

  Loss定义为

$$\min_{\phi} \mathbb{E}_{q_{0t}(\boldsymbol{x}_0,\boldsymbol{x}_t)}\left[\|f_\phi(\boldsymbol{x}_t,t)-\mathcal{E}_0(\boldsymbol{x}_0)\|_2^2\right]$$

该Loss的最优解为

$$f_\phi^{\mathrm{MSE}}(\boldsymbol{x}_t,t)=\mathbb{E}_{q_{0t}(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\mathcal{E}_0(\boldsymbol{x}_0)\right]$$

与真正的Intermediate Energy不相等

- **Diffusion Posterior Sampling**

  Train-free，复用Data Prediction Formulation

$$\mathbb{E}_{q_{0t}(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\boldsymbol{x}_0\right]\approx\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t):=\frac{\boldsymbol{x}_t-\sigma_t\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)}{\alpha_t}$$

  于是Intermediate Energy Function可由下式估计

$$f_\theta^{\mathrm{DPS}}(x_t,t):=\mathcal{E}_0\left(\hat{x}_\theta(x_t,t)\right)\approx\mathcal{E}_0\left(\mathbb{E}_{q_{0t}(x_0|x_t)}\left[x_0\right]\right)$$

  此即Diffusion Planner采用的方案

Relationship with Contrastive Learning and Classifier Guidance

这里取$\mathcal{E}_0(x_0)=-\log q_0(c|x_0),\beta=1$，则有

$$p_0(\boldsymbol{x}_0)\propto q_0(\boldsymbol{x}_0)q(c|\boldsymbol{x}_0)\propto q(\boldsymbol{x}_0|c)$$

- **Contrastive Learning**

  可以证明，此时(12)等价于下式

$$\mathbb{E}_{t,\boldsymbol{\epsilon}^{(1:K)}}\mathbb{E}_{\prod_{i=1}^K q_0(\boldsymbol{x}_0^{(i)},c^{(i)})}\left[-\sum_{i=1}^K\log\frac{e^{-f_\phi(\boldsymbol{x}_t^{(i)},c^{(i)},t)}}{\sum_{j=1}^K e^{-f_\phi(\boldsymbol{x}_t^{(j)},c^{(i)},t)}}\right]$$

  此即Contrastive Learning的Loss(注意分母中对所有正负例取和)

  GLIDE即用上式训练一个CLIP并用其梯度来指导Text2Image生成

- **Classifier Guidance**

  若Label c是离散值，一个替代的Conditional Generation方法是Classifier Guidance，其训练分类器如下

$$\mathbb{E}_{t,\boldsymbol{\epsilon}^{(1:K)}}\mathbb{E}_{\prod_{i=1}^{K}q_0(\boldsymbol{x}_0^{(i)},c^{(i)})}\left[-\sum_{i=1}^{K}\log\frac{e^{-f_\phi(\boldsymbol{x}_t^{(i)},c^{(i)},t)}}{\sum_{j=1}^{M}e^{-f_\phi(\boldsymbol{x}_t^{(i)},c^{(j)},t)}}\right]$$

此法无法推广至连续型的Label或Energy-Guided的情形，因此CEP更加一般

Q-Guided Policy Optimization for Offline Reinforcement Learning

Offline RL被表述为受(软)约束策略优化问题

$$\max_\pi \mathbb{E}_{\mathbf{s}\sim D^\mu}\left[\mathbb{E}_{\mathbf{a}\sim\pi(\cdot|\mathbf{s})}Q_\psi(\mathbf{s},\mathbf{a})-\frac{1}{\beta}D_{\mathrm{KL}}\left(\pi(\cdot|\mathbf{s})\|\mu(\cdot|\mathbf{s})\right)\right]$$

其中$\mu(\cdot|s)$表示Behavior Policy，Q_ψ为π的Q-Function的估计模型，可以证明，最优策略π^*满足

$$\pi^*(\mathbf{a}|\mathbf{s})\propto\mu(\mathbf{a}|\mathbf{s})\,e^{\beta Q_\psi(\mathbf{s},\mathbf{a})}$$

Problem Formulation

Behavior Policy $\mu(\cdot|s)$为Diffusion Model，$\mathcal{E}_0(s,a)=-\beta Q_\psi(\mathbf{s},\mathbf{a})$为Guidance Energy，**目标是从$\pi^*$中采样Action**

记π_t,μ_t,a_t分别为加噪后的$\pi^*=\pi_0,\mu=\mu_0,a_0$，则我们需要估计的Score Function为

$$\nabla_{a_t}\log\pi_t(a_t|s)=\underbrace{\nabla_{a_t}\log\mu_t(a_t|s)}_{\approx-\epsilon_\theta(a_t|s,t)/\sigma_t}+\nabla_{a_t}\underbrace{\mathcal{E}_t(s,a_t)}_{\approx f_\phi(s,a_t,t)}$$

因此，需要训练3个Models：

1. A state-conditioned diffusion model $\epsilon_\theta(a_t|s,t)/\sigma_t$ to model the behavior policy $\mu(\cdot|s)$
   相当于在数据集上用Diffusion Model作模仿学习
2. An action evaluation model $Q_\psi(s,a)$ to define the intermediate energy function $\mathcal{E}_0$
3. An energy model $f_\phi(s,a_t,t)$ to estimate $\mathcal{E}_t(s,a_t)$

In-Support Contrastive Energy Prediction

假设已有动作值函数的估计$Q_\psi(s,a)$，则f_ϕ训练的CEP Loss为

多了对状态s采样

$$\min_{\phi} \mathbb{E}_{p(t)}\mathbb{E}_{\mu(\boldsymbol{s})}\mathbb{E}_{\prod_{i=1}^{K}\mu(\boldsymbol{a}^{(i)}|\boldsymbol{s})p(\boldsymbol{\epsilon}^{(i)})}\Bigg[-\sum_{i=1}^{K}\frac{e^{\beta Q_\psi(\boldsymbol{s},\boldsymbol{a}^{(i)})}}{\sum_{j=1}^{K}e^{\beta Q_\psi(\boldsymbol{s},\boldsymbol{a}^{(j)})}}\log\frac{e^{f_\phi(\boldsymbol{s},\boldsymbol{a}_t^{(i)},t)}}{\sum_{j=1}^{K}e^{f_\phi(\boldsymbol{s},\boldsymbol{a}_t^{(j)},t)}}\Bigg], \tag{19}$$

where $t\sim\mathcal{U}(0,T)$, $\boldsymbol{a}_t=\alpha_t\boldsymbol{a}+\sigma_t\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I})$.

但是$\mu(a|s)$并不能直接从数据集中获取

estimate the objective in problem (19). This is because we require $K>1$ independent action samples from $\mu(\boldsymbol{a}|\boldsymbol{s})$ for a single $\boldsymbol{s}$ for contrastive learning, whereas we only have one such action in $\mathcal{D}^\mu$ given that $\boldsymbol{s}$ is a continuous variable.

为此，利用训练好的Behavior Model $\mu_\theta(\cdot|\boldsymbol{s})$，对数据集里的每个状态$\boldsymbol{s}$采样$K$个Actions $\{\hat{a}^{(i)}\}_K$构成$\mathcal{D}^{\mu_\theta}$

于是(19)可由下式估计(MC方法)

$$\min_{\phi}\mathbb{E}_{t,\boldsymbol{s},\boldsymbol{\epsilon}} -\sum_{i=1}^{K}\frac{e^{\beta Q_\psi(\boldsymbol{s},\hat{\boldsymbol{a}}^{(i)})}}{\sum_{j=1}^{K}e^{\beta Q_\psi(\boldsymbol{s},\hat{\boldsymbol{a}}^{(j)})}}\log\frac{e^{f_\phi(\boldsymbol{s},\hat{\boldsymbol{a}}_t^{(i)},t)}}{\sum_{j=1}^{K}e^{f_\phi(\boldsymbol{s},\hat{\boldsymbol{a}}_t^{(j)},t)}} \tag{20}$$

In-support Softmax Q-Learning

下面描述如何估计动作值函数$Q_\psi\approx Q^\pi$

一般来说，用TD(1)(即SARSA)来训练

$$\mathcal{T}^\pi Q_\psi(\boldsymbol{s},\boldsymbol{a})=r(\boldsymbol{s},\boldsymbol{a})+\gamma\mathbb{E}_{\boldsymbol{s}'\sim P(\cdot|\boldsymbol{s},\boldsymbol{a}),\boldsymbol{a}'\sim\pi(\cdot|\boldsymbol{s}')}Q_\psi(\boldsymbol{s}',\boldsymbol{a}'). \tag{21}$$

但其需要在训练过程中对π采样，过于耗时，因此利用$\mathcal{D}^{\mu_\theta}$和Importance Sampling来估计

$$
\begin{aligned}
\mathbb{E}_{a\sim\pi}[Q(a)] &= \mathbb{E}_{a\sim\mu}\left[\frac{\pi(a)}{\mu(a)}Q(a)\right] \\
&\approx \frac{\sum_a \frac{\pi(a)}{\mu(a)}Q(a)}{\sum_a \frac{\pi(a)}{\mu(a)}}
\end{aligned}
$$

其中

$$
\frac{\pi(a)}{\mu(a)} = e^{\beta Q_\psi(a)}
$$

注：以上省略状态s以保持简洁

因此，最终的TD(1) Target的估计式为

$$
\mathcal{T}^\pi Q_\psi(\boldsymbol{s},\boldsymbol{a}) \approx r(\boldsymbol{s},\boldsymbol{a}) + \gamma \frac{\sum_{\hat{\mathbf{a}}'} e^{\beta_Q Q_\psi(\boldsymbol{s}',\hat{\boldsymbol{a}}')} Q_\psi(\boldsymbol{s}',\hat{\boldsymbol{a}}')}{\sum_{\hat{\mathbf{a}}'} e^{\beta_Q Q_\psi(\boldsymbol{s}',\hat{\boldsymbol{a}}')}}. \tag{22}
$$

Summary

1. 从数据集中模仿学习策略μ_θ
2. 生成In-Support Dataset $\mathcal{D}^{\mu_\theta}$
3. 利用In-Support SARSA训练Q_ψ
4. 利用In-Support CEP训练f_ϕ
5. 从π^*中采样