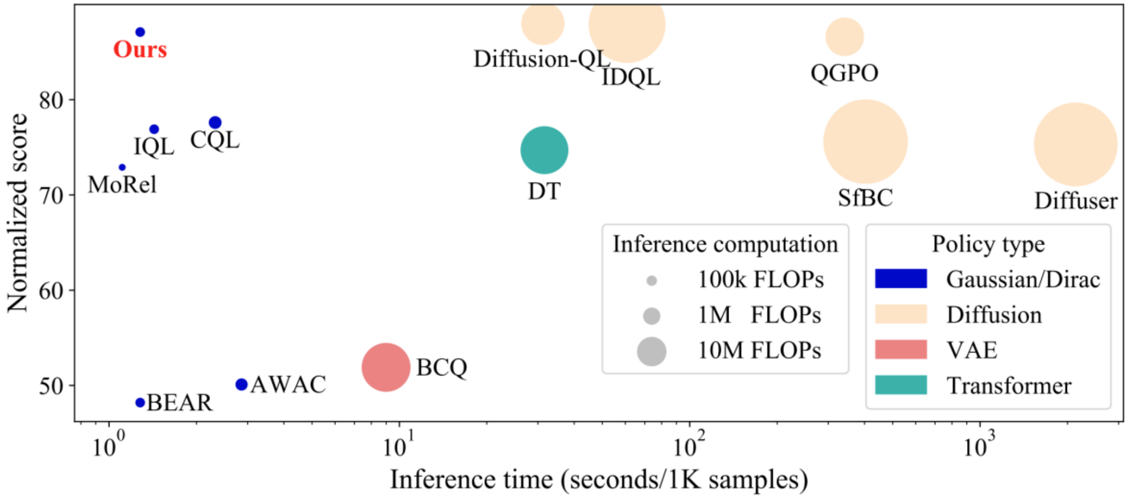
Score Regularized Policy Optimization through Diffusion Behavior

Huayu Chen, Cheng Lu, Zhengyi Wang, Hang Su, Jun Zhu
Tsinghua University

Code are available: https://github.com/thu-ml/SRPO

Overview

- We propose an offline RL algorithm for continuous control tasks.
- We leverage a powerful diffusion behavior model to regularize policy training.
- We entirely avoid iterative sampling from diffusion models in both during training and evaluation. This greatly increases computational efficiency



Key idea: diverse diffusion behavior but simple Dirac inference policy

Background

$$\min_\theta \mathbb{E}_{\boldsymbol{s}\sim\mathcal{D}^\mu} \underbrace{D_{\mathrm{KL}}\left[\pi^*(\cdot|\boldsymbol{s})||\pi_\theta(\cdot|\boldsymbol{s})\right]}_{\textit{Forward KL}} \Leftrightarrow \max_\theta \underbrace{\mathbb{E}_{(\boldsymbol{s},\boldsymbol{a})\sim\mathcal{D}^\mu}\left[\frac{1}{Z(\boldsymbol{s})}\log \pi_\theta(\boldsymbol{a}|\boldsymbol{s})\, e^{\beta Q_\phi(\boldsymbol{s},\boldsymbol{a})}\right]}_{\textit{Weighted Regression}}, \quad (2)$$

$$\updownarrow$$

$$\min_\theta \mathbb{E}_{\boldsymbol{s}\sim\mathcal{D}^\mu} \underbrace{D_{\mathrm{KL}}\left[\pi_\theta(\cdot|\boldsymbol{s})||\pi^*(\cdot|\boldsymbol{s})\right]}_{\textit{Reverse KL}} \Leftrightarrow \boxed{\underbrace{\max_\theta \mathbb{E}_{\boldsymbol{s}\sim\mathcal{D}^\mu,\boldsymbol{a}\sim\pi_\theta} Q_\phi(\boldsymbol{s},\boldsymbol{a}) - \frac{1}{\beta} D_{\mathrm{KL}}\left[\pi_\theta(\cdot|\boldsymbol{s})||\mu(\cdot|\boldsymbol{s})\right]}_{\textit{Behavior-Regularized Policy Optimization}}}. \quad (3)$$

Our initial loss function

Method Derivation

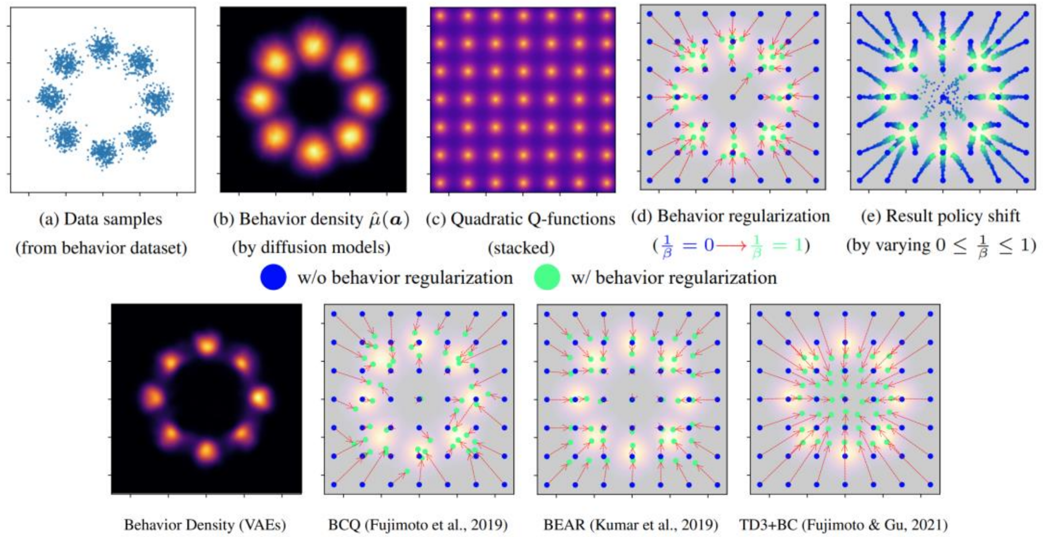
Decomposing the KL term:

$$\mathcal{L}_\pi(\theta) = \underbrace{\mathbb{E}_{\boldsymbol{s}\sim\mathcal{D}^\mu,\boldsymbol{a}\sim\pi_\theta} Q_\phi(\boldsymbol{s},\boldsymbol{a})}_{\text{Policy optimization}} + \frac{1}{\beta}\underbrace{\mathbb{E}_{\boldsymbol{s}\sim\mathcal{D}^\mu,\boldsymbol{a}\sim\pi_\theta} \log \mu(\boldsymbol{a}|\boldsymbol{s})}_{\text{Behavior regularization}} + \frac{1}{\beta}\underbrace{\mathbb{E}_{\boldsymbol{s}\sim\mathcal{D}^\mu}\mathcal{H}(\pi_\theta(\cdot|\boldsymbol{s}))}_{\text{Entropy (often constant}^1\text{)}}.$$

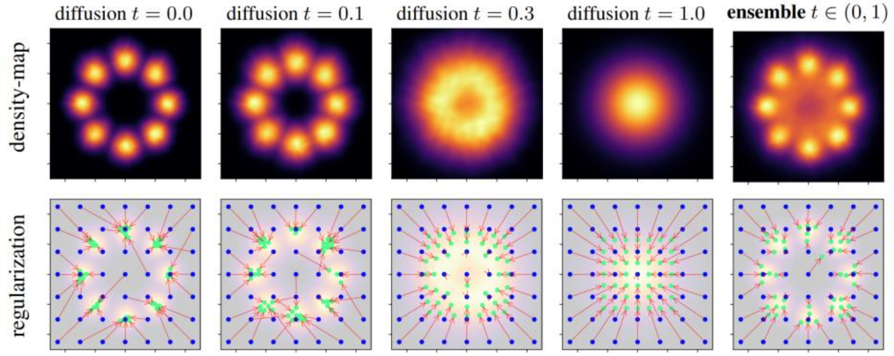
Applying the chain rule and the reparameterization trick:

$$\nabla_\theta \mathcal{L}_\pi(\theta) = \mathbb{E}_{\boldsymbol{s}\sim\mathcal{D}^\mu}\left[\nabla_{\boldsymbol{a}} Q_\phi(\boldsymbol{s},\boldsymbol{a})|_{\boldsymbol{a}=\pi_\theta(\boldsymbol{s})} + \frac{1}{\beta}\underbrace{\nabla_{\boldsymbol{a}}\log\mu(\boldsymbol{a}|\boldsymbol{s})|_{\boldsymbol{a}=\pi_\theta(\boldsymbol{s})}}\right]\nabla_\theta \pi_\theta(\boldsymbol{s}).$$

can be estimated by diffusion models



Ensembling various diffusion times t for better regularization result:



More Experimental Reuslt