Reinforcement Learning via High-Fidelity Generative Behavior Modeling

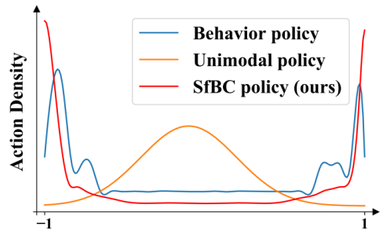
Huayu Chen[1], Cheng Lu[1], Chengyang Ying[1], Hang Su[1], Jun Zhu[1]
1. Tsinghua University

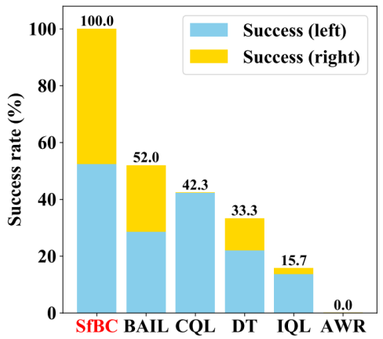
Contact: chenhuay21@mails.tsinghua.edu.cn

Code are available:
https://github.com/ChenDRAG/SfBC

Motivation

- Traditional weighted regression methods generally use Gaussian policies which lack distributional expressivity.
- The behavior dataset are usually highly diverse, and the optimal decision space may be multimodal.
- Limited expressivity may lead to the OOD problem during dynamic programming.





- Diffusion models are powerful generative models, which may potentially be helpful to modeling a heterogeneous behavior dataset.

Challenges

- Diffusion models is an implicit generative model, which means that calculation of log probability is not tractable.
- Weighted regression method cannot be directly applied.

$$\arg\max_{\theta} \quad \mathbb{E}_{(\boldsymbol{s},\boldsymbol{a})\sim\mathcal{D}^{\mu}}\left[\frac{1}{Z(\boldsymbol{s})}\log \pi_{\theta}(\boldsymbol{a}|\boldsymbol{s})\exp\left(\alpha Q_{\phi}(\boldsymbol{s},\boldsymbol{a})\right)\right]$$

Difficult to analytically calculate

Method

Constrained policy search:

$$\arg\max_{\pi} \quad \int_{\mathcal{S}} \rho_{\mu}(\boldsymbol{s}) \int_{\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) Q_{\phi}(\boldsymbol{s},\boldsymbol{a})\, \mathrm{d}\boldsymbol{a}\, \mathrm{d}\boldsymbol{s} - \frac{1}{\alpha}\int_{\mathcal{S}} \rho_{\mu}(\boldsymbol{s}) D_{\mathrm{KL}}\left(\pi(\cdot|\boldsymbol{s})||\mu(\cdot|\boldsymbol{s})\right) \mathrm{d}\boldsymbol{s}.$$

Optimal solution:

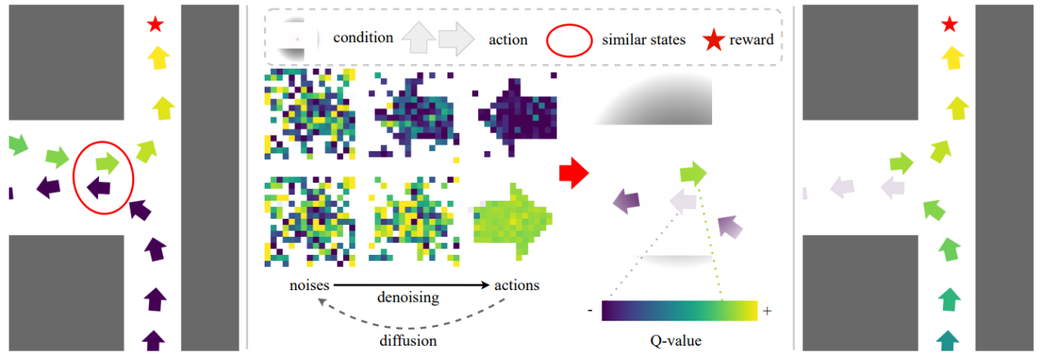
$$\pi^{*}(\boldsymbol{a}|\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{s})}\mu(\boldsymbol{a}|\boldsymbol{s})\exp\left(\alpha Q_{\phi}(\boldsymbol{s},\boldsymbol{a})\right)$$

Policy decoupling:

$$\pi(\boldsymbol{a}|\boldsymbol{s}) \propto \mu_{\theta}(\boldsymbol{a}|\boldsymbol{s})\exp\left(\alpha Q_{\phi}(\boldsymbol{s},\boldsymbol{a})\right)$$

Diffusion modeling:

$$\theta = \arg\min_{\theta} \quad \mathbb{E}_{(\boldsymbol{s},\boldsymbol{a})\sim D^{\mu},\boldsymbol{\epsilon},t}[\|\sigma_t \mathbf{s}_{\theta}(\alpha_t \boldsymbol{a} + \sigma_t \boldsymbol{\epsilon}, \boldsymbol{s}, t) + \boldsymbol{\epsilon}\|_2^2]$$



D4RL Experiments

Algorithm	MuJoCo Locomotion	Antmaze	Maze2d	Kitchen
IQL	76.9	63.0	50.0	53.3
BAIL	71.6	46.7	-	-
DT	74.7	18.7	-	-
Diffuser	75.3	-	119.5	-
SfBC (ours)	75.6	74.2	74.0	57.1

Env Decision space