Diffusion Spectral Representation for Reinforcement Learning

Sampling-Free Diffusion Representation for Efficient Reinforcement Learning

Dmitry Shribak[*1], Chen-Xiao Gao[*2], Yitong Li[1], Chenjun Xiao[3], Bo Dai[1]

[1] Georgia Institute of Technology, [2]Nanjing University, [3]CUHK(SZ)

NEURAL INFORMATION PROCESSING SYSTEMS

Georgia Tech

Outline

- SOTA Diffusion for RL
- MDP framework
- Fundamental question
- Spectral Representation from EBM
- Loss function and algorithms
- Empirical performance
- Summary and conclusion

Current State of Diffusion for RL

Diffusion models are powerful in modeling complex distributions

- As policy: DQL, IDQL, SfBC
- As planner: Diffuser, DD, UniPi
- As world models: Diffusion World Model, PolyGRAD

Diffusion models come with substantial inference cost, where typical sampling method requires > 100 Langevin dynamics calls to generate one sample

MDP Framework

Markov Decision Process: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, r, T, \mu, H \rangle$

- State space: $\mathcal{S}$
- Action space: $\mathcal{A}$
- Reward function: $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$
- Transition: $T : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$
- Initial state distribution μ

$$V_h^\pi(s_h) := \mathbb{E}_{T,\pi}\left[\sum_{t=h}^{H-1} r(s_t, a_t) \mid s_h = s\right]$$

$$Q_h^\pi(s_h, a_h) = \mathbb{E}_{T,\pi}\left[\sum_{t=h}^{H-1} r(s_t, a_t) \mid s_h = s, a_h = a\right]$$

$$\pi(\cdot \mid s) : \mathcal{S} \rightarrow \Delta(\mathcal{A})$$

Fundamental Question

Can we exploit the flexibility of diffusion models with efficient planning and exploration for RL while bypassing the sampling cost?

Spectral Representation

We can represent the dynamics using spectral representation:

SVD Factorization of transition matrix:

$$\mathbb{P}(s' \mid s, a) = \langle \phi^*(s, a), \mu^*(s') \rangle$$

Linear representation of the Q function:

$$\begin{aligned} Q^\pi(s, a) &= r(s, a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot \mid s, a)}\left[V^\pi(s')\right] \\ &= r(s, a) + \left\langle \phi^*(s, a), \underbrace{\int_{\mathcal{S}} \mu^*(s') V^\pi(s') ds'}_{w^\pi} \right\rangle . \end{aligned}$$

Energy Based Model to Spectral Representation

Consider transition probability as an EBM:

$$\mathbb{P}(s'|s,a) = \exp\left(\psi(s,a)^\top \nu(s') - \log Z(s,a)\right), \; Z(s,a) = \int \exp\left(\psi(s,a)^\top \nu(s')\right) ds'$$

Algebra manipulation allows us to get:

$$\mathbb{P}(s'|s,a) \propto \exp\left(\|\psi(s,a)\|^2/2\right) \exp\left(-\|\psi(s,a) - \nu(s')\|^2/2\right) \exp\left(\|\nu(s')\|^2/2\right)$$

Decomposing central term with Random Fourier Feature (RFF):

$$\mathbb{P}(s'|s,a) = \langle \phi_\omega(s,a), \mu_\omega(s') \rangle_{\mathcal{N}(\omega)}$$

$$\Rightarrow$$

$$\phi_\omega(s,a) = \exp\left(-\mathbf{i}\omega^\top \psi(s,a)\right) \exp\left(\|\psi(s,a)\|^2/2 - \log Z(s,a)\right)$$
$$\mu_\omega(s') = \exp\left(-\mathbf{i}\omega^\top \nu(s')\right) \exp\left(\|\nu(s')\|^2/2\right).$$

Diffusion Model connection to Energy Based Model

Consider perturbation on next state: $\mathbb{P}(\tilde{s}'|s', \beta) = \mathcal{N}(\sqrt{1-\beta}s', \beta I)$

We get perturbed transition function: $\mathbb{P}\left(\tilde{s}' \mid s, a; \beta\right) = \int \mathbb{P}\left(\tilde{s}' \mid s'; \beta\right) \mathbb{P}\left(s' \mid s, a\right) ds' \propto \exp\left(\psi(s,a)^\top \nu\left(\tilde{s}', \beta\right)\right)$

Using score-matching objective: $\mathbb{E}_{p(\mathbf{x})}[\|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x})\|_2^2]$

We get a final objective:

$$\min_\theta \mathbb{E}_\beta \mathbb{E}_{(s,a,s')} \left[\|s_\theta(s, a, \tilde{s}'; \beta) - \nabla_{\tilde{s}'} \log \mathbb{P}(\tilde{s}'|s, a; \beta)\|^2\right]$$

$$\min_\theta \mathbb{E}_\beta \mathbb{E}_{(s,a,s')} \left[\|\psi_\theta(s, a)^\top \mu_\theta(\tilde{s}'; \beta) + \frac{\tilde{s}' - \sqrt{1-\beta}s'}{\beta}\|^2\right]$$

Algorithm 1 - Diff-SR Training

Initialize representation networks ψ, ζ,noise levels $\{\beta^k\}_{k=1}^T$,data buffer $\mathcal{D} = \emptyset$

for update step t = 1 to N_{rep} :

Sample a batch of n transitions $\{(s_i, a_i, s'_i)\}_{i=1}^n$

Sample noise schedules for each transition $\{\beta_i\}_{i=1}^n \sim \text{Uniform}(\beta^1, \beta^2, \ldots, \beta^T)$

Corrupt the next states $\tilde{s}'_i \leftarrow \sqrt{1-\beta_i} s'_i + \sqrt{\beta_i}\epsilon_i$

Optimize ψ, ζ via gradient descent by minimizing

$$\min_\theta \mathbb{E}_\beta \mathbb{E}_{(s,a,s')} \left[\| \psi_\theta(s,a)^\top \mu_\theta(\tilde{s}'; \beta) + \frac{\tilde{s}' - \sqrt{1-\beta} s'}{\beta} \|^2 \right]$$

Algorithm 2 - Online RL with Diff-SR

Initialize policy π, double Q (critic) network $(\xi_1, \theta_1), (\xi_2, \theta_2)$, data buffer $\mathcal{D} = \emptyset$

for timestep t = 1 to T:

Sample $a_t \sim \pi(\cdot|s_t)$ $r_t = r(s_t, a_t)$, $s'_t \sim \mathbb{P}(\cdot|s_t, a_t)$

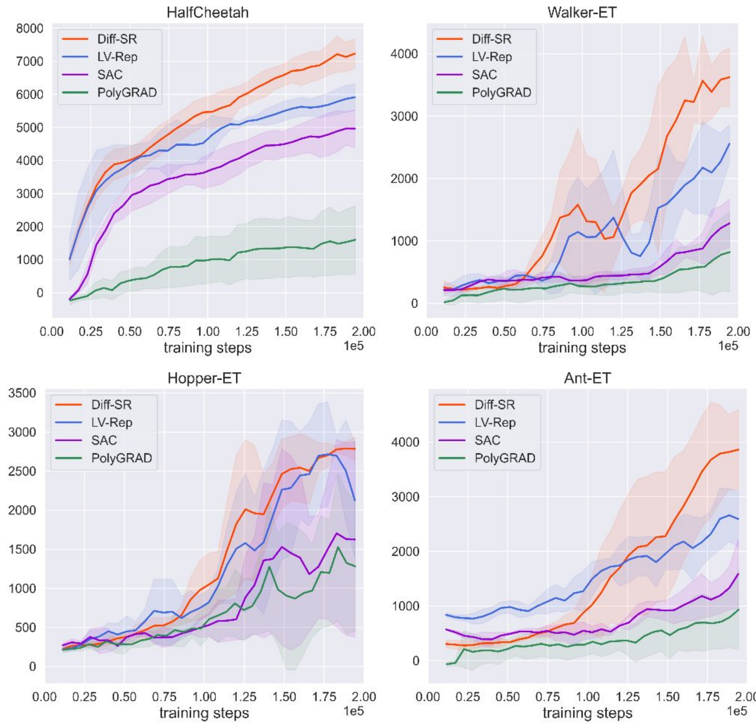
Update data buffer $\mathcal{D} \leftarrow \mathcal{D} \cup (s_t, a_t, r_t, s'_t)$

Train representation ψ with $\mathcal{D}$ by obj. func.

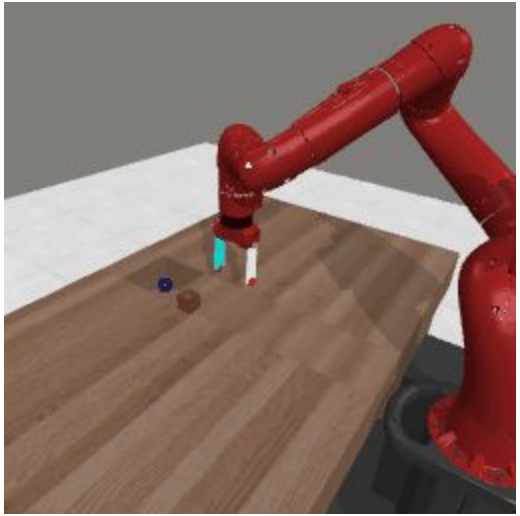
Update critic $(\xi_1, \theta_1), (\xi_2, \theta_2)$ by standard TD loss

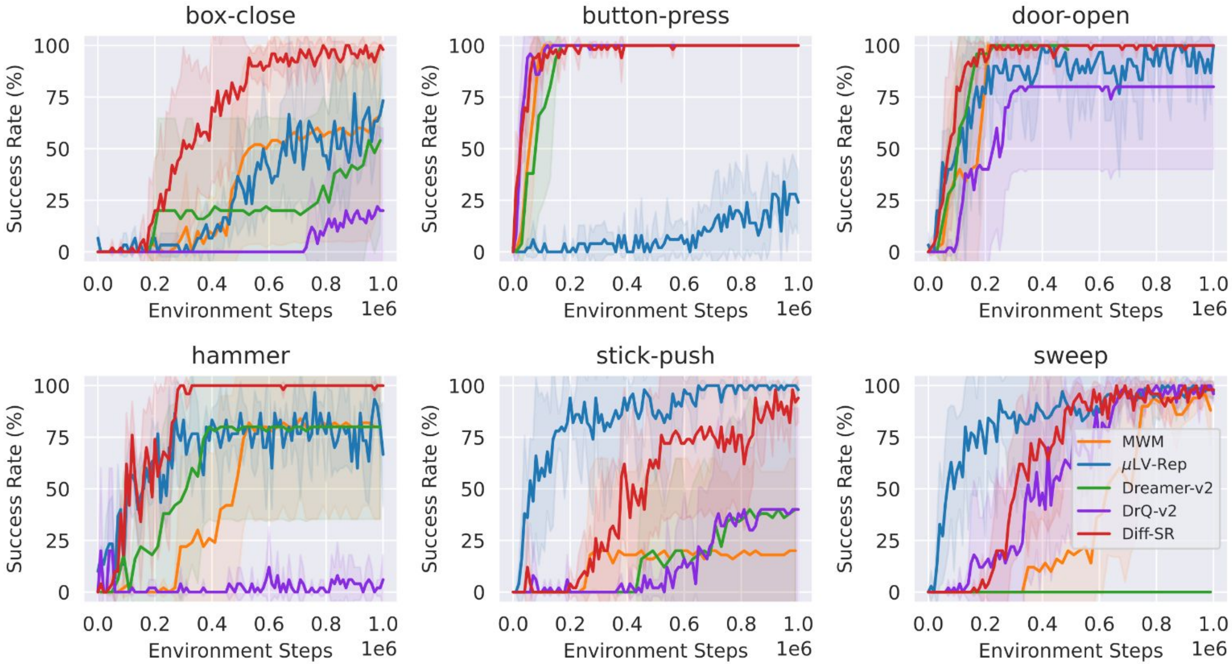
Update policy π with $\max_\pi \mathbb{E}_{s \sim \mathcal{D}, a \sim \pi}[\min_{i \in \{1,2\}} Q_{\xi_i, \theta}(s, a)]$

Empirical Performance - Mujoco



Empirical Performance - MetaWorld





Thanks !

Diffusion Spectral Representation for Reinforcement Learning

Dmitry Shribak[*1], Chen-Xiao Gao[*2], Yitong Li[1], Chenjun Xiao[3], Bo Dai[1]

[1] Georgia Institute of Technology, [2]Nanjing University, [3]CUHK(SZ)

NEURAL INFORMATION PROCESSING SYSTEMS

Georgia Tech