Learning a Diffusion Model Policy from Rewards via Q-Score Matching

Michael Psenka*, Alejandro Escontrela*, Pieter Abbeel, Yi Ma

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Score matching for diffusion model policies, purely from rewards.

Problem motivation & setup

Diffusion models have risen as an interesting way to parametrize policies in reinforcement learning, due to their expressivity. While such policies in the behavior cloning setting are naturally posed, since we have samples of the distribution we want to sample from, optimizing diffusion model policies from rewards is a fundamentally different scenario that requires special attention.

Our paper sets up a theoretical framework to analyze this regime, proves a certain class of training methods can effectively perform Q learning from diffusion model policies, and demonstrate the practical benefits of such an algorithm.

Our method

Core idea: matching denoising model to $\nabla_a Q$ effectively learns optimal policies.

By representing our system as a joint SDE:

$$ds = F(s,a)dt + \Sigma_s(s,a)dB_t^s$$
$$da = \Psi(s,a)dt + \Sigma_a(s,a)dB_t^a$$

we can prove that any update that pushes the denoising model $\Psi(s,a)$ towards the action gradient of the policy's Q-function $\nabla_a Q(s,a)$ will strictly increase expected rewards.

Key results

- Comparable results to popular baselines that do not use as expressive of policy classes.
- Uniquely learns explorative policies compared to alternative methods for training diffusion model policies.

Key takeaways

- Diffusion model policies can be effectively and efficiently optimized for Q-learning by matching the denoising model against $\nabla_a Q$.
- Converged policies from such training are multi-modal and still explore, without explicit entropy regularization or exploring terms.

Website: