Reflected Flows for On-support offline RL via noise Manipulation

ICLR 2026

Background

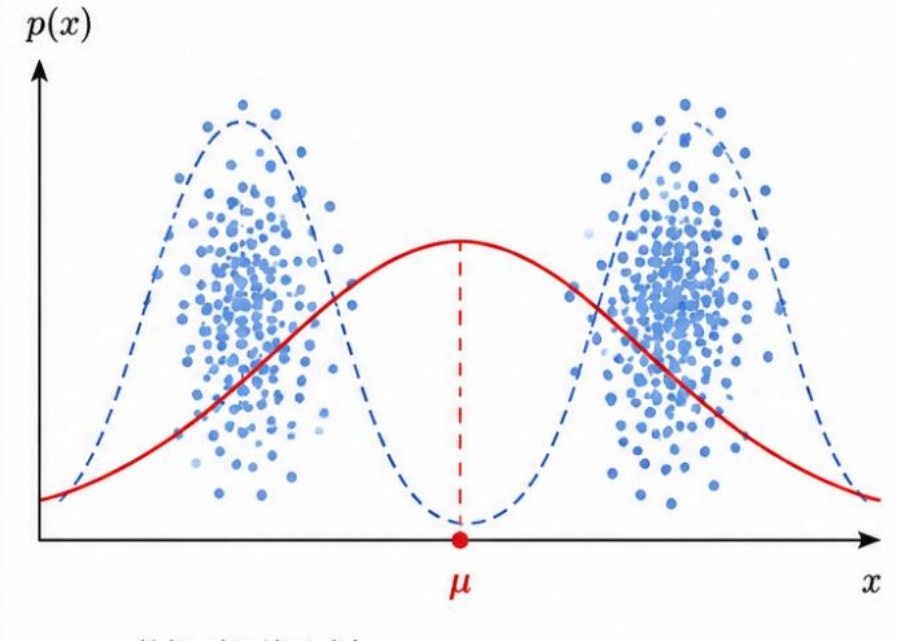
- **The Status Quo:**

Offline RL learns from static, sub-optimal datasets (e.g., human demonstrations, mixed policies).

- **The Expressivity Bottleneck:**

Traditional parametric policies (e.g., Single Gaussian) fail in complex environments.

- **Multi-modal Data:** Divergent actions for the same state (e.g., turning left vs. turning right).
- **Mode-Averaging Disaster:** Averaging multimodal actions leads to disastrous, out-of-distribution (OOD) execution (e.g., crashing straight into a wall).

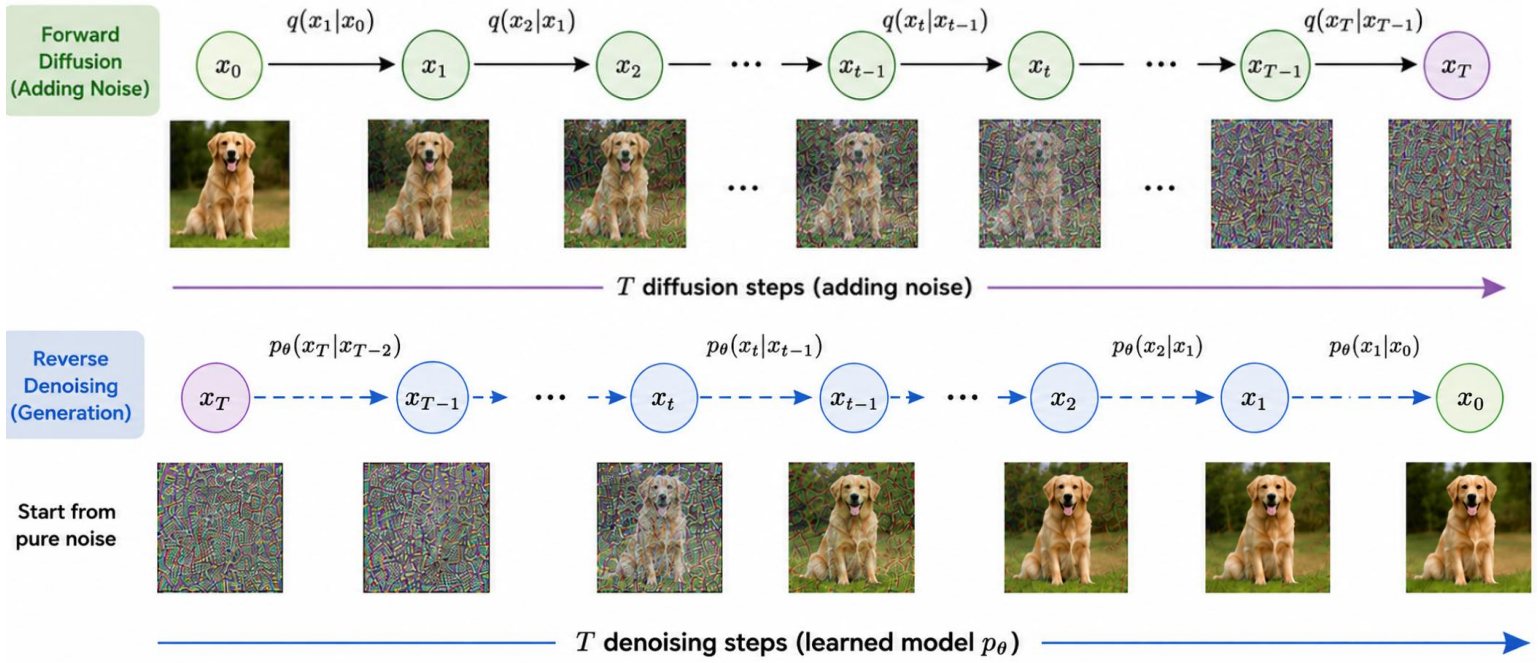


Background

- **The Paradigm Shift:**

Generative Offline RL (e.g., Diffusion-QL) perfectly models multi-modal distributions.

- **The Denoising Tax (Fatal Flaws):**
  - **Prohibitive Inference Latency:** Requires dozens/hundreds of iterative denoising steps. Unusable for high-frequency real-world robotics (e.g., 50Hz+ control).
  - **Training Instability:** Backpropagation Through Time (BPTT) across Markov chains makes Q-learning gradients noisy and computationally expensive.
  - **The Dilemma:** Can we achieve Diffusion-level expressivity with MLP-level speed?



Background

The Vulnerability: The Soft Penalty Tug-of-War

The Loss balances two forces: Q-loss (greedy reward maximization) vs. BC-loss (a soft penalty pulling actions back to the clover shape).

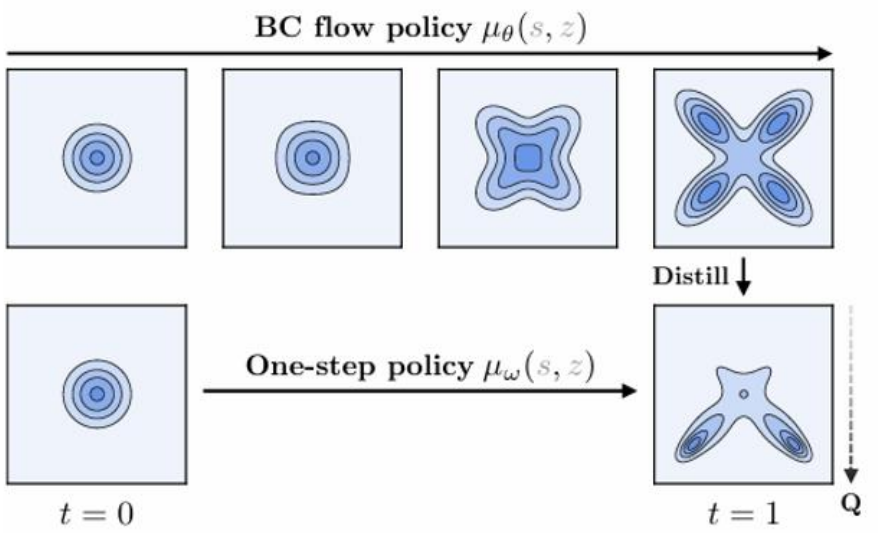
The Achilles' Heel (α):

α too large → Overly conservative (fails to reach optimal rewards).

α too small → OOD Catastrophe! The Q-gradient snaps the constraint, leading to severe hallucinations.

$$\mathcal{L}_{\text{Flow}}(\theta) = \mathbb{E}_{\substack{s,a=x^1\sim\mathcal{D},\\ x^0\sim\mathcal{N}(0,I_d),\\ t\sim\text{Unif}([0,1])}} \left[\|v_\theta(t,s,x^t) - (x^1 - x^0)\|_2^2\right], \tag{5}$$

$$\mathcal{L}_\pi(\theta) = \underbrace{\mathbb{E}_{s\sim\mathcal{D},a^\pi\sim\pi_\theta}[-Q_\phi(s,a^\pi)]}_{\texttt{Q loss}} + \underbrace{\alpha\mathcal{L}_{\text{Flow}}(\theta)}_{\texttt{BC loss}}. \tag{6}$$



Method

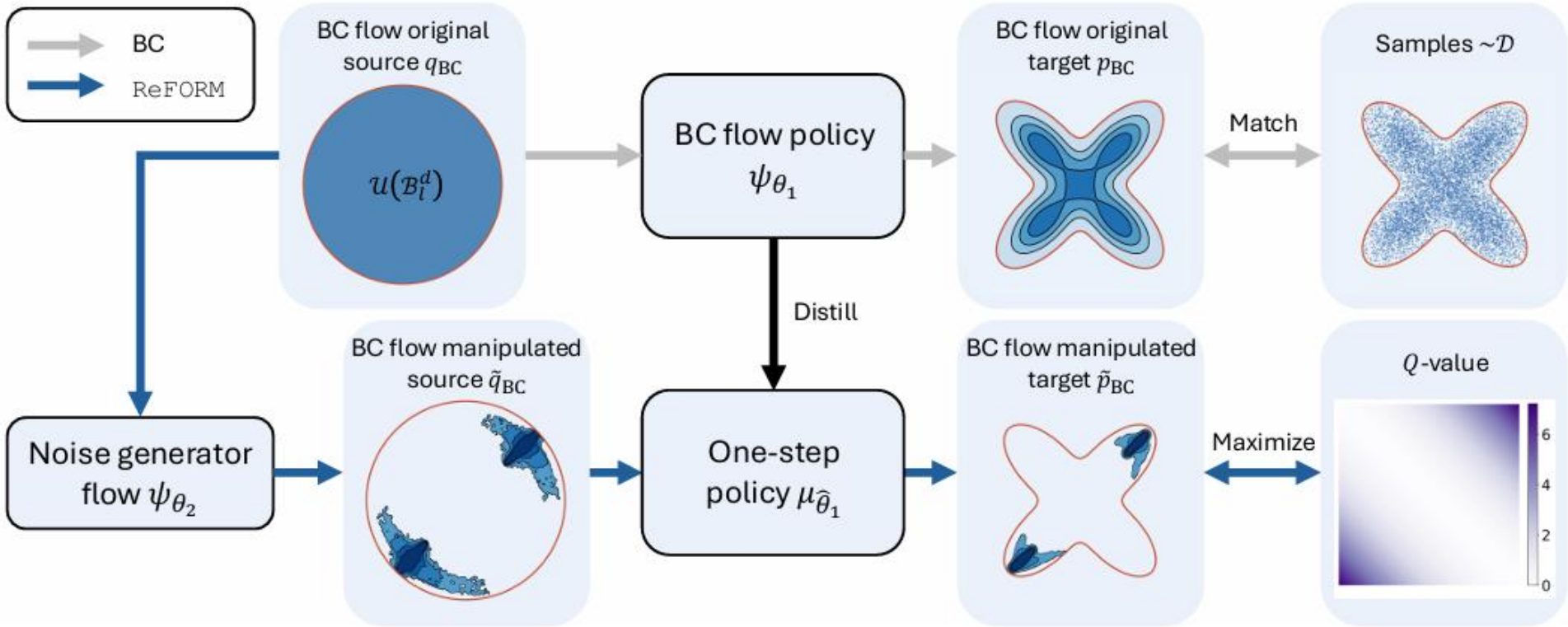


Figure 1: **ReFORM algorithm.** The process with gray arrows indicates the BC flow policy, learned to transform a simple source distribution $q_{\text{BC}} = \mathcal{U}(\mathcal{B}_l^d)$ to a target distribution p_{BC} that matches the dataset $\mathcal{D}$. The blue arrows indicate the ReFORM process, where we learn a flow noise generator to generate a manipulated source distribution $\tilde{q}_{\text{BC}}$ for the BC policy so that the manipulated target $\tilde{p}_{\text{BC}}$ maximizes the Q value while staying inside the support (denoted in red) of the BC policy.

Method

The "Iron Cage" (Eq. 7):

- Replaces standard unbounded Gaussian noise with a uniform distribution over a hypersphere of radius l.

$$\mathrm{supp}(q_{\mathrm{BC}}) = \mathcal{B}_l^d := \{z \in \mathbb{R}^d \mid \|z\| \leq l\}. \tag{7}$$

The Fast Proxy (Eq. 13):

- Distills the slow multi-step ODE into a one-step network
- Crucial: Solves the Backpropagation Through Time (BPTT) bottleneck for downstream Q-gradient flow.

$$\mathcal{L}_{\mathrm{Distill}}(\hat{\theta}_1) = \mathbb{E}_{s\sim\mathcal{D}, z\sim\mathcal{U}(\mathcal{B}_l^d)} \left[\|\mu_{\hat{\theta}_1}(z; s) - \mu_{\theta_1}(z; s)\|^2 \right]. \tag{13}$$

Method

The Reflected ODE (Eq. 9):

- The reflection term dL_t is the "invisible wall". It pushes the motion back whenever the noise tries to escape the hypersphere boundary q_{BC}.

$$d\psi_{\theta_2}(t, w; s) = v_{\theta_2}(t, \psi_{\theta_2}(t, w; s); s)dt + dL_t, \quad \psi_{\theta_2}(0, w; s) = w, \tag{9}$$

The Objective (Eq. 10):

- Maximize Q-value by manipulating the initial noise ω.

$$\mathcal{L}_{\mathrm{NG}}(\theta_2) = \mathbb{E}_{s\sim\mathcal{D}, w\sim\mathcal{U}(\mathcal{B}_l^d)}\left[- Q^{\mu_\theta}\left(s, \mu_{\theta_1}(\mu_{\theta_2}(w; s); s)\right)\right], \tag{10}$$

Method

Standard Euler (Eq. 11 - The Problem):

- Discrete steps can easily overshoot the boundary → Catastrophic OOD!

$$z_{k+1} = z_k + v_{\theta_2}(k\Delta t, w; s)\Delta t, \quad k \in \{0, \ldots, N-1\}, \quad \psi_{\theta_2}(1, w; s) \leftarrow z_N, \tag{11}$$

Reflected Euler (Eq. 12 - The Solution):

- Elegant Projection: If the particle tries to step outside, mathematically subtract its outward velocity component $< \nu\Delta t, n > n$.
- Result: The particle glides smoothly along the inner wall. Fully differentiable!

$$z_{k+1} = \mathbf{1}\{\hat{z}_{k+1} \in \mathcal{B}_l^d\}\hat{z}_{k+1} + (1 - \mathbf{1}\{\hat{z}_{k+1} \in \mathcal{B}_l^d\})\left(\hat{z}_{k+1} - \langle v_{\theta_2}(k\Delta t, w; s)\Delta t, n_{k+1}\rangle n_{k+1}\right), \tag{12}$$

Experiment

Environments

We evaluate ReFORM and the baselines on 40 tasks from the OGBench offline RL benchmark (Park et al., 2025a) designed in 4 environments, including locomotion tasks and manipulation tasks. We use two kinds of datasets, CLEAN and NOISY. The CLEAN dataset consists of random environment trajectories generated by an expert policy. The NOISY dataset consists of random trajectories generated by a highly suboptimal and noisy policy.

Baselines.

We compare ReFORM with the state-of-the-art offline RL algorithms with flow policies,including Flow Q-Learning (FQL) (Park et al., 2025b), Implicit Flow Q-Learning (IFQL) (Parket al., 2025b), and Diffusion Steering via RL (DSRL) (Wagenmaker et al., 2025). Since FQL'sperformance highly depends on the α hyperparameter (Eq. (3)), we consider three variants of FQL:FQL(M) uses the $\alpha*$ that is hand-tuned for each environment using the CLEAN dataset by Parket al. (2025b), FQL(S) uses $\alpha = \alpha*/10$, and FQL(L) uses $\alpha = 10 \cdot \alpha*$. IFQL is the flow version of IDQL (Hansen-Estruch et al., 2023b) implemented in Park et al. (2025b). For DSRL, we use the hand-tuned noise bound by Wagenmaker et al. (2025). Note that ReFORM uses the same hyperparameters across all tasks.

Experiment

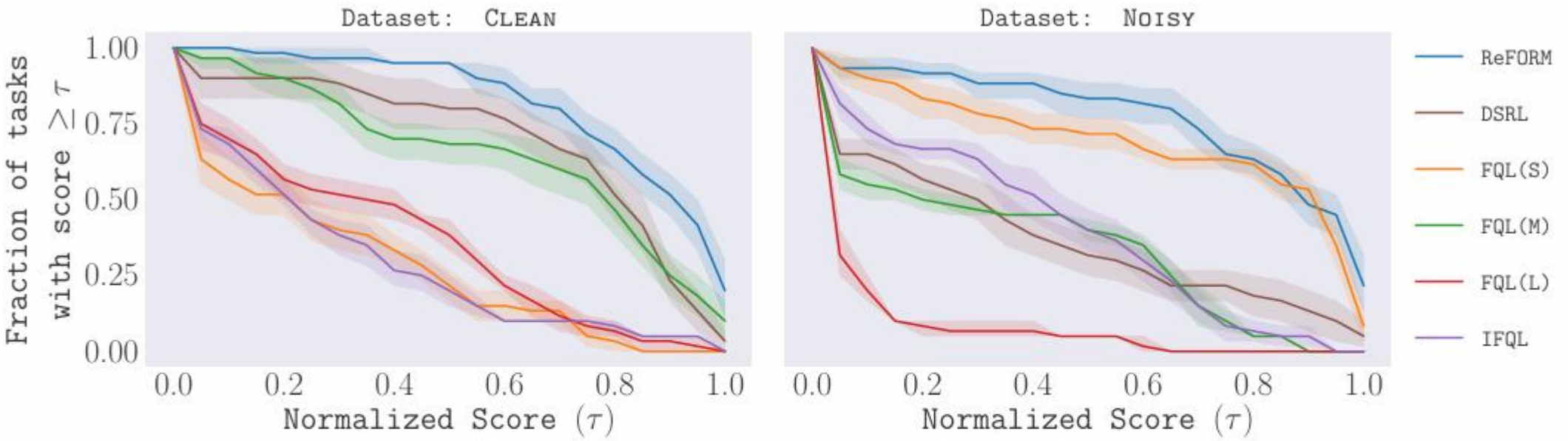
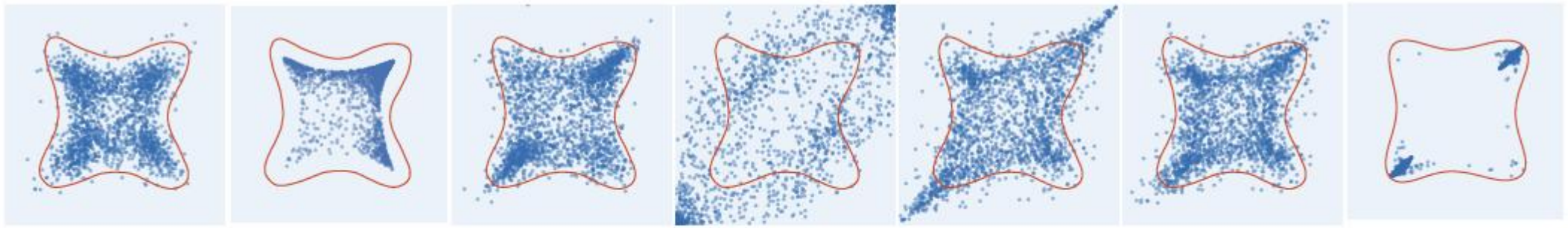


Figure 2: **Performance profile over CLEAN and NOISY datasets.** For a given normalized score τ (x-axis), the performance profile shows the probability that a given method achieves a score $\geq \tau$ (see Agarwal et al. (2021) for details). On the CLEAN dataset, ReFORM achieves greater scores with higher probabilities than all other baselines. The same is true on the NOISY dataset except for a small set of normalized scores around 0.9 where ReFORM and FQL(S) have similar probabilities within the statistical margins.

Experiment



(a) BC. (b) DSRL. (c) IFQL. (d) FQL(S). (e) FQL(M). (f) FQL(L). (g) ReFORM[3].

Figure 3: **Learned policy distributions with the toy example.** The Q-value reaches the maximum at the lower left and upper right corners (See the Q-value plot in Figure 1). The red boundaries denote the estimated $\mathrm{supp}(\pi_{\mathrm{BC}})$[4].

Thanks