On Pathologies in KL-Regularized Reinforcement Learning from Expert Demonstrations





Tim G. J. Rudner* **Cong Lu*** Michael A. Osborne Yarin Gal Yee Whye Teh

Neural Information Processing Systems 2021

Correspondence to

tim.rudner@cs.ox.ac.uk

How can we use expert demonstrations to effectively accelerate online training in RL?

KL-regularization balances fitting online data

and matching a behavioral expert policy.

KL-regularization balances fitting online data

and matching a behavioral expert policy.

KL-regularization balances fitting online data

and matching a behavioral expert policy.

Problem:

Problem:

KL-regularized RL can suffer from pathological behavior during training.

Problem:

KL-regularized RL can suffer from pathological behavior during training.

Problem:

KL-regularized RL can suffer from **pathological behavior** during training.

Expert Demonstration

Learned Behavior

Problem:

KL-regularized RL can suffer from **pathological behavior** during training.

Expert Demonstration

Learned Behavior

How can we avoid pathological behavior

that may result in poor policies?

We show:

We show:

why such pathologies may occur in theory;

We show:

why such pathologies may occur in theory;

why they occur in practice;

We show:

why such pathologies may occur in theory;

why they occur in practice;

how to prevent them.

Goal

Goal

- Learn a good policy in **as few environment interactions as possible**

Reinforcement Learning from Expert Demonstrations

Goal

- Learn a good policy in **as few environment interactions as possible**

How?

- Use expert demonstrations to **give agents a head start**

Reinforcement Learning from Expert Demonstrations

Goal

- Learn a good policy in **as few environment interactions as possible**

How?

- Use expert demonstrations to **give agents a head start**
- Common approach: **Behavioral cloning**
  - offline: $\mathcal{D}_0 = \{(\mathbf{s}_n, \mathbf{a}_n)\}_{n=1}^{N} = \{\overline{\mathbf{S}}, \overline{\mathbf{A}}\} \longrightarrow \pi_0(\cdot|\mathbf{s})$

Reinforcement Learning from Expert Demonstrations

Goal

- Learn a good policy in **as few environment interactions as possible**

How?

- Use expert demonstrations to **give agents a head start**
- Common approach: Behavioral cloning + **KL regularization**
  - offline: $\mathcal{D}_0 = \{(\mathbf{s}_n, \mathbf{a}_n)\}_{n=1}^{N} = \{\overline{\mathbf{S}}, \overline{\mathbf{A}}\} \longrightarrow \pi_0(\cdot|\mathbf{s})$
  - online: $\tilde{R}(\boldsymbol{\tau}_t) = \sum_{k=t}^{\infty} \gamma^k \left[r(\mathbf{s}_k, \mathbf{a}_k) - \alpha \mathbb{D}_{\mathrm{KL}}\left(\pi(\cdot \mid \mathbf{s}_k) \| \pi_0(\cdot \mid \mathbf{s}_k)\right)\right]$

Kullback-Leibler divergence

$$\tilde{R}\left(\boldsymbol{\tau}_t\right) = \sum_{k=t}^{\infty} \gamma^k \left[r\left(\mathbf{s}_k, \mathbf{a}_k\right) - \alpha \mathbb{D}_{\mathrm{KL}}\left(\pi\left(\cdot \mid \mathbf{s}_k\right) \| \pi_0\left(\cdot \mid \mathbf{s}_k\right)\right)\right]$$

Note!

KL-Regularized Reinforcement Learning

Kullback-Leibler divergence

$$\tilde{R}(\boldsymbol{\tau}_t) = \sum_{k=t}^{\infty} \gamma^k \left[r(\mathbf{s}_k, \mathbf{a}_k) - \alpha \mathbb{D}_{\mathrm{KL}}\left(\pi(\cdot \mid \mathbf{s}_k) \,\|\, \pi_0(\cdot \mid \mathbf{s}_k)\right)\right]$$

Note!

- KL divergence is well-defined (i.e., finite) **if and only if** learned policy is **absolutely continuous** w.r.t. behavioral policy

KL-Regularized Reinforcement Learning

Kullback-Leibler divergence

$$\tilde{R}\left(\boldsymbol{\tau}_t\right)=\sum_{k=t}^{\infty} \gamma^k\left[r\left(\mathbf{s}_k, \mathbf{a}_k\right)-\alpha \mathbb{D}_{\mathrm{KL}}\left(\pi\left(\cdot \mid \mathbf{s}_k\right) \| \pi_0\left(\cdot \mid \mathbf{s}_k\right)\right)\right]$$

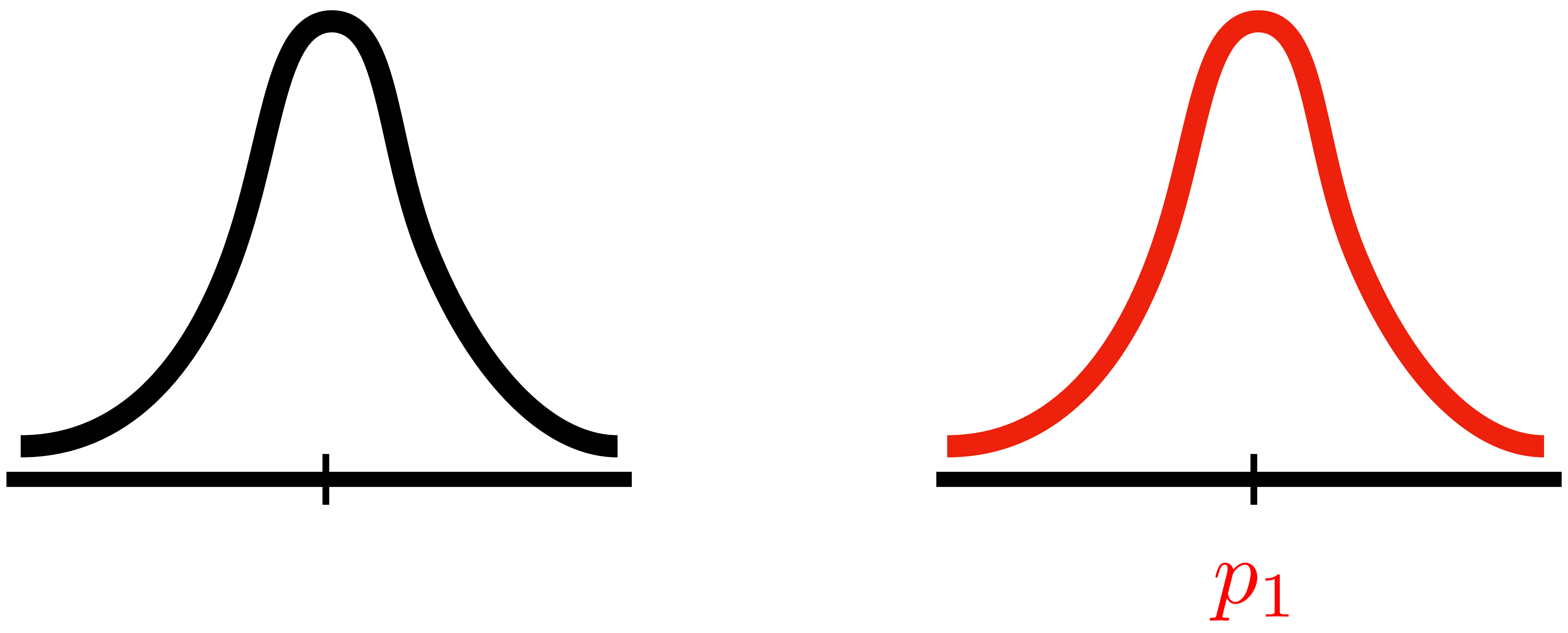
Note!

- KL divergence is well-defined (i.e., finite) **if and only if** learned policy is **absolutely continuous** w.r.t. behavioral policy
- Potential failure mode: **degenerate behavioral policies**

Could this be an issue in practice?

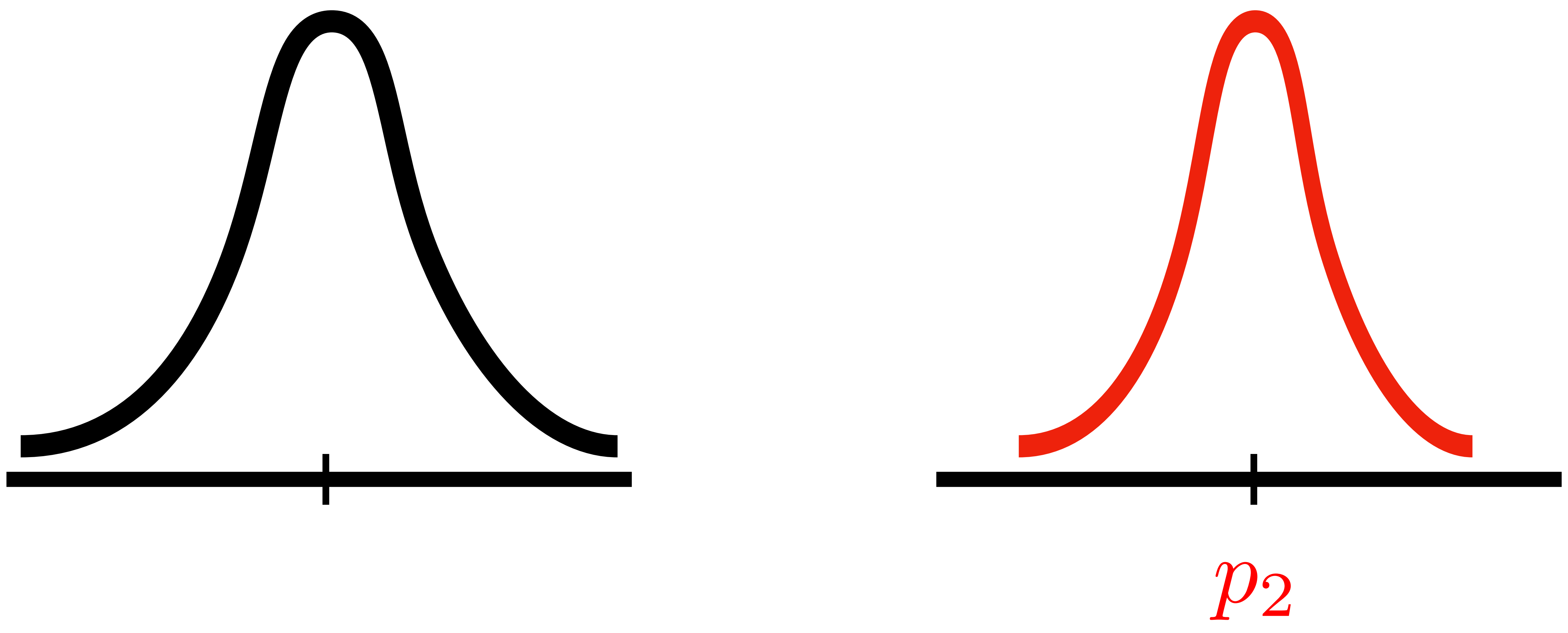
When Is The KL Divergence Well-Defined?

$$\mathbb{D}_{\mathrm{KL}}\left(q \| p_1\right) = 0$$



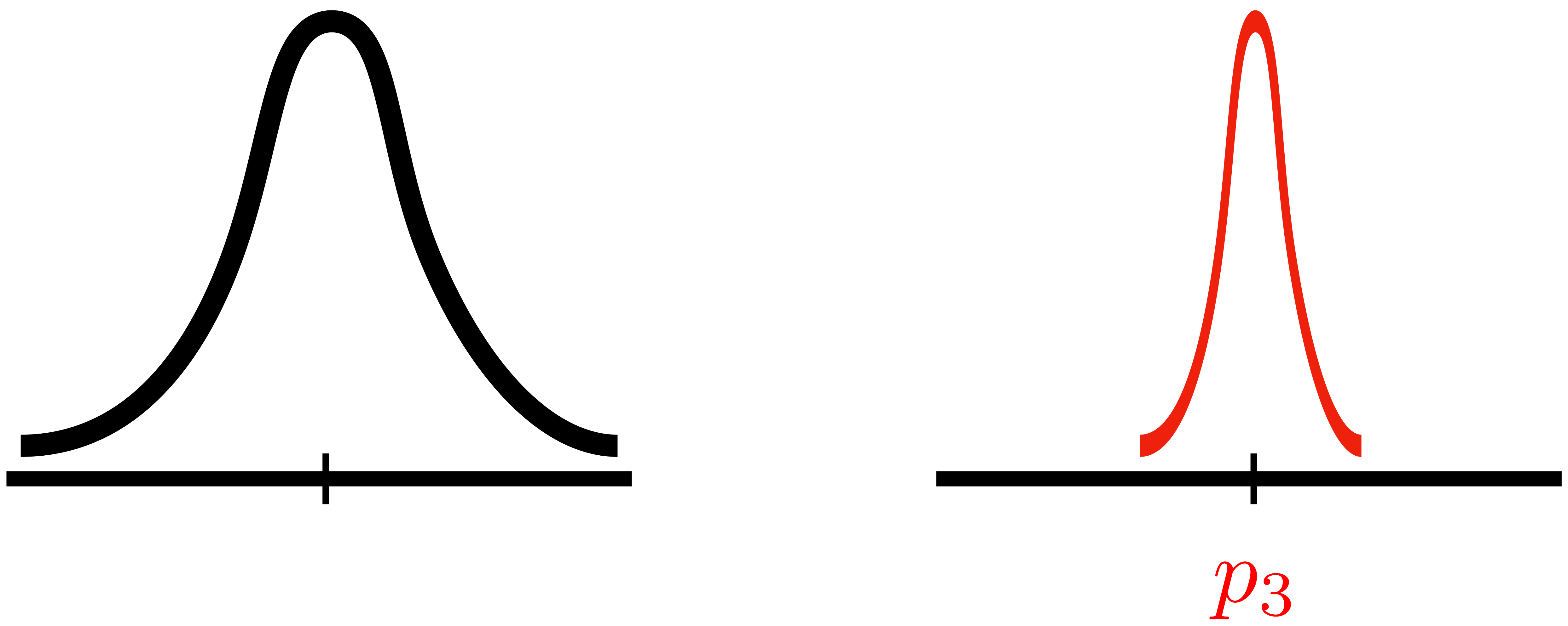
When Is The KL Divergence Well-Defined?

$$\mathbb{D}_{\mathrm{KL}}\left(q \| p_1\right) = 0 < \mathbb{D}_{\mathrm{KL}}\left(q \| p_2\right)$$



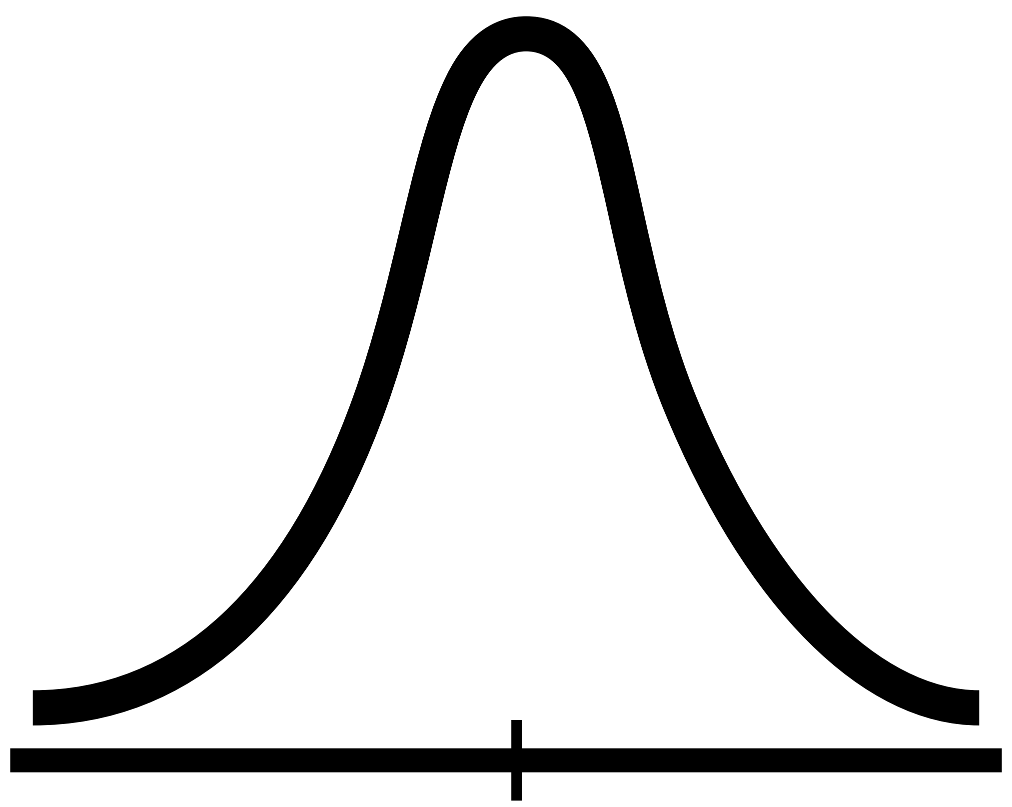
When Is The KL Divergence Well-Defined?

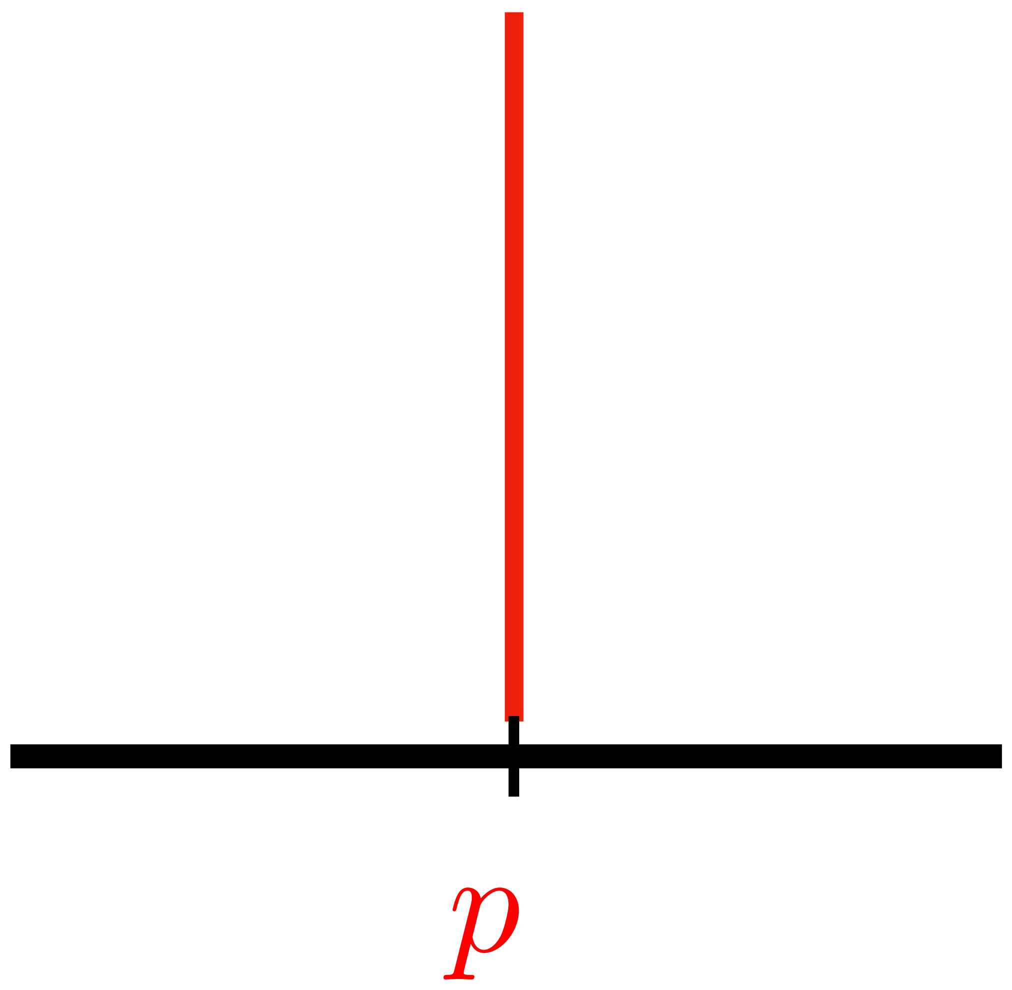
$$\mathbb{D}_{\mathrm{KL}}\left(q \| p_1\right) = 0 < \mathbb{D}_{\mathrm{KL}}\left(q \| p_2\right) < \mathbb{D}_{\mathrm{KL}}\left(q \| p_3\right)$$



When Is The KL Divergence Well-Defined?

$$\mathbb{D}_{\mathrm{KL}}\left(q \| p_1\right) = 0 < \mathbb{D}_{\mathrm{KL}}\left(q \| p_2\right) < \mathbb{D}_{\mathrm{KL}}\left(q \| p_3\right) < \mathbb{D}_{\mathrm{KL}}\left(q \| p\right) = \infty$$





When Is KL-Regularized Meaningful?

Potential Failure Mode

- If variance of behavioral policy tends to zero, KL blows up

When Is KL-Regularized Meaningful?

Potential Failure Mode

- If variance of behavioral policy tends to zero, KL blows up

Is this a problem in practice?

- How fast does the KL divergence blow up?
- Do commonly used behavioral policy have vanishingly small variance?

Estimating Behavioral Policies via Maximum Likelihood

Parametric policy predictive variance

- Collapse in predictive variance away from expert trajectories

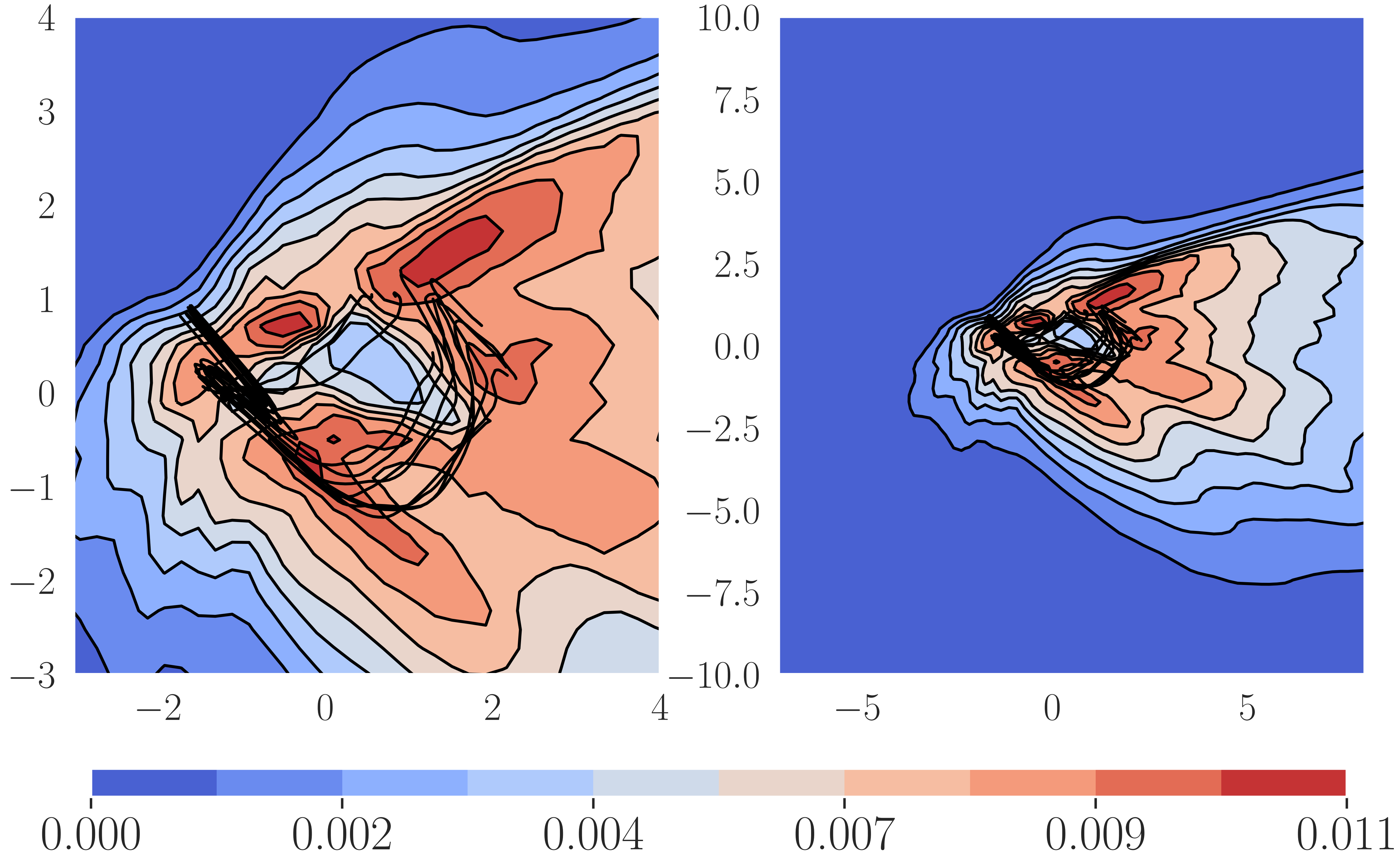
Parametric policy predictive variance

- Collapse in predictive variance away from expert trajectories

Estimating Behavioral Policies via Maximum Likelihood

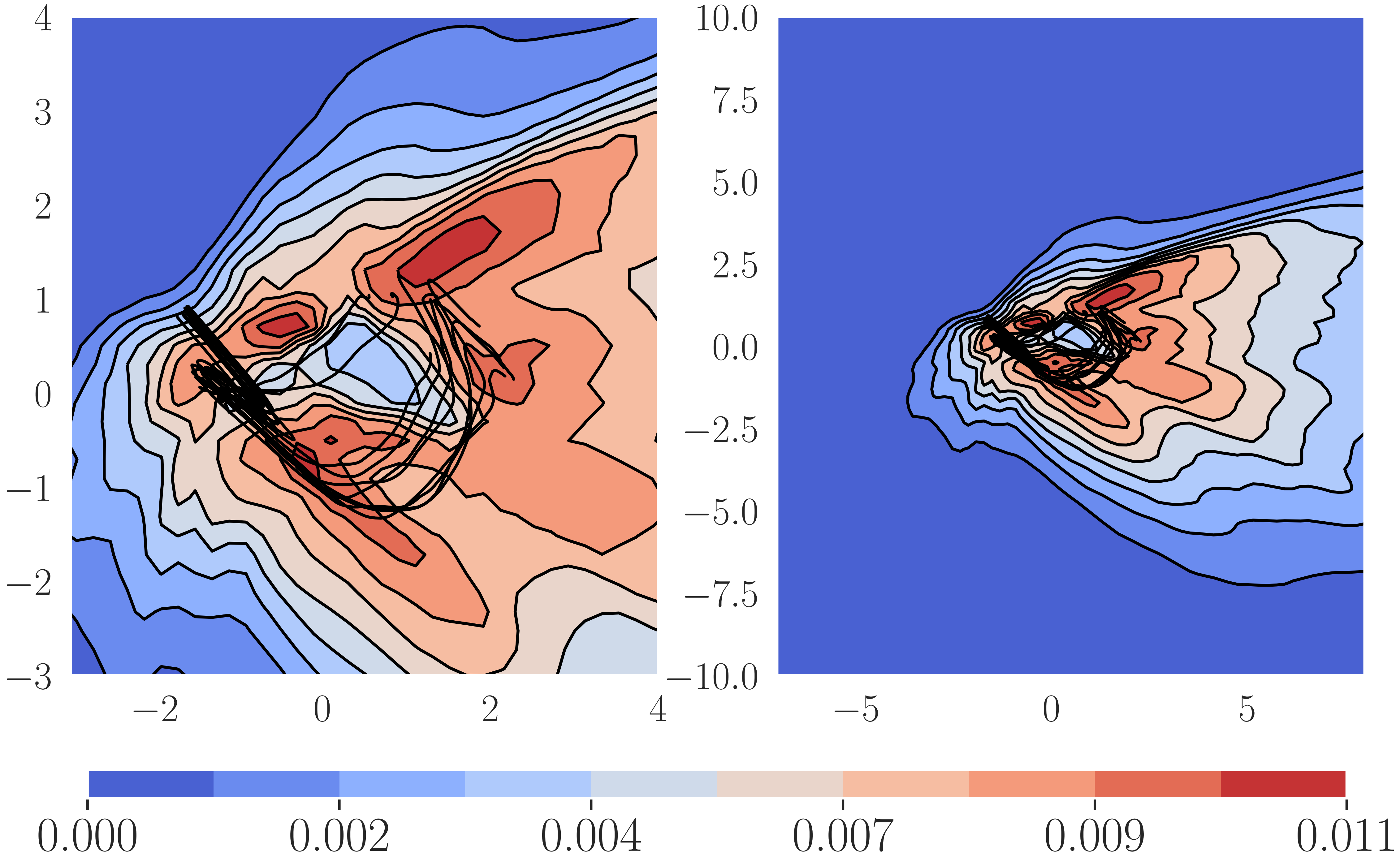
Parametric policy predictive variance

- Collapse in predictive variance away from expert trajectories

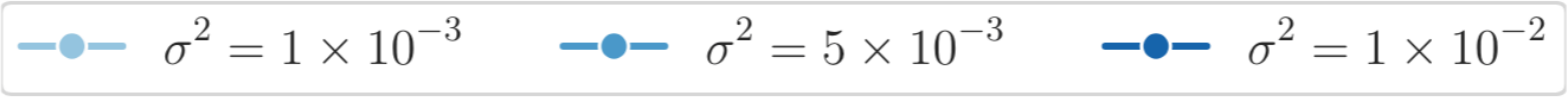


Parametric policy predictive variance

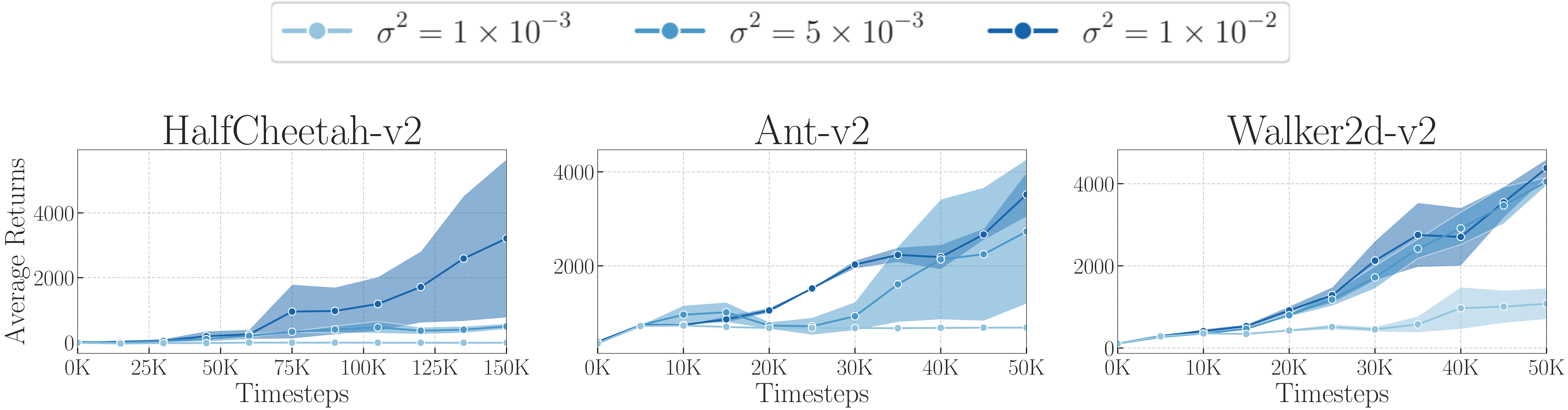
- Collapse in predictive variance away from expert trajectories



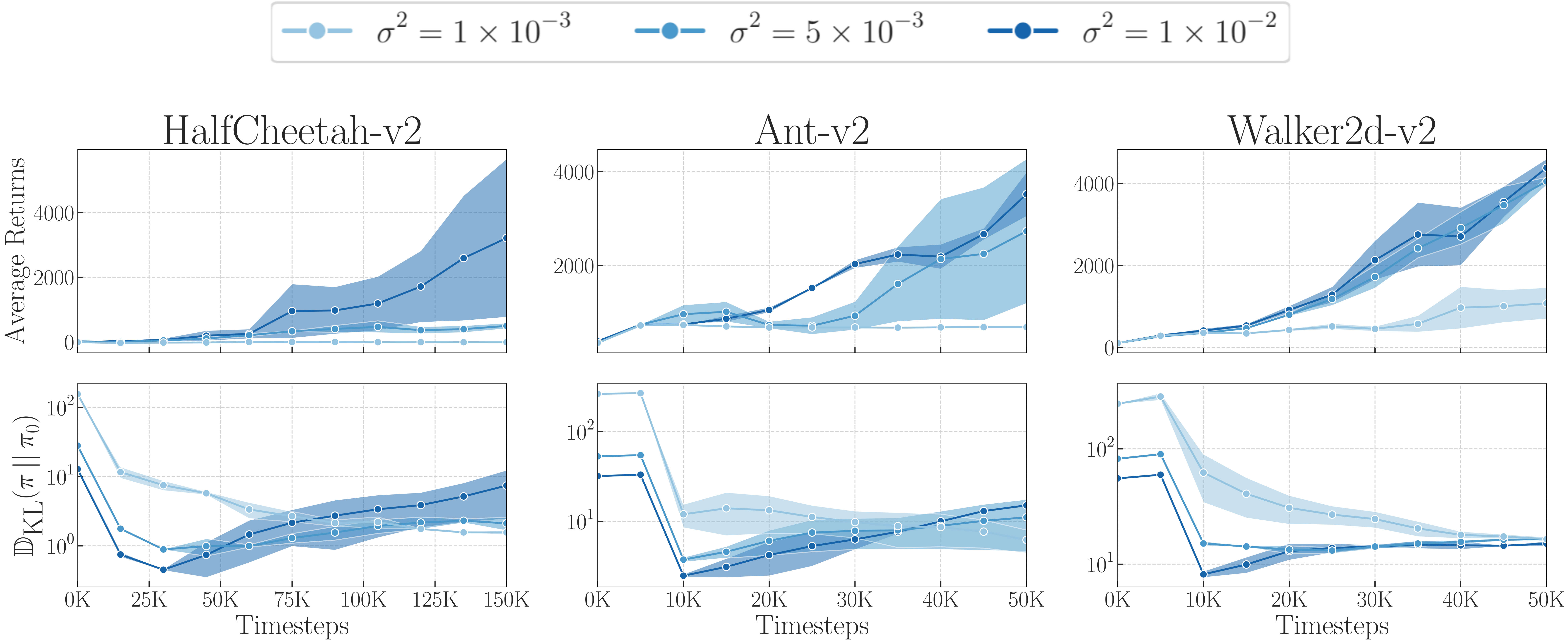
Effect of Decreasing Prior Variance on Performance



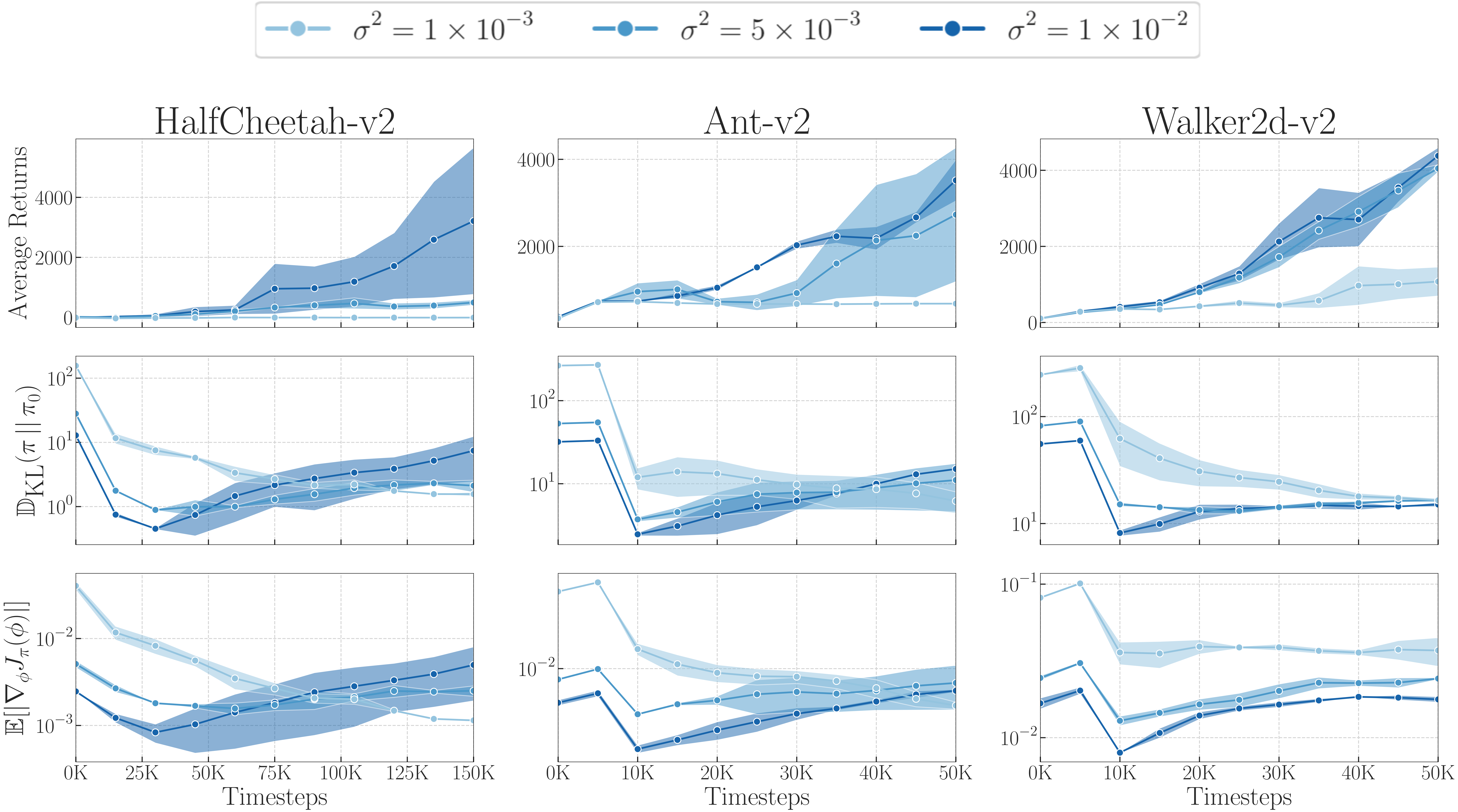
Effect of Decreasing Prior Variance on Performance



Effect of Decreasing Prior Variance on Performance



Effect of Decreasing Prior Variance on Performance



Exploding Gradients in KL-Regularized RL Objectives

Proposition 1 (informal).

Proposition 1 (informal).

- Let the objective function be given by

$$\tilde{R}(\boldsymbol{\tau}_t) = \sum_{k=t}^{\infty} \gamma^k \left[r(\mathbf{s}_k, \mathbf{a}_k) - \alpha \mathbb{D}_{\mathrm{KL}} \left(\pi(\cdot \mid \mathbf{s}_k) \,\|\, \pi_0(\cdot \mid \mathbf{s}_k) \right) \right]$$

Exploding Gradients in KL-Regularized RL Objectives

Proposition 1 (informal).

- Let the objective function be given by

$$\tilde{R}\left(\boldsymbol{\tau}_t\right)=\sum_{k=t}^{\infty} \gamma^k\left[r\left(\mathbf{s}_k, \mathbf{a}_k\right)-\alpha \mathbb{D}_{\mathrm{KL}}\left(\pi\left(\cdot \mid \mathbf{s}_k\right) \| \pi_0\left(\cdot \mid \mathbf{s}_k\right)\right)\right]$$

- Let the online and behavioral policies be Gaussian distributions

Exploding Gradients in KL-Regularized RL Objectives

Proposition 1 (informal).

- Let the objective function be given by

$$\tilde{R}\left(\boldsymbol{\tau}_t\right) = \sum_{k=t}^{\infty} \gamma^k \left[r\left(\mathbf{s}_k, \mathbf{a}_k\right) - \alpha \mathbb{D}_{\mathrm{KL}}\left(\pi\left(\cdot \mid \mathbf{s}_k\right) \| \pi_0\left(\cdot \mid \mathbf{s}_k\right)\right)\right]$$

- Let the online and behavioral policies be Gaussian distributions
- Let the online policy be parametrized by $\mathbf{a}_t = f_\phi\left(\epsilon_t; \mathbf{s}_t\right)$

Exploding Gradients in KL-Regularized RL Objectives

Proposition 1 (informal).

- Let the objective function be given by

$$\tilde{R}(\boldsymbol{\tau}_t) = \sum_{k=t}^{\infty} \gamma^k \left[r(\mathbf{s}_k, \mathbf{a}_k) - \alpha \mathbb{D}_{\mathrm{KL}}\left(\pi(\cdot \mid \mathbf{s}_k) \,\|\, \pi_0(\cdot \mid \mathbf{s}_k)\right)\right]$$

- Let the online and behavioral policies be Gaussian distributions
- Let the online policy be parametrized by $\mathbf{a}_t = f_\phi(\epsilon_t; \mathbf{s}_t)$
- Then:

$$\left|\hat{\nabla}_\phi J_\pi(\phi)\right| \rightarrow \infty \;\text{ as }\; \sigma_0^2 \rightarrow 0 \;\text{ with }\; \mathcal{O}\left(\sigma_0^{-2}(\mathbf{s}_t)\right)$$

Prevent predictive uncertainty collapse in behavioral policies

- Goal: increase predictive variance away from expert demonstrations

Prevent predictive uncertainty collapse in behavioral policies

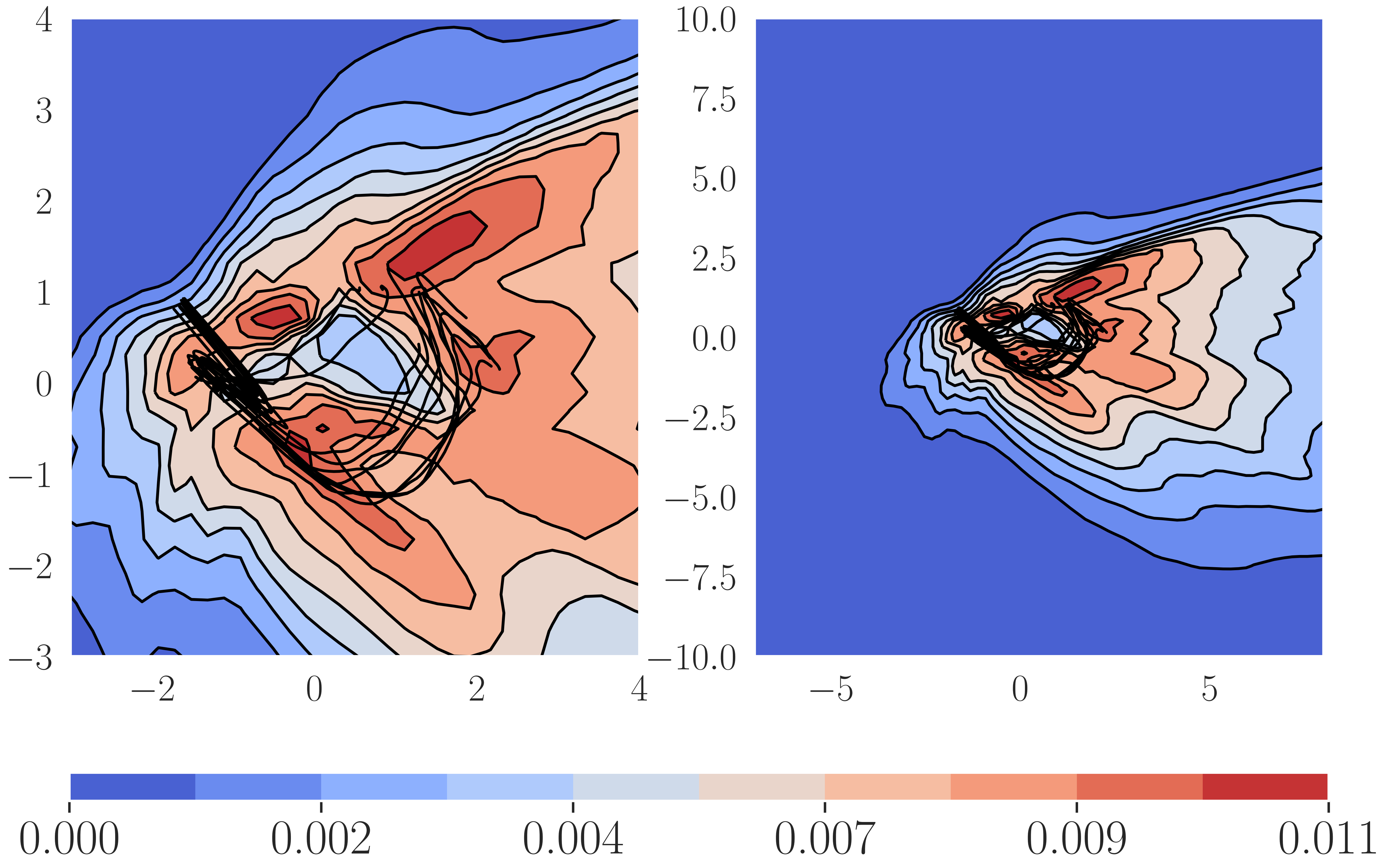
- Goal: increase predictive variance away from expert demonstrations
- Non-parametric Gaussian process behavioral policy

Prevent predictive uncertainty collapse in behavioral policies

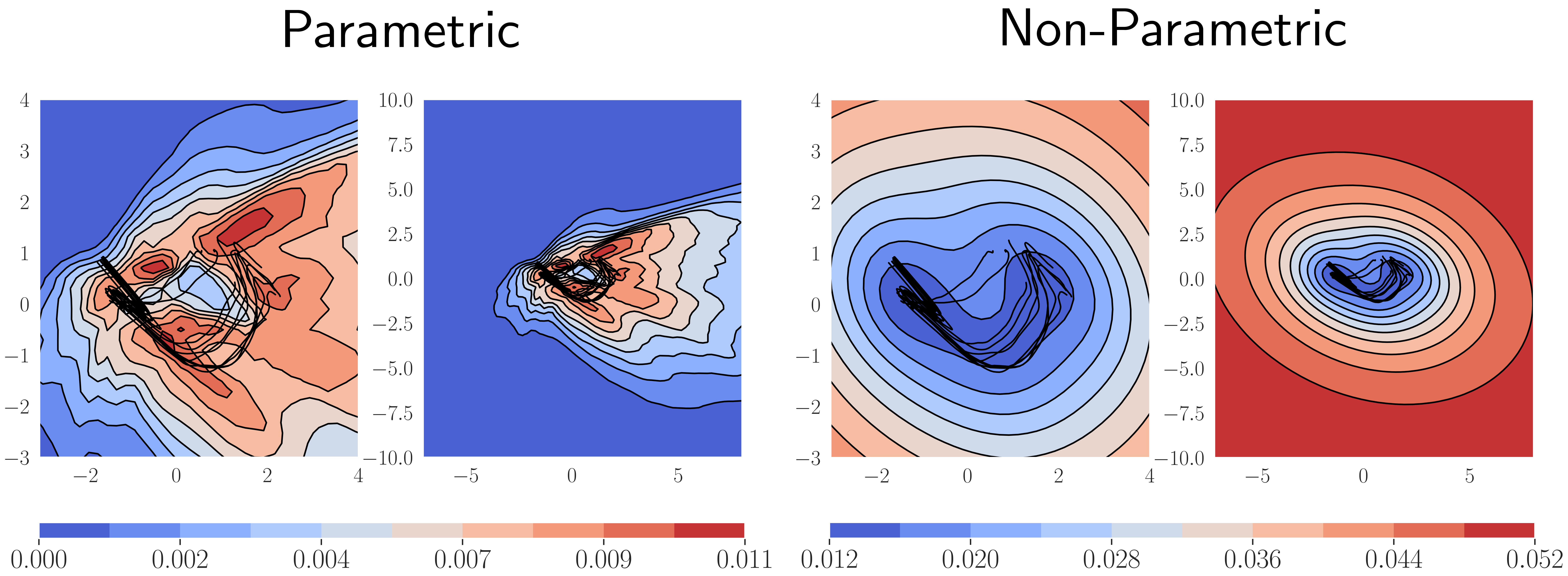
- Goal: increase predictive variance away from expert demonstrations
- Non-parametric Gaussian process behavioral policy
  - Prior: $A|s \sim \pi_0(\cdot|s) = \mathcal{GP}(m(s), k(s, s'))$
  - Posterior: $A|s, \mathcal{D}_0 \sim \pi_0(\cdot|s, \mathcal{D}_0) = \mathcal{GP}(\mu_0(s), \Sigma_0(s, s'))$
    - Mean: $\mu_0(s) = m(s) + k(s, \bar{S})(k(\bar{S}, \bar{S}))^{-1}(\bar{A} - m(\bar{A}))$
    - Covariance: $\Sigma_0(s, s') = k(s, s') + k(s, \bar{S})k(\bar{S}, \bar{S})^{-1}k(\bar{S}, s')$

Well-Calibrated Predictive Uncertainty

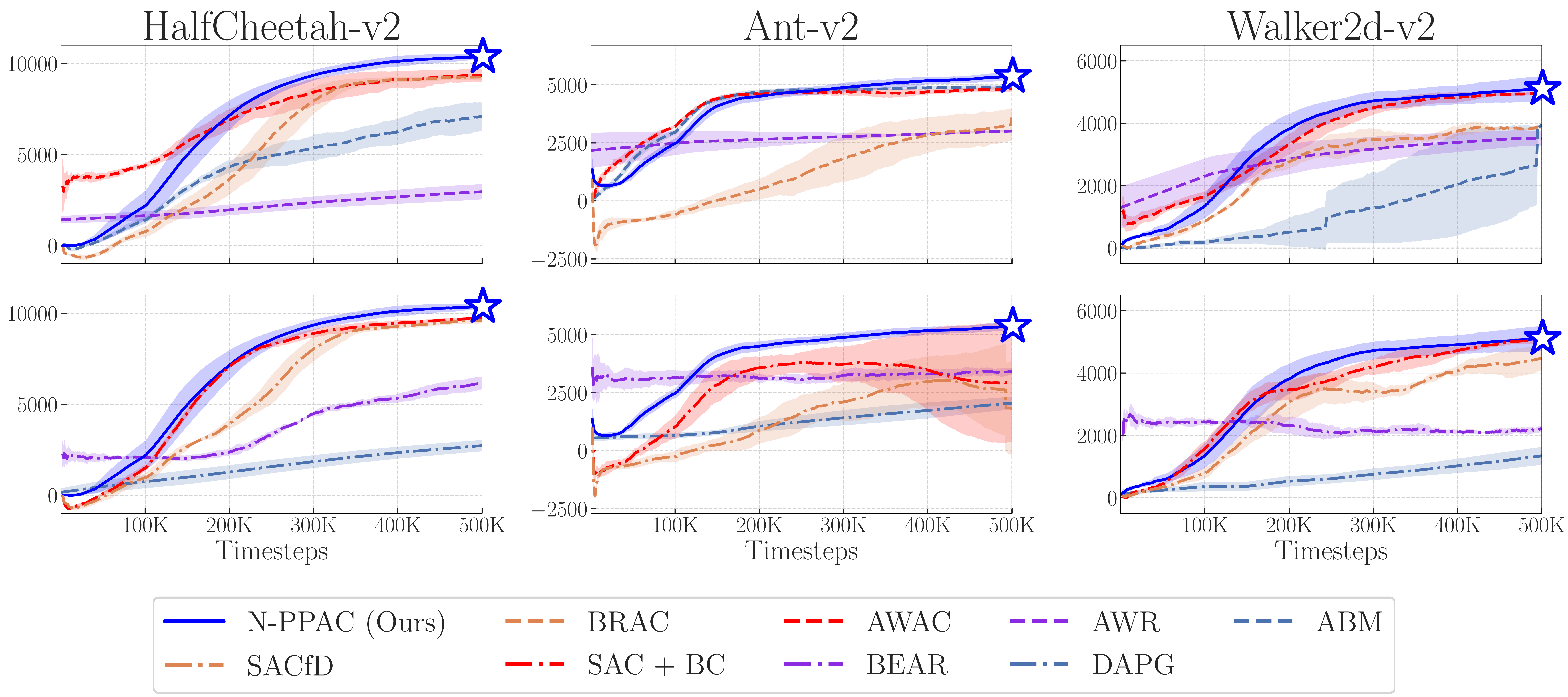
Parametric



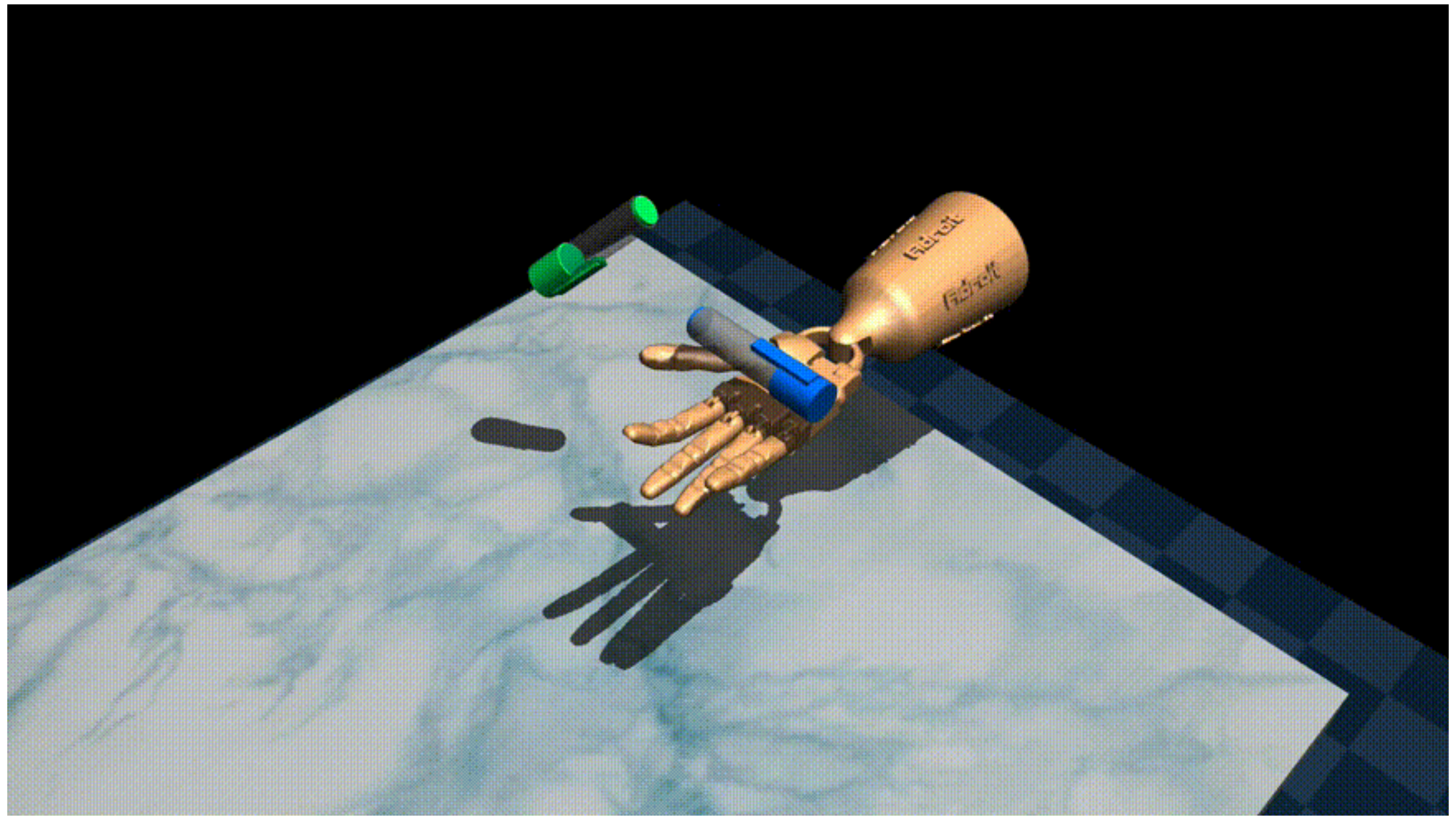
Well-Calibrated Predictive Uncertainty



MuJoCo Locomotion Tasks



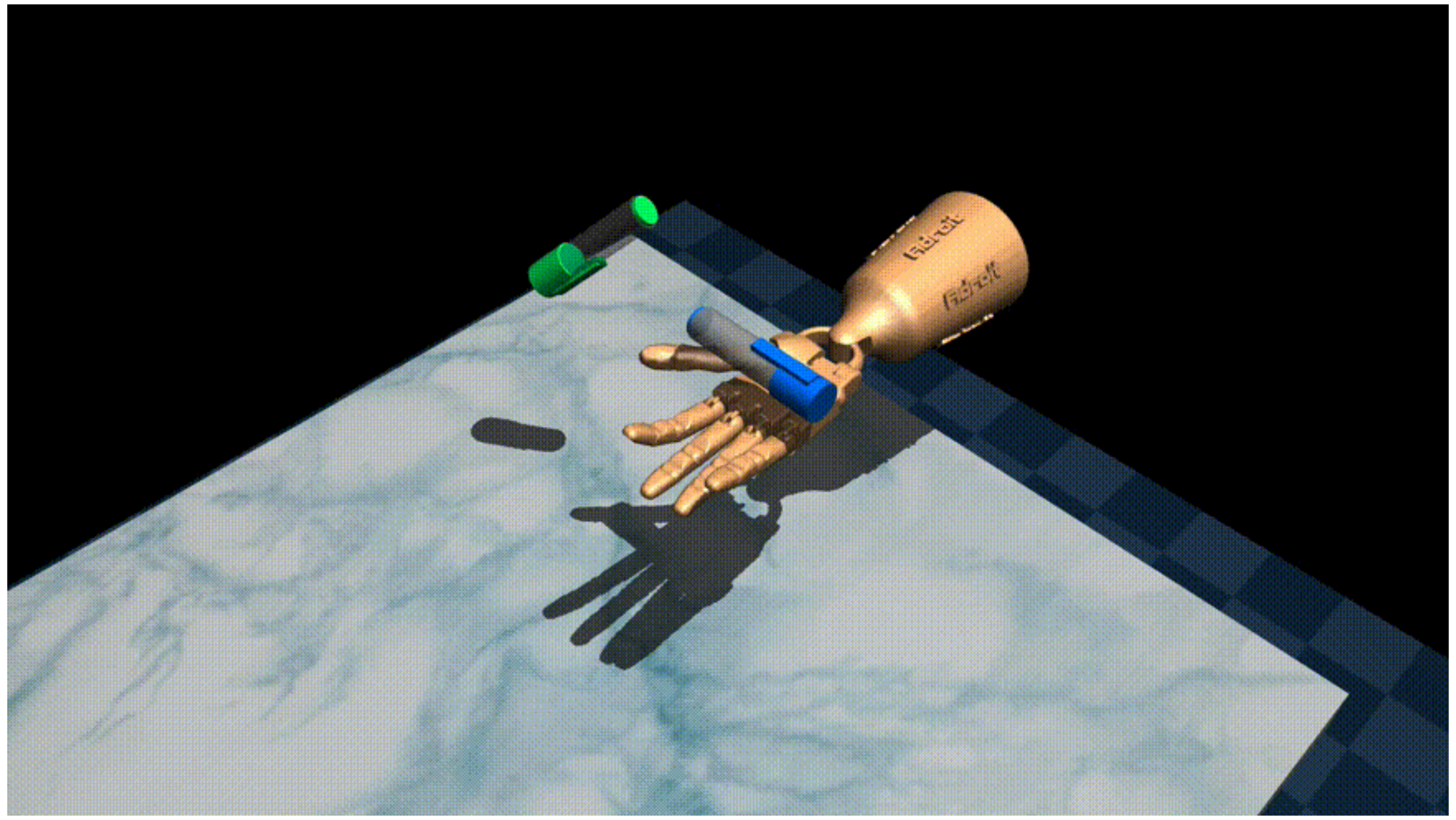
Dexterous Hand Manipulation Tasks



pen-binary-v0

door-binary-v0

Dexterous Hand Manipulation Tasks

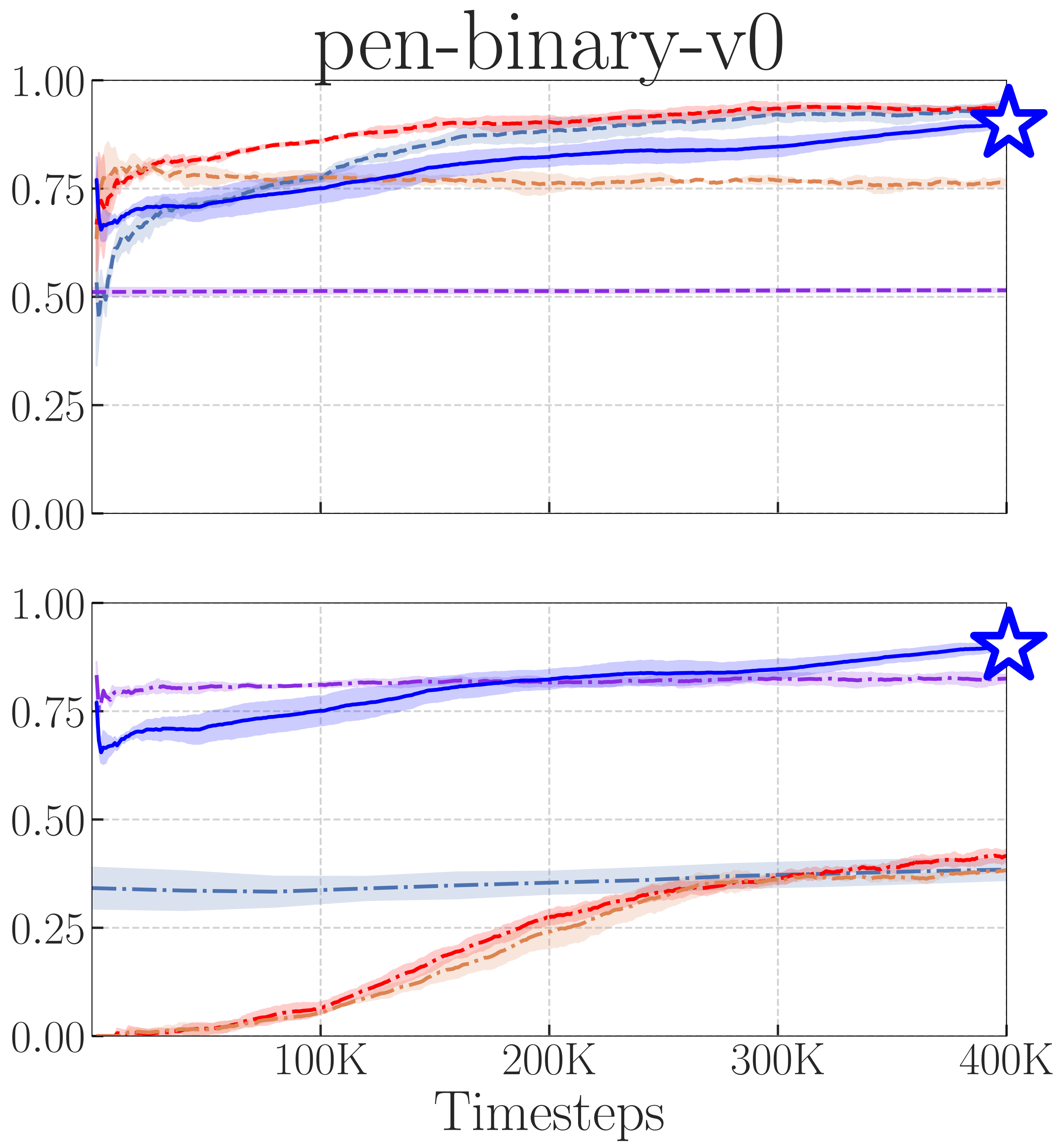


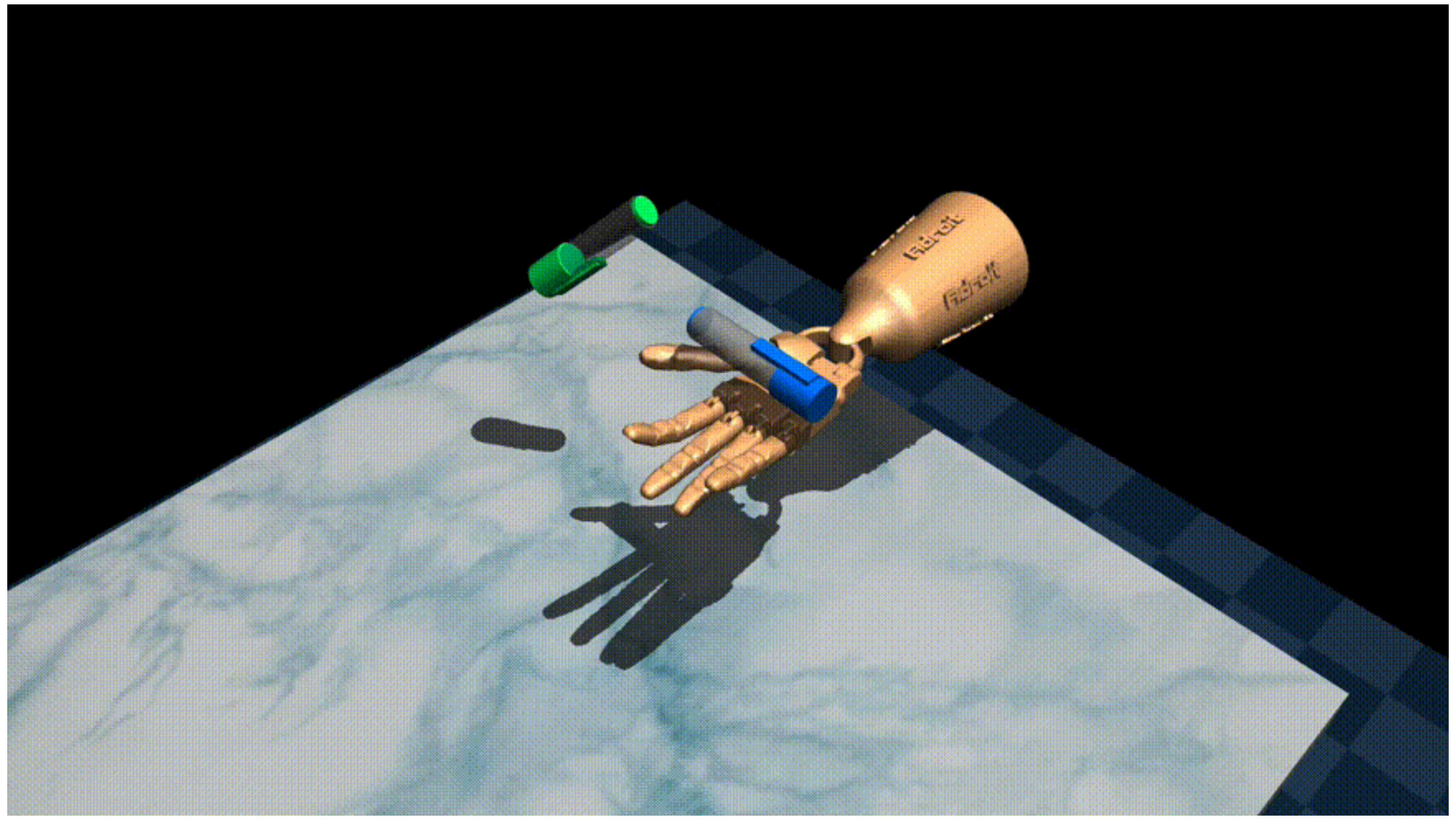
pen-binary-v0

door-binary-v0

KL-Regularized RL with Non-Parametric Behavioral Policies

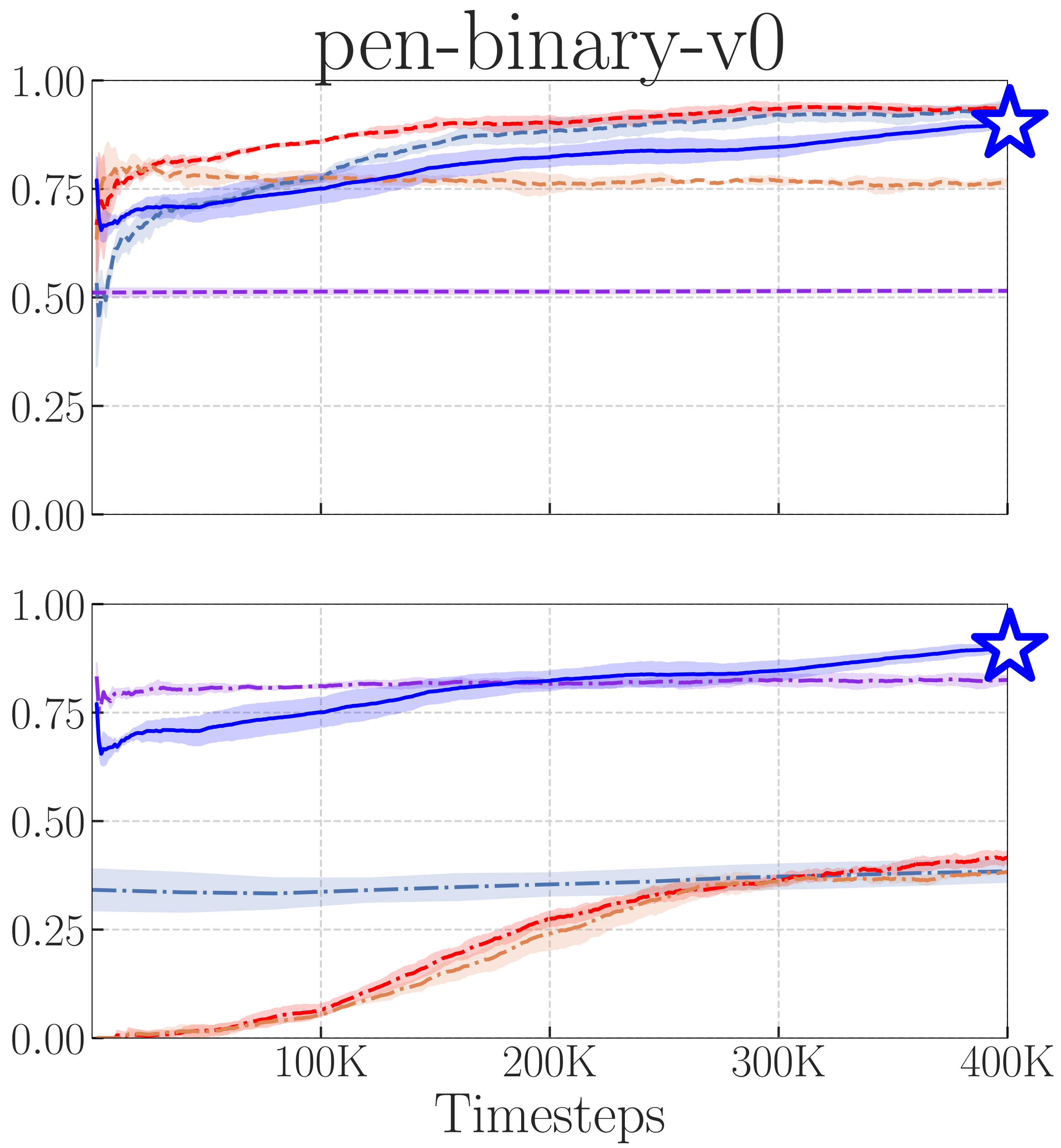
Dexterous Hand Manipulation: pen-binary-v0

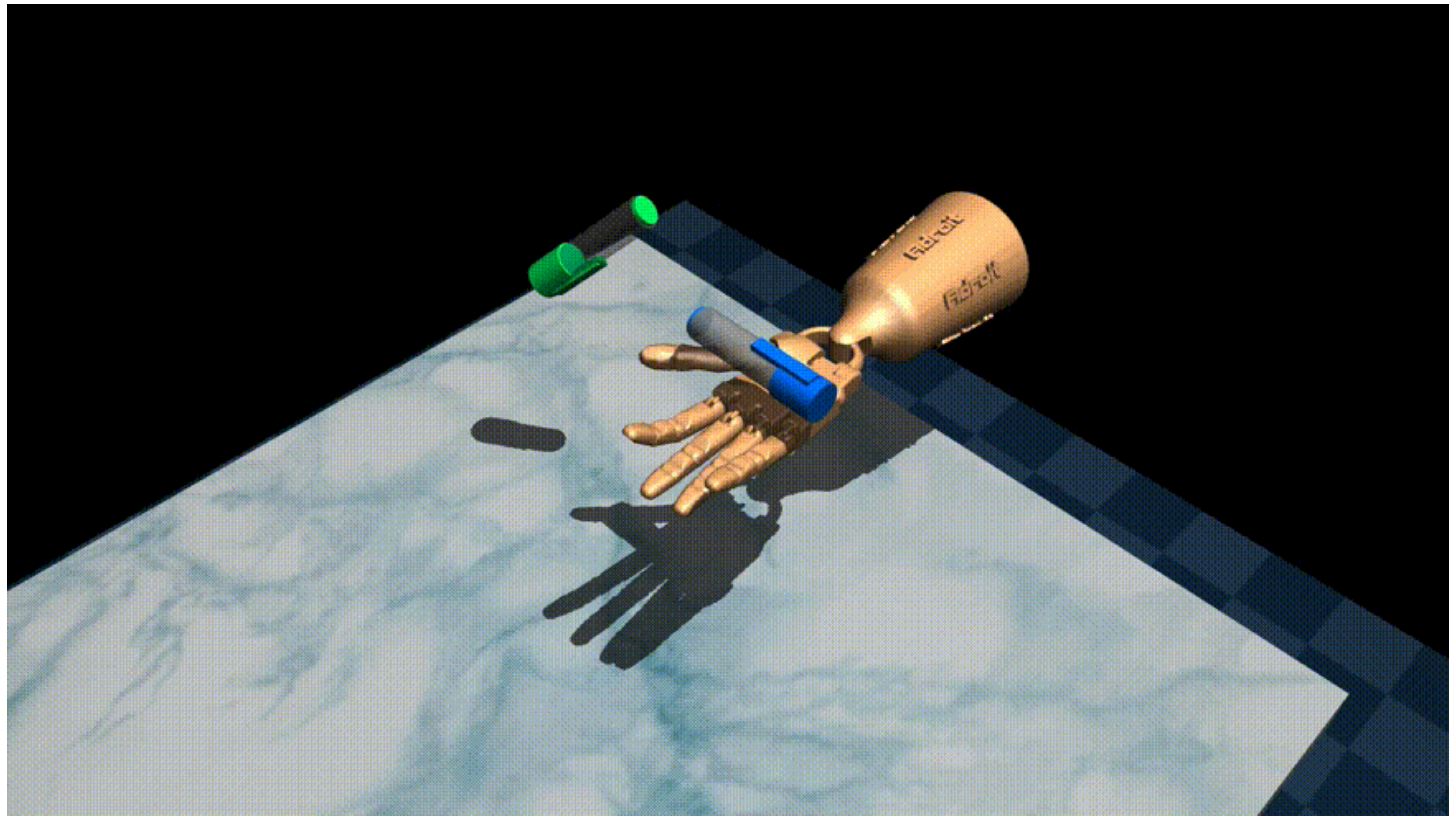




N-PPAC (Ours) | BRAC | AWAC | AWR | ABM | SACfD | SAC + BC | BEAR | DAPG

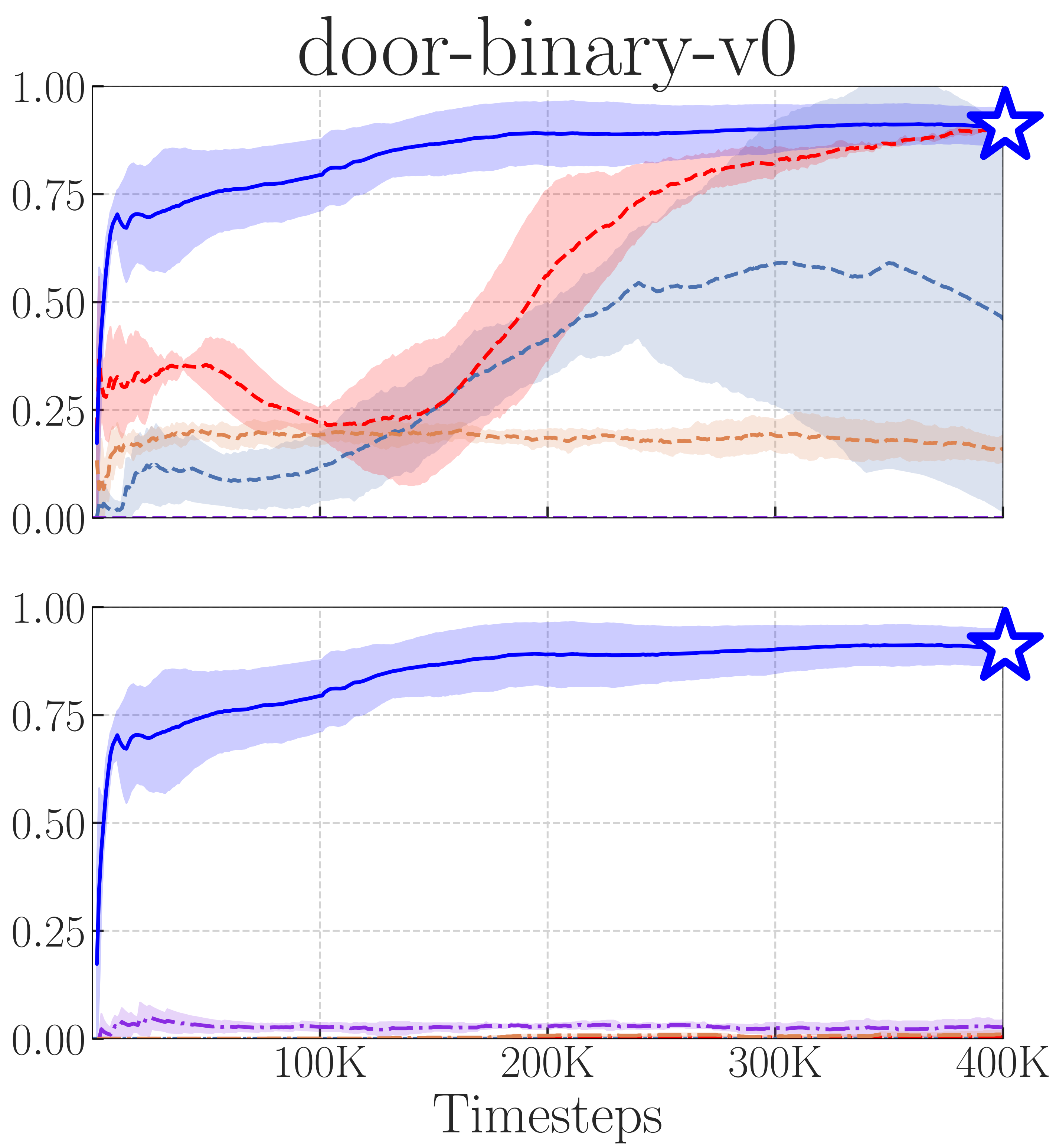
Dexterous Hand Manipulation: pen-binary-v0



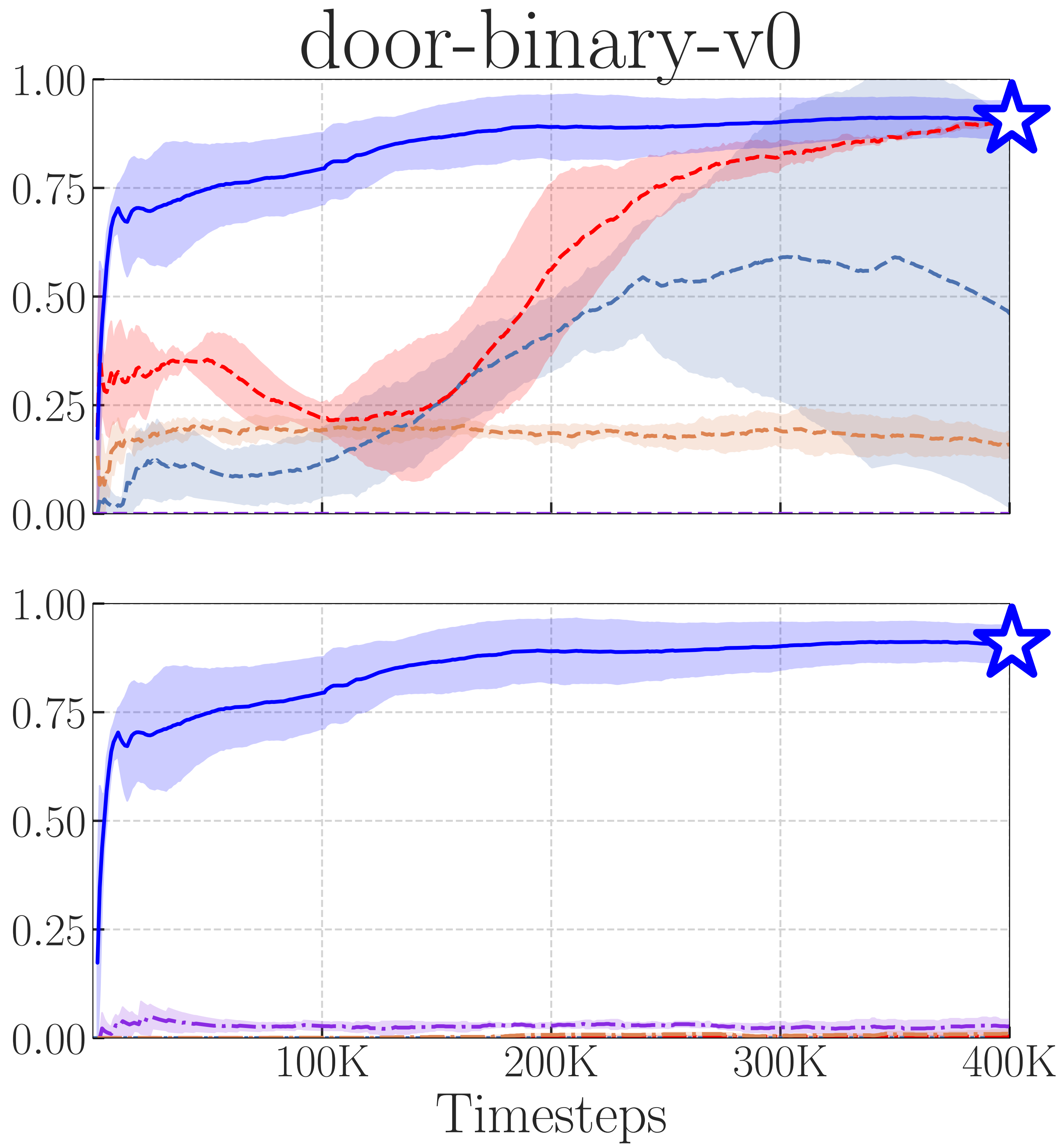


KL-Regularized RL with Non-Parametric Behavioral Policies

Dexterous Hand Manipulation Example: door-binary-v0



Dexterous Hand Manipulation Example: door-binary-v0

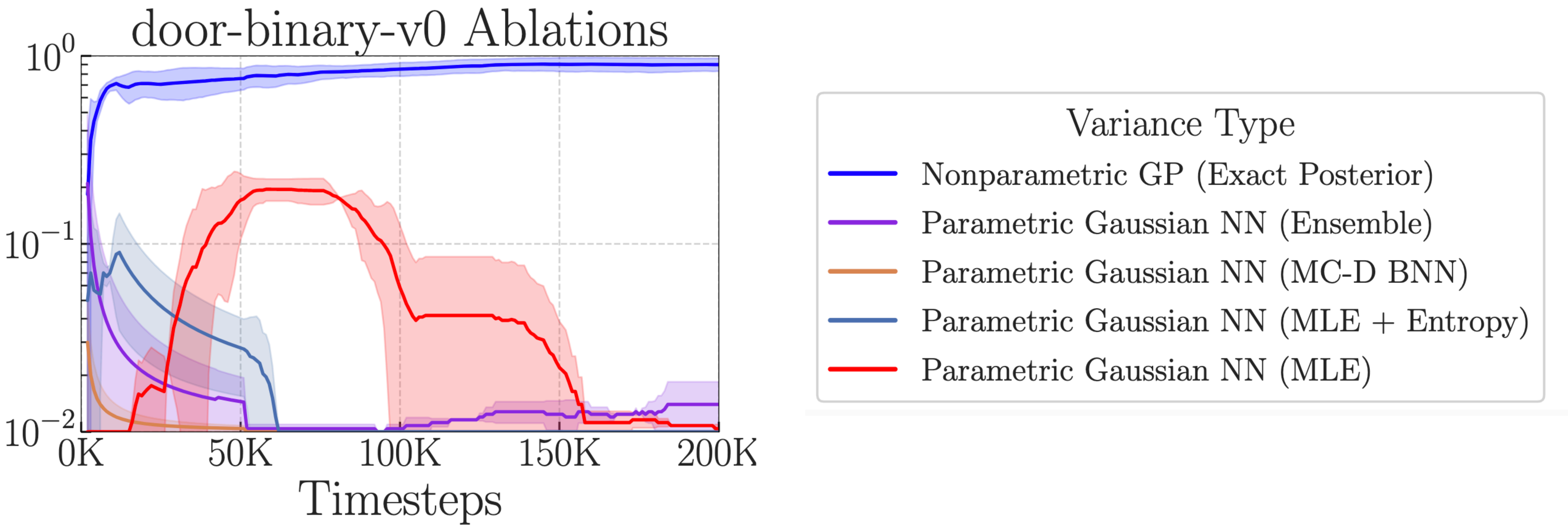


Takeaway

Fixing the pathological training dynamics in KL-regularized RL leads to state-of-the-art performance

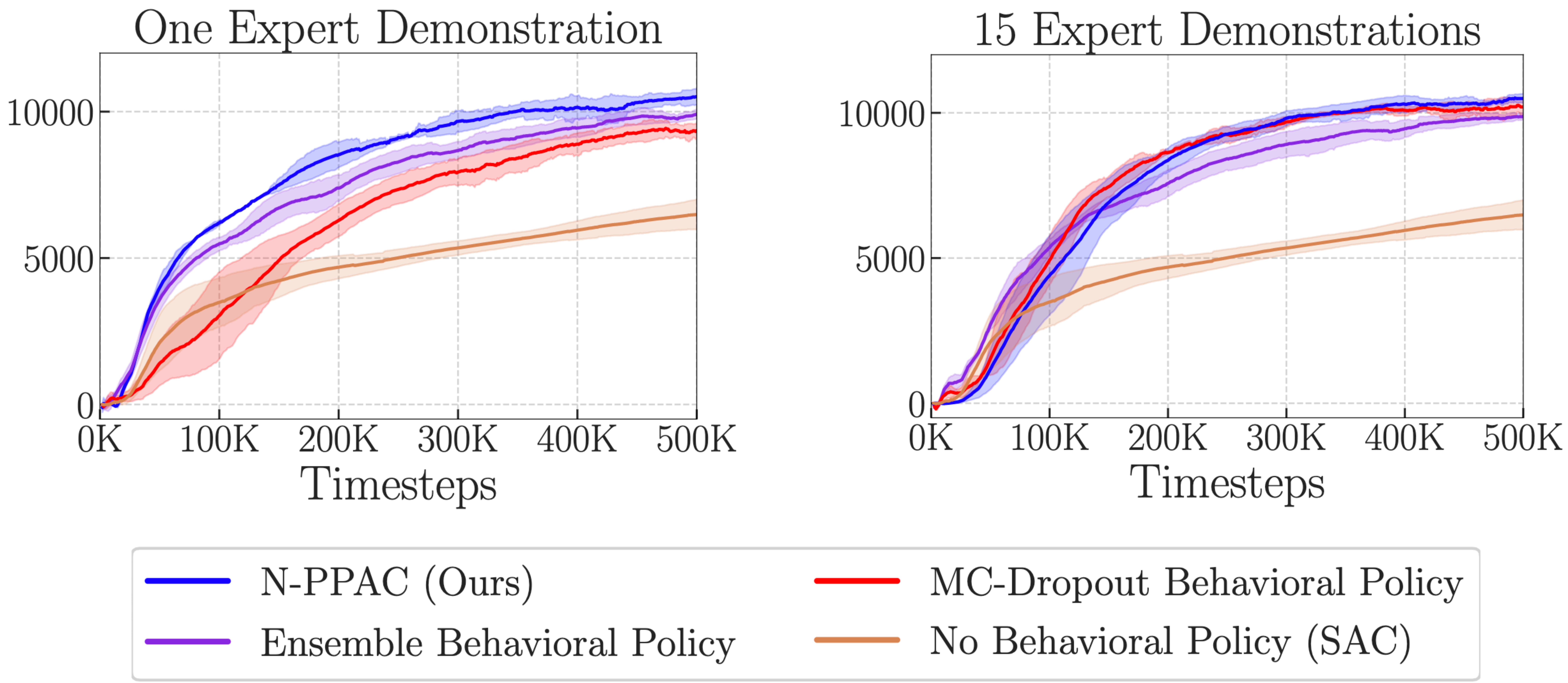
Could Better Uncertainty Quantification Fix the Pathology?

- Bayesian Neural Networks
- Deep Ensembles
- Lower-bounding Parametric Behavioral Policy Variance



Can a Single Expert Demonstration Be Sufficient?

MuJoCo Locomotion Example: HalfCheetah



Main Takeaways

Main Takeaways

KL-regularized RL can suffer from pathological behavior during training.

Main Takeaways

KL-regularized RL can suffer from **pathological behavior** during training.

Main Takeaways

KL-regularized RL can suffer from pathological behavior during training.

The pathology can be remedied by non-parametric behavioral policies.

Main Takeaways

KL-regularized RL can suffer from pathological behavior during training.

The pathology can be remedied by non-parametric behavioral policies.

Main Takeaways

KL-regularized RL can suffer from pathological behavior during training.

The pathology can be remedied by non-parametric behavioral policies.

Fixing the pathology leads to state-of-the-art policies and data-efficient online training.

Main Takeaways

KL-regularized RL can suffer from pathological behavior during training.

The pathology can be remedied by non-parametric behavioral policies.

Fixing the pathology leads to state-of-the-art policies and data-efficient online training.

Thank You!





Tim G. J. Rudner*   **Cong Lu***   Michael A. Osborne   Yarin Gal   Yee Whye Teh

Correspondence: tim.rudner@cs.ox.ac.uk

Project Website: https://sites.google.com/view/nppac