On Pathologies in KL-Regularized Reinforcement Learning from Expert Demonstrations

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

[†] University of Oxford * Equal contribution. Corresponding author: `tim.rudner@cs.ox.ac.uk`. @timrudner @cong_ml

TL;DR

(i) We show that KL-regularized RL with behavioral reference policies derived from expert demonstrations can suffer from **pathological training dynamics** caused by a collapse in the predictive variance of behavioral reference policies about states away from the expert demonstrations.

(ii) We demonstrate that this pathology can lead to instability and sub-optimality in online learning, but that it can be prevented by specifying **non-parametric behavioral reference policies** whose predictive variance is guaranteed not to collapse about previously unseen states.

(iii) We show that fixing the pathology allows KL-regularized RL to significantly outperform state-of-the-art approaches on a range of challenging locomotion and dexterous manipulation tasks.

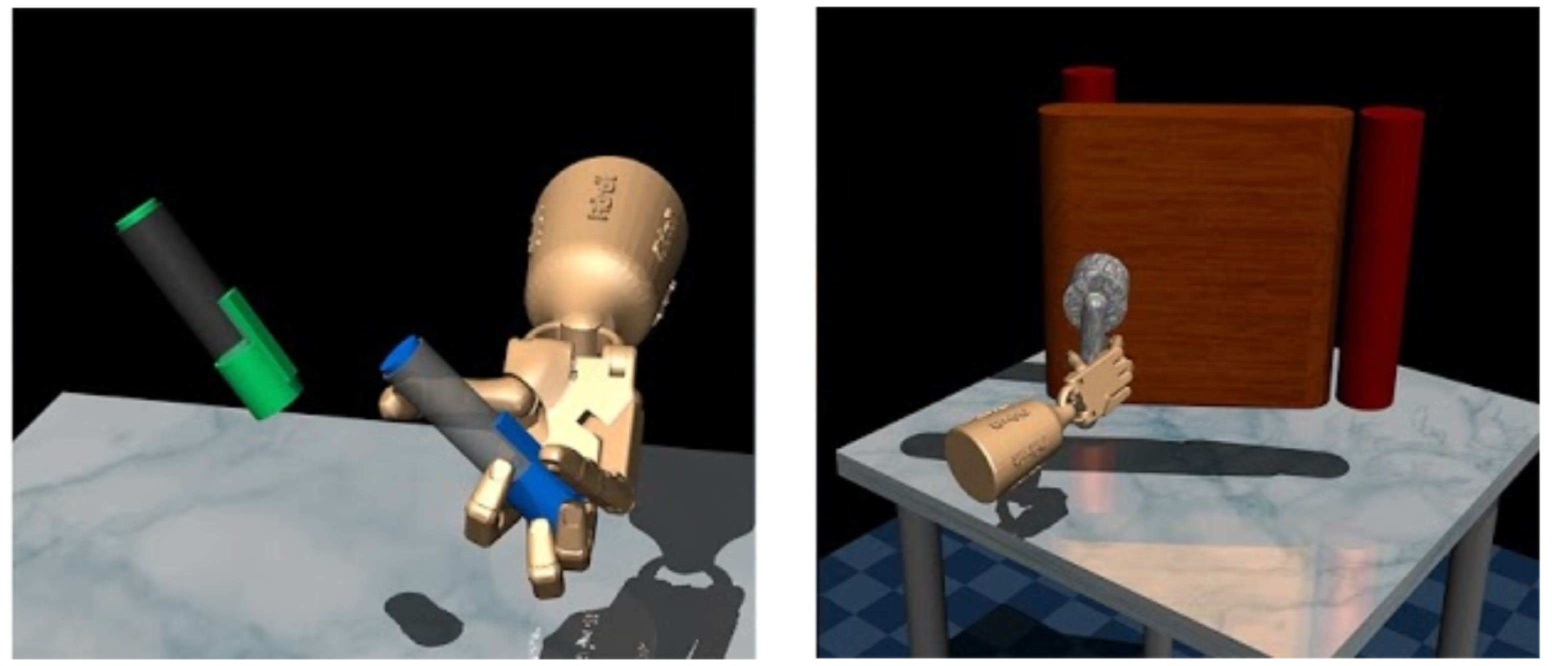

Figure 1: Dexterous hand manipulation tasks on which our fix leads to a significant acceleration in training and improvement in performance.

Reinforcement Learning & Behavioral Cloning

- An agent interacts with a discounted Markov Decision Process $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$. $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $p(\cdot \,|\, \mathbf{s}_t, \mathbf{a}_t)$ are the transition dynamics, $r(\mathbf{s}_t, \mathbf{a}_t)$ is the reward function, and γ is a discount factor. The agent learns a policy $\pi(\mathbf{a} \,|\, \mathbf{s})$.
- In behavioral cloning, a mapping $\pi_0 : \mathcal{S} \to \mathcal{A}$ is learned from an offline dataset $\mathcal{D}_0 = \{(\bar{\mathbf{s}}_i, \bar{\mathbf{a}}_i)\}_{i=1}^n$ of expert demonstrations, with n typically in the order of $1\text{k} - 10\text{k}$.

Identifying the Pathology

KL-Regularized Reinforcement Learning

- Given a reference policy π_0 and temperature α, KL-regularized RL augments the reward with a KL-penalty:

$$\sum_{t=0}^{\infty} \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[\gamma^t \left(r(\mathbf{s}_t, \mathbf{a}_t) - \alpha \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \,|\, \mathbf{s}_t) \,||\, \pi_0(\cdot \,|\, \mathbf{s}_t)) \right) \right]$$

- For the KL divergence to be defined, we require the support of π to be contained with in the support of π_0.
- Behaviorally cloned stochastic policies parameterized by a neural network via MLE experience a collapse in predictive variance about states off the offline data manifold, effectively leading to a loss in support between π and π_0.

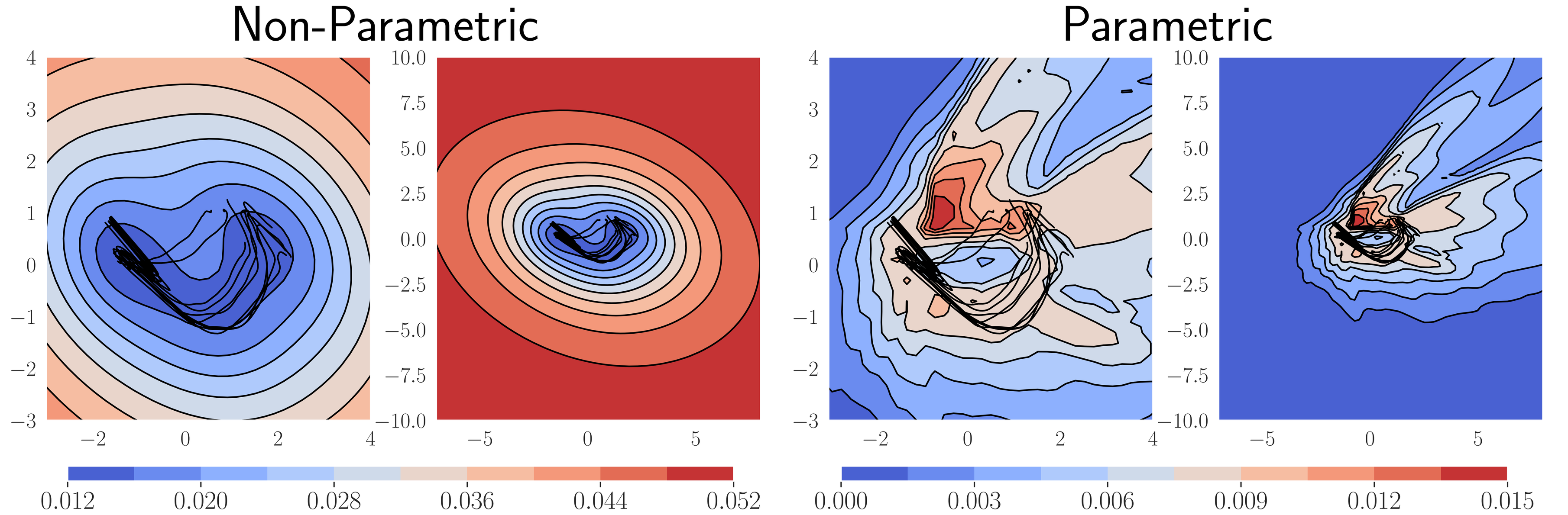

Figure 2: Predictive variances of non-parametric and parametric behavioral policies on a low dimensional representation of a 39-dimensional dexterous hand manipulation state space (door-binary-v0). **Left**: Non-parametric Gaussian process posterior behavioral policy $\pi_{\mathcal{GP}}(\cdot \,|\, \mathbf{s}, \mathcal{D}_0) = \mathcal{GP}(\boldsymbol{\mu}_0(\mathbf{s}), \boldsymbol{\Sigma}_0(\mathbf{s}, \mathbf{s}'))$. **Right**: Parametric neural network Gaussian behavioral policy $\pi_\psi(\cdot \,|\, \mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}_\psi(\mathbf{s}), \boldsymbol{\sigma}_\psi(\mathbf{s}))$. Expert trajectories $\mathcal{D}$ used to train the behavioral policies are shown in black.

Proposition 1 (Exploding Gradients in KL-Regularized RL;). *Let $\pi_0(\cdot \,|\, \mathbf{s})$ be a Gaussian behavioral policy with mean $\boldsymbol{\mu}_0(\mathbf{s}_t)$ and variance $\boldsymbol{\sigma}_0^2(\mathbf{s}_t)$, and let $\pi_\phi(\cdot \,|\, \mathbf{s})$ be an online policy with reparameterization $\mathbf{a}_t = f_\phi(\epsilon_t; \mathbf{s}_t)$ and random vector ϵ_t. Let the gradient of the policy loss with respect to the online policy's parameters ϕ be denoted $\nabla_\phi J_\pi(\phi)$. For fixed $|\mathbf{a}_t - \boldsymbol{\mu}_0|$, $|\hat{\nabla}_\phi J_\pi(\phi)| \to \infty$ as $\boldsymbol{\sigma}_0^2 \to 0$, when $\nabla_\phi f_\phi(\epsilon_t; \mathbf{s}_t) \neq 0$.*

Full paper: `timrudner.com/rl-pathologies`

Fixing the Pathology

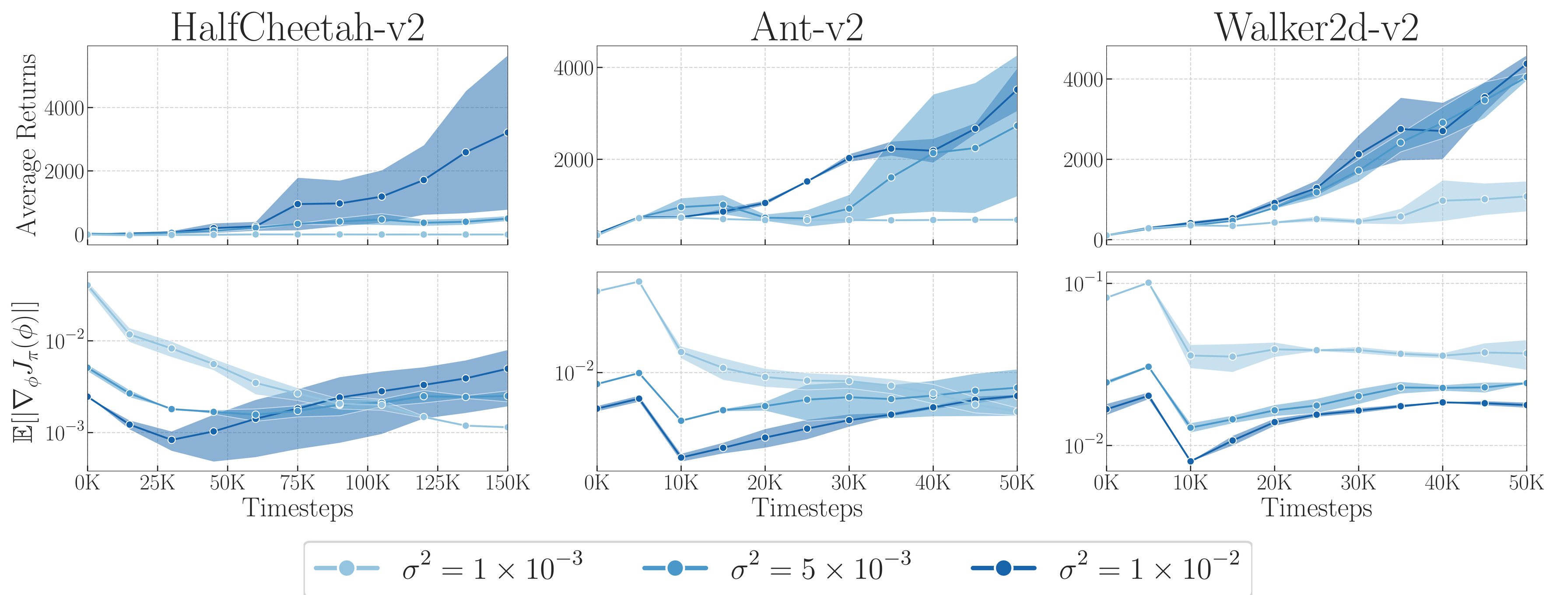

Figure 3: Effect of decrease in predictive variance on performance.

We **fix the pathology** by specifying a **non-parametric behavioral reference policy** whose variance is guaranteed not to collapse about unseen states.

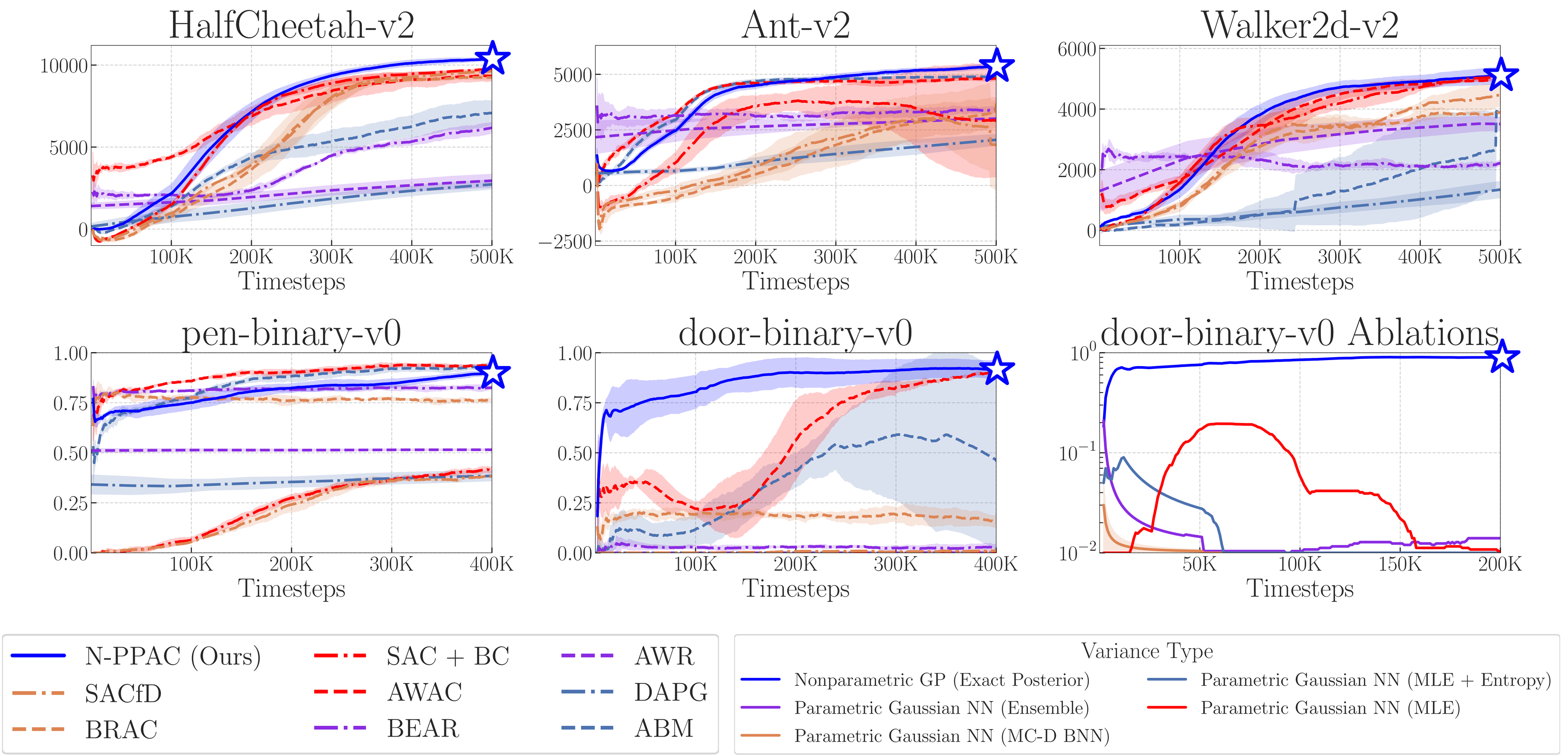

Figure 4: MuJoCo and dexterous hand manipulation tasks. **Bottom Right**: Comparison of behavioral reference policies with the same GP predictive mean but different predictive variances.

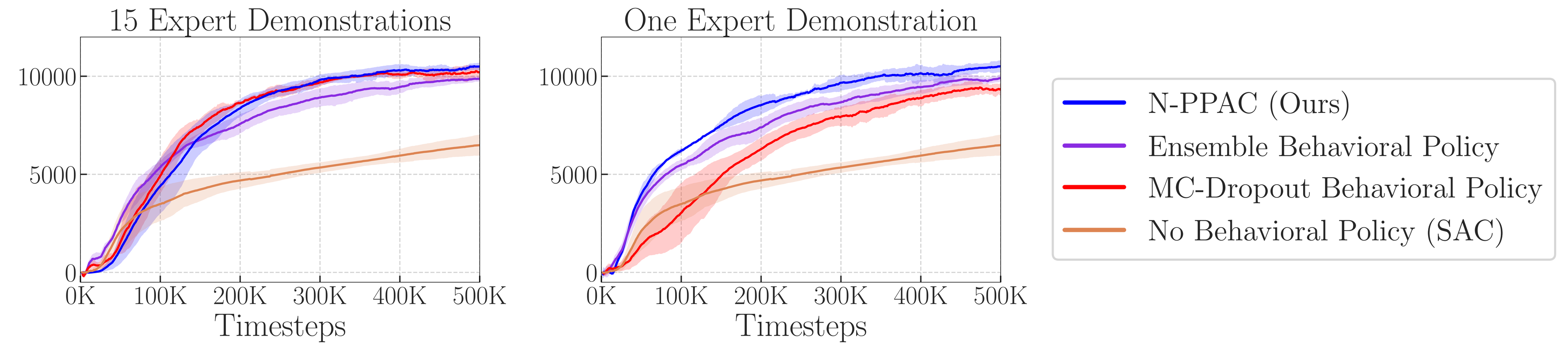

Figure 5: Online training returns for different numbers of expert demonstrations on the HalfCheetah-v2 environment using different behavioral policies.