On Pathologies in KL-Regularized Reinforcement Learning from Expert Demonstrations

Tim G. J. Rudner[*] [†]
University of Oxford

Cong Lu[*]
University of Oxford

Michael A. Osborne
University of Oxford

Yarin Gal
University of Oxford

Yee Whye Teh
University of Oxford

Abstract

KL-regularized reinforcement learning from expert demonstrations has proved successful in improving the sample efficiency of deep reinforcement learning algorithms, allowing them to be applied to challenging physical real-world tasks. However, we show that KL-regularized reinforcement learning with behavioral policies derived from expert demonstrations suffers from hitherto unrecognized pathological behavior that can lead to slow, unstable, and suboptimal online training. We show empirically that the pathology occurs for commonly chosen behavioral policy classes and demonstrate its impact on sample efficiency and online policy performance. Finally, we show that the pathology can be remedied by specifying *non-parametric* behavioral policies and that doing so allows KL-regularized RL to significantly outperform state-of-the-art approaches on a variety of challenging locomotion and dexterous hand manipulation tasks.

1 Introduction

Reinforcement learning [14, 23, 44, 45] is a powerful paradigm for learning complex behaviors. Unfortunately, many modern reinforcement learning algorithms require agents to carry out millions of interactions with their environment to learn desirable behaviors, making them of limited use for a wide range of practical applications that cannot be simulated [8, 27]. This limitation has motivated the study of algorithms that can incorporate pre-collected offline data into the training process either fully offline or with online exploration to improve sample efficiency and performance [2, 6, 15, 22, 50, 51]. An important and well-motivated subset of these methods are approaches for efficiently incorporating expert demonstrations into the learning process [5, 10, 17, 40].

KL-regularized RL is a particularly successful approach for doing so [3, 26, 28, 30, 42, 49]. In KL-regularized RL, the standard reinforcement learning objective is augmented by a Kullback-Leibler (KL) divergence penalty that expresses the dissimilarity between the online policy and a behavioral reference policy derived from expert demonstrations. The resulting regularized objective pulls the agent's online policy towards the behavioral policy, while also allowing it to improve upon the behavioral policy by exploring and interacting with the environment. Recent advances that leverage explicit or implicit KL-regularized objectives, such as BRAC [49], ABM [42], and AWAC [26], have shown that KL-regularized RL from expert demonstrations is able to significantly improve the sample efficiency of online training and solve challenging environments previously unsolved by standard deep RL algorithms.

[*]Equal contribution.
[†]Corresponding author: `tim.rudner@cs.ox.ac.uk`.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

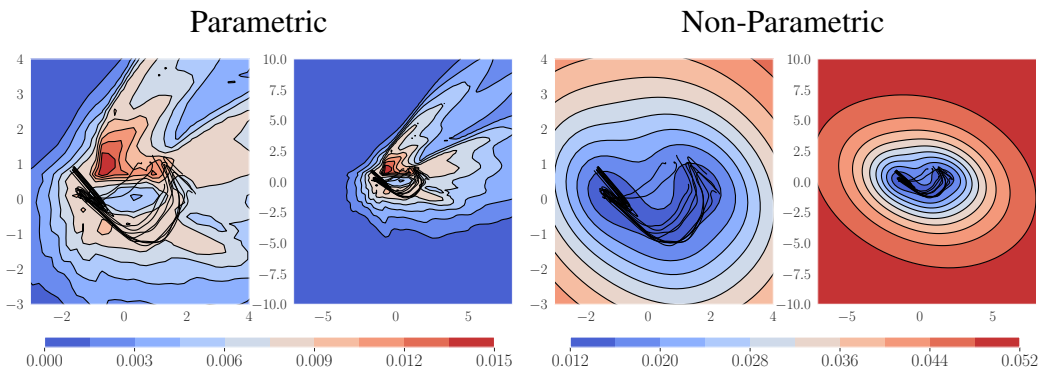

Figure 1: Predictive variances of non-parametric and parametric behavioral policies on a low dimensional representation (the first two principal components) of a 39-dimensional dexterous hand manipulation state space (see "door-binary-v0" in Figure 5). **Left**: Parametric neural network Gaussian behavioral policy $\pi_\psi(\cdot \mid \mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}_\psi(\mathbf{s}), \boldsymbol{\sigma}^2_\psi(\mathbf{s}))$. **Right**: Non-parametric Gaussian process posterior behavioral policy $\pi_{\mathcal{GP}}(\cdot \mid \mathbf{s}, \mathcal{D}_0) = \mathcal{GP}(\boldsymbol{\mu}_0(\mathbf{s}), \boldsymbol{\Sigma}_0(\mathbf{s}, \mathbf{s}'))$. Expert trajectories $\mathcal{D}$ used to train the behavioral policies are shown in black. The GP predictive variance is well-calibrated: It is small near the expert trajectories and large in other parts of the state space. In contrast, the neural network predictive variance is poorly calibrated: It is relatively small on the expert trajectories, and collapses to near zero elsewhere. Note the significant difference in scales.

Contributions. In this paper, we show that despite empirical successes, KL-regularized RL from expert demonstrations suffers from previously unrecognized pathologies that lead to instability and sub-optimality in online learning. To summarize, our core contributions are as follows:

- We illustrate empirically that commonly used classes of parametric behavioral policies experience a collapse in predictive variance about states away from the expert demonstrations.
- We demonstrate theoretically and empirically that KL-regularized RL algorithms suffer from pathological behavior in online training when regularized against behavioral policies that exhibit such a collapse in predictive variance.
- We show that the pathology can be remedied by *non-parametric* behavioral policies, whose predictive variances are well-calibrated and also guaranteed not to collapse about previously unseen states, and that fixing the pathology results in online policies that significantly outperform state-of-the-art approaches on a range of challenging locomotion and dexterous hand manipulation tasks.

The left panel of Figure 1 shows an example of the collapse in predictive variance away from the expert trajectories in parametric behavioral policies. In contrast, the right panel of Figure 1 shows the predictive variance of a non-parametric behavioral policy, which—unlike in the case of the parametric policy—increases off the expert trajectories. By avoiding the pathology, we obtain a stable and reliable approach to sample-efficient reinforcement learning, applicable to a wide range of reinforcement learning algorithms that leverage KL-regularized objectives.[3]

2 Background

We consider the standard reinforcement learning setting where an agent interacts with a discounted Markov Decision Process (MDP) [44] given by a 5-tuple $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $p(\cdot \mid \mathbf{s}_t, \mathbf{a}_t)$ are the transition dynamics, $r(\mathbf{s}_t, \mathbf{a}_t)$ is the reward function, and γ is a discount factor. $\rho_\pi(\boldsymbol{\tau}_t)$ denotes the state–action trajectory distribution from time t induced by a policy $\pi(\cdot \mid \mathbf{s}_t)$. The discounted return from time step t is given by $R(\boldsymbol{\tau}_t) = \sum_{k=t}^{\infty} \gamma^k r(\mathbf{s}_k, \mathbf{a}_k)$ for $t \in \mathbb{N}_0$. The standard reinforcement learning objective to be maximized is the expected discounted return $J_\pi(\boldsymbol{\tau}_0) = \mathbb{E}_{\rho_\pi(\boldsymbol{\tau}_0)}[R(\boldsymbol{\tau}_0)]$ under the policy trajectory distribution.

2.1 Improving and Accelerating Online Training via Behavioral Cloning

We consider settings where we have a set of expert demonstrations *without reward*, $\mathcal{D}_0 = \{(\mathbf{s}_n, \mathbf{a}_n)\}_{n=1}^N = \{\bar{\mathbf{S}}, \bar{\mathbf{A}}\}$, which we would like to use to speed up and improve online learning [5, 40]. One approach for turning the expert trajectories into a policy is behavioral cloning [1, 4]

[3] Code and visualizations of our results can be found at `https://sites.google.com/view/nppac`.

which involves learning a mapping from states in the expert demonstrations to their corresponding actions, that is, $\pi_0 : \mathcal{S} \to \mathcal{A}$. As such, behavioral cloning does not assume or require access to a reward function and only involves learning a mapping from states to action in a supervised fashion.

Since expert demonstrations are costly to obtain and often only available in small number, behavioral cloning alone is typically insufficient for agents to learn good policies in complex environments and has to be complemented by a method that enables the learner to build on the cloned behavior by interacting with the environment. A particularly successful and popular class of algorithms used for incorporating behavioral policies into online training is KL-regularized reinforcement learning [9, 36, 41, 46].

2.2 KL-Regularized Objectives in Reinforcement Learning

KL-regularized RL modifies the standard reinforcement learning objective by augmenting the return with a negative Kullback-Leibler (KL) divergence term from the learned policy π to a reference policy π_0, given a temperature parameter α. The resulting discounted return from time step $t \in \mathbb{N}_0$ is then given by

$$\tilde{R}(\boldsymbol{\tau}_t) = \sum_{k=t}^{\infty} \gamma^k \big[r(\mathbf{s}_k, \mathbf{a}_k) - \alpha \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \,|\, \mathbf{s}_k) \parallel \pi_0(\cdot \,|\, \mathbf{s}_k)) \big] \tag{1}$$

and the reinforcement learning objective becomes $\tilde{J}_\pi(\boldsymbol{\tau}_0) = \mathbb{E}_{\rho_\pi(\boldsymbol{\tau}_0)}[\tilde{R}(\boldsymbol{\tau}_0)]$. When the reference policy π_0 is given by a uniform distribution, we recover the entropy-regularized reinforcement learning objective used in Soft Actor–Critic (SAC) [12] up to an additive constant.

Under a uniform reference policy π_0, the resulting objective encourages exploration, while also choosing high-reward actions. In contrast, when π_0 is non-uniform, the agent is discouraged to explore areas of the state space $\mathcal{S}$ where the variance of $\pi_0(\cdot \,|\, \mathbf{s})$ is low (that is, more certain) and encouraged to explore areas of the state space where the variance of $\pi_0(\cdot \,|\, \mathbf{s})$ is high. The KL-regularized reinforcement learning objective can be optimized via policy–gradient and actor–critic algorithms.

2.3 KL-Regularized Actor–Critic

An optimal policy π that maximizes the expected KL-augmented discounted return $\tilde{J}_\pi$ can be learned by directly optimizing the policy gradient $\nabla_\pi \tilde{J}_\pi$. However, this policy gradient estimator exhibits high variance, which can lead to unstable learning. Actor–critic algorithms [7, 16, 31, 37] attempt to reduce this variance by making use of the state value function $V^\pi(\mathbf{s}_t) = \mathbb{E}_{\rho_\pi(\boldsymbol{\tau}_t)}[\tilde{R}(\boldsymbol{\tau}_t) \,|\, \mathbf{s}_t]$ or the state–action value function $Q^\pi(\mathbf{s}_t, \mathbf{a}_t) = \mathbb{E}_{\rho_\pi(\boldsymbol{\tau}_t)}[\tilde{R}(\boldsymbol{\tau}_t) \,|\, \mathbf{s}_t, \mathbf{a}_t]$ to stabilize training.

Given a reference policy $\pi_0(\mathbf{a}_t \,|\, \mathbf{s}_t)$, the state value function can be shown to satisfy the modified Bellman equation

$$V^\pi(\mathbf{s}_t) \doteq \mathbb{E}_{\mathbf{a}_t \sim \pi(\cdot|\mathbf{s}_t)}[Q^\pi(\mathbf{s}_t, \mathbf{a}_t)] - \alpha \mathbb{D}_{\mathrm{KL}}\big(\pi(\cdot \,|\, \mathbf{s}_t) \,||\, \pi_0(\cdot \,|\, \mathbf{s}_t)\big)$$

with a recursively defined Q-function

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t) \doteq r(\mathbf{s}_t, \mathbf{a}_t) + \gamma \, \mathbb{E}_{\mathbf{s}_{t+1} \sim p(\cdot|\mathbf{s}_t, \mathbf{a}_t)}[V^\pi(\mathbf{s}_{t+1})].$$

Instead of directly optimizing the objective function $\tilde{J}_\pi$ via the policy gradient, actor–critic methods alternate between policy evaluation and policy improvement [7, 12]:

Policy Evaluation. During the policy evaluation step, $Q^\pi_\theta(\mathbf{s}, \mathbf{a})$, parameterized by parameters θ, is trained by minimizing the Bellman residual

$$J_Q(\theta) \doteq \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mathcal{D}} \Big[\big(Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - (r(\mathbf{s}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim p(\cdot|\mathbf{s}_t, \mathbf{a}_t)}[V_{\bar{\theta}}(\mathbf{s}_{t+1})]) \big)^2 \Big], \tag{2}$$

where $\mathcal{D}$ is a replay buffer and $\bar{\theta}$ is a stabilizing moving average of parameters.

Policy Improvement. In the policy improvement step, the policy π_ϕ, parameterized by parameters ϕ, is updated towards the exponential of the KL-augmented Q-function,

$$J_\pi(\phi) \doteq \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[\alpha \mathbb{D}_{\mathrm{KL}}(\pi_\phi(\cdot \,|\, \mathbf{s}_t) \parallel \pi_0(\cdot \,|\, \mathbf{s}_t)) \right] - \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[\mathbb{E}_{\mathbf{a}_t \sim \pi_\phi(\cdot|\mathbf{s}_t)} \left[Q_\theta(\mathbf{s}_t, \mathbf{a}_t) \right] \right], \tag{3}$$

with states sampled from a replay buffer $\mathcal{D}$ and actions sampled from the parameterized online policy π_ϕ. The following sections will focus on the policy improvement objective and how certain types of references policies can lead to pathologies when optimizing $J_\pi(\phi)$ with respect to ϕ.

3 Identifying the Pathology

In this section, we investigate the effect of KL-regularization on the training dynamics. To do so, we first consider the properties of the KL divergence to identify a potential failure mode for KL-regularized reinforcement learning. Next, we consider parametric Gaussian behavioral policies commonly used in practice for continuous control tasks [12, 49] and show that for Gaussian behavioral policies with small predictive variance, the policy improvement objective suffers from exploding gradients with respect to the policy parameters ϕ. We confirm that this failure occurs empirically and demonstrate that it results in slow, unstable, and suboptimal online training. Lastly, we show that various regularization techniques used for estimating behavioral policies are unable to prevent this failure and also lead to suboptimal online policies.

3.1 When Are KL-Regularized RL Objectives Meaningful?

We start by considering the properties of the KL divergence and discuss how these properties can lead to potential failure modes in KL-regularized objectives. A well-known property of KL-regularized objectives in the variational inference literature is the occurrence of singularities when one distribution loses effective support with the other.

To illustrate this problem, we consider the case of Gaussian behavioral and learned policies, as are often chosen in practice. In theory, the KL divergence between two full Gaussian distributions is always finite and well-defined. Hence, we might hope KL-regularized RL with Gaussian behavioral and online policies to be unaffected by the failure mode described above. However, a Gaussian online policy $\pi_\phi(\cdot \mid \mathbf{s}_t)$ loses effective support with $\pi_0(\cdot \mid \mathbf{s}_t)$ as the predictive variance $\boldsymbol{\sigma}_0^2(\mathbf{s}_t)$ tends to zero, and hence: $\mathbb{D}_{\mathrm{KL}}(\pi_\phi(\cdot \mid \mathbf{s}_t) \parallel \pi_0(\cdot \mid \mathbf{s}_t)) \to \infty$ as $\boldsymbol{\sigma}_0^2(\mathbf{s}_t) \to 0$. In other words, in the limit of zero variance in the behavioral policy, the distribution becomes degenerate and will cause the KL divergence to be infinite [24]. While in practice, no behavioral policy will operate in the limit of zero variance, the functional form of the KL divergence between (univariate) Gaussians,

$$\mathbb{D}_{\mathrm{KL}}(\pi_\phi(\cdot \mid \mathbf{s}_t) \parallel \pi_0(\cdot \mid \mathbf{s}_t)) \propto \log \frac{\boldsymbol{\sigma}_0(\mathbf{s}_t)}{\boldsymbol{\sigma}_\phi(\mathbf{s}_t)} + \frac{\boldsymbol{\sigma}_\phi^2(\mathbf{s}_t) + (\boldsymbol{\mu}_\phi(\mathbf{s}_t) - \boldsymbol{\mu}_0(\mathbf{s}_t))^2}{2\boldsymbol{\sigma}_0^2(\mathbf{s}_t)},$$

implies a continuous, quadratic increase in the magnitude of the divergence as $\boldsymbol{\sigma}_0(\mathbf{s}_t)$ decreases, further exacerbated by a large difference in predictive means, $|\boldsymbol{\mu}_\phi(\mathbf{s}_t) - \boldsymbol{\mu}_0(\mathbf{s}_t)|$.

As a result, for Gaussian behavioral policies $\pi_0(\cdot \mid \mathbf{s}_t)$ that assign very low probability to sets of points in sample space far away from the distribution's mean $\boldsymbol{\mu}_0(\mathbf{s}_t)$, computing the KL divergence can result in divergence values so large to cause numerical instabilities and arithmetic overflow. Hence, even for a suitably chosen behavioral policy class, we see that vanishingly small behavioral policy predictive variances—which effectively cause the support of the online policy to not be contained in the support of the behavioral policy—can result in the KL divergence blowing up at evaluation points far away from states in the expert demonstrations.

One way of remedying this failure mode in practice would be to constrain the output of the variance network (e.g., by adding a small constant bias). However, to encourage effectively learning guaranteeing a flat, non-zero predictive variance far from the offline dataset is not enough. Rather, we require calibrated predictive variance estimates that *increase on data points away from the expert demonstrations* and can thus reduce the regularization towards poor predictive means on such points. Another solution could be to use distributions with heavier tails, e.g., a Laplace distribution over a Gaussian to avoid the quadratic scaling on the difference of means. However, we show in Appendix B.3 that doing so also does not fully resolve the pathology.

In the following sections, we explain how an explosion in $\mathbb{D}_{\mathrm{KL}}(\pi_\phi(\cdot \mid \mathbf{s}_t) \parallel \pi_0(\cdot \mid \mathbf{s}_t))$ caused by small $\boldsymbol{\sigma}_0^2(\mathbf{s}_t)$ affects the gradients of $J_\pi(\phi)$ in KL-regularized RL and discuss of how and why $\boldsymbol{\sigma}_0^2(\mathbf{s}_t)$ may tend to zero in practice.

3.2 Exploding Gradients in KL-Regularized Reinforcement Learning Objectives

To understand how small predictive variances in behavioral policies can affect—and possibly destabilize—online training in KL-regularized RL, we consider the contribution of the behavioral policy's variance to the gradient of the policy objective in Equation (3). Compared to entropy-regularized actor–critic methods (SAC, Haarnoja et al. [12]), which implicitly regularize against a uniform policy,

the gradient estimator $\hat{\nabla}_\phi J_\pi(\phi)$ in KL-regularized RL gains an extra scaling term $\nabla_{\mathbf{a}_t} \log \pi_0(\mathbf{a}_t \mid \mathbf{s}_t)$, the gradient of the prior log-density evaluated actions $\mathbf{a}_t$ sampled from the online policy:

Proposition 1 (Exploding Gradients in KL-Regularized RL)**.** *Let $\pi_0(\cdot \mid \mathbf{s})$ be a Gaussian behavioral policy with mean $\boldsymbol{\mu}_0(\mathbf{s}_t)$ and variance $\boldsymbol{\sigma}_0^2(\mathbf{s}_t)$, and let $\pi_\phi(\cdot \mid \mathbf{s})$ be an online policy with reparameterization $\mathbf{a}_t = f_\phi(\epsilon_t; \mathbf{s}_t)$ and random vector ϵ_t. The gradient of the policy loss with respect to the online policy's parameters ϕ is then given by*

$$\begin{aligned}\hat{\nabla}_\phi J_\pi(\phi) = \big(&\alpha \nabla_{\mathbf{a}_t} \log \pi_\phi(\mathbf{a}_t \mid \mathbf{s}_t) - \alpha \nabla_{\mathbf{a}_t} \log \pi_0(\mathbf{a}_t \mid \mathbf{s}_t) \\ &- \nabla_{\mathbf{a}_t} Q(\mathbf{s}_t, \mathbf{a}_t)\big) \nabla_\phi f_\phi(\epsilon_t; \mathbf{s}_t) + \alpha \nabla_\phi \log \pi_\phi(\mathbf{a}_t \mid \mathbf{s}_t)\end{aligned} \tag{4}$$

with $\nabla_{\mathbf{a}_t} \log \pi_0(\mathbf{a}_t \mid \mathbf{s}_t) = -\frac{\mathbf{a}_t - \boldsymbol{\mu}_0(\mathbf{s}_t)}{\boldsymbol{\sigma}_0^2(\mathbf{s}_t)}$. *For fixed* $|\mathbf{a}_t - \boldsymbol{\mu}_0(\mathbf{s}_t)|$, $\nabla_{\mathbf{a}_t} \log \pi_0(\mathbf{a}_t|\mathbf{s}_t)$ *grows as* $\mathcal{O}(\boldsymbol{\sigma}_0^{-2}(\mathbf{s}_t))$*; thus,*

$$|\hat{\nabla}_\phi J_\pi(\phi)| \to \infty \quad \textit{as} \quad \boldsymbol{\sigma}_0^2(\mathbf{s}_t) \to 0 \quad \textit{whenever} \quad \nabla_\phi f_\phi(\epsilon_t; \mathbf{s}_t) \neq 0.$$

Proof. See Appendix A.1. □

This result formalizes the intuition presented in Section 3.1 that a behavioral policy with a sufficiently small predictive variance may cause KL-regularized RL to suffer from pathological behavior in gradient-based optimization. The smaller the behavioral policy's predictive variance, the more sensitive the policy objective's gradients will be to differences in the means of the online and behavioral policies. As a result, for behavioral policies with small predictive variance, the KL divergence will heavily penalize online policies whose predictive means diverge from the predictive means of the behavioral policy—even at states off the expert trajectory where the behavioral policy's mean prediction is poor.

3.3 Predictive Uncertainty Collapse Under Parametric Policies

The most commonly used method for training behaviorally cloned stochastic policies is maximum likelihood estimation (MLE) [42, 49], where we seek $\pi_0 \doteq \pi_{\psi^\star}$ with $\psi^\star \doteq \arg\max_\psi \left\{\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_0}[\log \pi_\psi(\mathbf{a} \mid \mathbf{s})]\right\}$ for a parametric behavioral policy π_ψ. In practice, π_ψ is often assumed to be Gaussian, $\pi_\psi(\cdot \mid \mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}_\psi(\mathbf{s}), \boldsymbol{\sigma}_\psi^2(\mathbf{s}))$, with $\boldsymbol{\mu}_\psi(\mathbf{s})$ and $\boldsymbol{\sigma}_\psi^2(\mathbf{s})$ parameterized by a neural network.

While maximizing the likelihood of the expert trajectories under the behavioral policy is a sensible choice for behavioral cloning, the limited capacity of the neural network parameterization can produce unwanted behaviors in the resulting policy. The maximum likelihood objective ensures that the behavioral policy's predictive mean reflects the expert's actions and the predictive variance the (aleatoric) uncertainty inherent in the expert trajectories.

However, the maximum likelihood objective encourages parametric policies to use their model capacity toward fitting the expert demonstrations and reflecting the aleatoric uncertainty in the data. As a result, for states off the expert trajectories, the policy can become degenerate and collapse to point predictions instead of providing meaningful predictive variance estimates that reflect that the behavioral policy ought to be highly uncertain about its predictions in previously unseen parts of the state space. Similar behaviors are well-known in parametric probabilistic models and well-documented in the approximate Bayesian inference literature [32, 38].

Figure 1 demonstrates the collapse in predictive variance under maximum likelihood estimation in a low-dimensional representation of the "door-binary-v0" dexterous hand manipulation environment. It shows that while the predictive variance is small close to the expert trajectories (depicted as black lines), it rapidly decreases further away from them. Examples of variance collapse in other environments are presented in Appendix B.6. Figure 2 shows that the predictive variance off the expert trajectories consistently decreases during training. As shown in Proposition 1, such a collapse in predictive variance

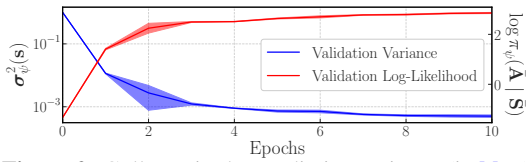

Figure 2: Collapse in the predictive variance (in **blue**) of a Gaussian behavioral policy parameterized by a neural network when training via maximum likelihood estimation. Lines and shaded regions denote means and standard deviations over five random seeds, respectively.

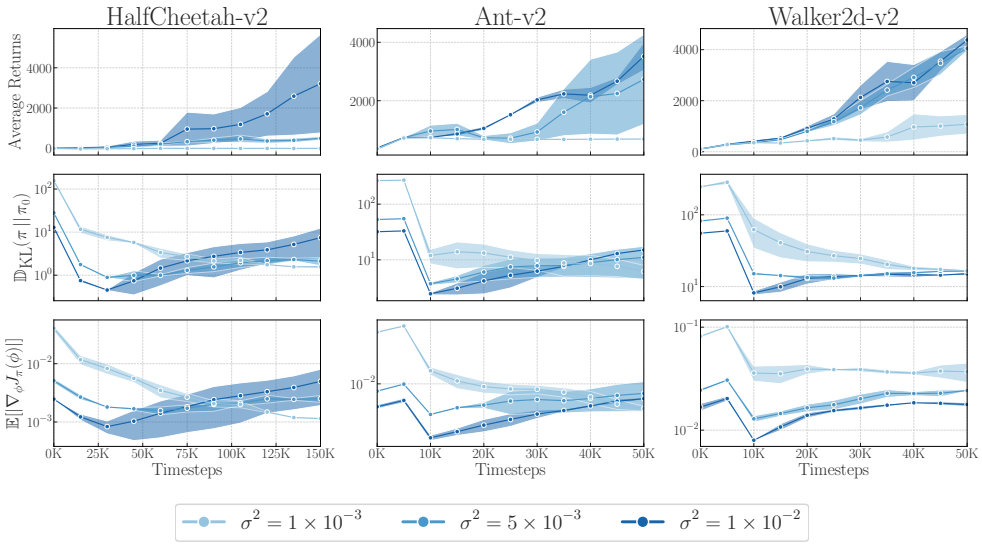

Figure 3: Ablation study showing the effect of predictive variance collapse on the performance of KL-regularized RL on MuJoCo environments. The plots show the average return of the learned policy, the magnitude of the KL penalty, and the magnitude of the average absolute gradients of the policy loss during online training. The lighter the shading, the lower the behavioral policy's predictive variance.

can result in pathological behavior in KL-regularized online training—steering the online policy towards suboptimal trajectories away from the expert demonstrations and deteriorating performance.

Effect of regularization on uncertainty collapse. To prevent a collapse in the behavioral policy's predictive variance, prior work proposed adding entropy or Tikhonov regularization to the MLE objective [49]. However, doing so does not succeed in preventing a collapse in predictive variance off the expert demonstration trajectories, as we show in Appendix A.3. Deep ensembles [19], whose predictive mean and variance are computed from the predictive means and variances of multiple Gaussian neural networks, are a widely used method for uncertainty quantification in regression settings. However, model ensembling can be costly and unreliable, as it requires training multiple neural networks from scratch and does not guarantee well-calibrated uncertainty estimates [38, 47]. We provide visualizations in Appendix B.5 which show that ensembling multiple neural network policies can slightly improve performance relative to single models but does not fully prevent a collapse in the predictive variance.

3.4 Empirical Confirmation of Uncertainty Collapse

To confirm Proposition 1 empirically and assess the effect of the collapse in predictive variance on the performance of KL-regularized RL, we perform an ablation study where we fix the predictive mean function of a behavioral policy to a mean function that attains 60% of the optimal performance and vary the magnitude of the policy's predictive variance. Specifically, we set the behavioral policy's predictive variance to different constant values in the set $\{1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}\}$ (following a similar implementation in Nair et al. [26]).[4] The results of this experiment are shown in Figure 3, which shows the average returns, the KL divergence, and the average absolute gradients of the policy loss over training. The plots confirm that as the predictive variance of the offline behavioral policy tends to zero, the KL terms and average policy gradient magnitude explode as implied by Proposition 1, leading to unstable training and a collapse or dampening in average returns.

In other words, even for behavioral policies with accurate predictive means, smaller predictive variances slow down or even entirely prevent learning good behavioral policies. This observation confirms that the pathology identified in Proposition 1 occurs in practice and that it can have a significant impact on KL-regularized RL from expert demonstrations, calling into question the usefulness of KL regularization as a means for accelerating and improving online training. In Appendix B.1, we show that an analogous relationship exists for the gradients of the Q-function loss.

[4] We attempted to use smaller values, but the gradients grew too large and caused arithmetic overflow.

4 Fixing the Pathology

In order to address the collapse in predictive uncertainty for behavioral policies parameterized by a neural network trained via MLE, we specify a *non-parametric* behavioral policy whose predictive variance is *guaranteed* not to collapse about previously unseen states. Noting that KL-regularized RL with a behavioral policy can be viewed as approximate Bayesian inference with an empirical prior policy [12, 20, 39], we propose *Non-Parametric Prior Actor–Critic* (N-PPAC), an off-policy temporal difference algorithm for improved, accelerated, and stable online learning with behavioral policies.

4.1 Non-Parametric Gaussian Processes Behavioral Policies

Gaussian processes (GPs) [35] are models over functions defined by a mean $m(\cdot)$ and covariance function $k(\cdot,\cdot)$. When defined in terms of a non-parametric covariance function, that is, a covariance function constructed from infinitely many basis functions, we obtain a non-degenerate GP, which has sufficient capacity to prevent a collapse in predictive uncertainty away from the training data. Unlike parametric models, whose capacity is limited by their parameterization, a non-parametric model's capacity *increases* with the amount of training data.

Considering a non-parametric GP behavioral policy, $\pi_0(\cdot \mid \mathbf{s})$, with

$$\mathbf{A} \mid \mathbf{s} \sim \pi_0(\cdot \mid \mathbf{s}) = \mathcal{GP}\big(m(\mathbf{s}), k(\mathbf{s}, \mathbf{s}')\big), \tag{5}$$

we can obtain a *non-degenerate* posterior predictive distribution over actions conditioned on the offline data $\mathcal{D}_0 = \{\bar{\mathbf{S}}, \bar{\mathbf{A}}\}$ with actions sampled according to $\mathbf{A} \mid \mathbf{s}, \mathcal{D}_0 \sim \pi_0(\cdot \mid \mathbf{s}, \mathcal{D}_0) = \mathcal{GP}\big(\boldsymbol{\mu}_0(\mathbf{s}), \boldsymbol{\Sigma}_0(\mathbf{s}, \mathbf{s}')\big)$, and

$$\mu(\mathbf{s}) = m(\mathbf{s}) + k(\mathbf{s}, \bar{\mathbf{S}})k(\bar{\mathbf{S}}, \bar{\mathbf{S}})^{-1}(\bar{\mathbf{A}} - m(\bar{\mathbf{A}})) \quad \text{and} \quad \Sigma(\mathbf{s}, \mathbf{s}') = k(\mathbf{s}, \mathbf{s}') + k(\mathbf{s}, \bar{\mathbf{S}})k(\bar{\mathbf{S}}, \bar{\mathbf{S}})^{-1}k(\bar{\mathbf{S}}, \mathbf{s}').$$

To obtain this posterior distribution, we perform exact Bayesian inference, which naively scales as $\mathcal{O}(N^3)$ in the number of training points N, but Wang et al. [48] show that exact inference in GP regression can be scaled to $N > 1,000,000$. Since expert demonstrations usually contain less than 100k datapoints, non-parametric GP behavioral policies are applicable to a wide array of real-world tasks. For an empirical evaluation of the time complexity of using a GP prior, see Section 5.5.

Figure 1 confirms that the non-parametric GP's predictive variance is well-calibrated: It is small near the expert trajectories and large in other parts of the state space. While actor–critic algorithms like SAC implicitly use a uniform prior to explore the state space, using a behavioral policy with well-calibrated predictive variance has the benefit that, close to the offline training data, the online policy learns to match the expert, whereas away from the data, the predictive variance increases and encourages exploration.

Algorithmic details. In our experiments, we use a KL-regularized objective with a standard actor–critic implementation and Double DQN [13]. Pseudocode for the algorithm is provided in Algorithm 1 (Appendix C.1).

5 Empirical Evaluation

We carry out a comparative empirical evaluation of our proposed approach vis-à-vis related methods that integrate offline data into online training. We provide a detailed description of the algorithms we compare against in Appendix A.4. We perform experiments on the MuJoCo benchmark suite and the substantially more challenging dexterous hand manipulation suite with sparse rewards.

We show that KL-regularized RL with a non-parametric behavioral policy can rapidly learn to solve difficult high-dimensional continuous control problems given only a small set of expert demonstrations and (often significantly) outperforms state-of-the-art methods, including ones that use offline reward information—which our approach does not require. Furthermore, we demonstrate that the GP behavioral policy's predictive variance is crucial for KL-regularized objectives to learn good online policies with the help of expert demonstrations. Finally, we perform ablation studies that illustrate that non-parametric GP behavioral policies also outperform parametric behavioral policies with improved uncertainty quantification, such as deep ensembles and Bayesian neural networks (BNNs) with Monte

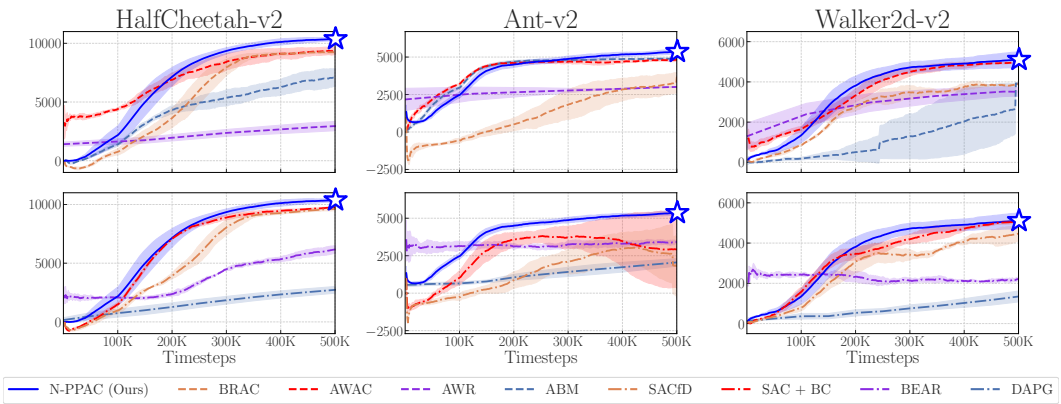

Figure 4: Comparison of N-PPAC (ours) vs. previous baselines on standard MuJoCo benchmark tasks. **Top**: KL-based methods (dashed lines), **Bottom**: Non-KL-based methods (dash-dotted lines). Both top and bottom plots include N-PPAC (**blue**). BRAC uses the same actor–critic algorithm as N-PPAC, but uses a parametric behavioral policy, and results in slower learning and worse final performance.

Carlo dropout, and that the differential between non-parametric and parametric models is exacerbated the fewer expert demonstrations are available. For our empirical evaluation, we use the expert data from Nair et al. [26]. Every experiment uses six random seeds, and we use a fixed KL-temperature for each environment class. For further implementation details, see Appendix C.2.

5.1 Environments

MuJoCo locomotion tasks. We evaluate N-PPAC on three representative tasks: "Ant-v2", "HalfCheetah-v2", and "Walker2d-v2". For each task, we use 15 demonstration trajectories collected by a pre-trained expert, each containing 1,000 steps. The behavioral policy is specified as the posterior distribution of a GP with a squared exponential kernel, which is well-suited for modeling smooth functions.

Dexterous hand manipulation tasks. Real-world robot learning is a setting where human demonstration data is readily available, and many deep RL approaches fail to learn efficiently. We study this setting in a suite of challenging dexterous manipulation tasks [34] using a 28-DoF five-fingered simulated ADROIT hand. The tasks simulate challenges common to real-world settings with high-dimensional action spaces, complex physics, and a large number of intermittent contact forces. We consider two tasks in particular: in-hand rotation of a pen to match a target and opening a door by unlatching and pulling a handle. We use binary rewards for task completion, which is significantly more challenging than the original setting considered in Rajeswaran et al. [34]. 25 expert demonstrations were provided for each task, each consisting of 200 environment steps which are not fully optimal but do successfully solve the task. The behavioral policy is specified as the posterior distribution of a GP with a Matérn kernel, which is more suitable for modeling non-smooth data.

5.2 Results

On MuJoCo environments, KL-regularized RL with a non-parametric behavioral policy consistently outperforms all related methods across all three tasks, successfully accelerating learning from offline data, as shown in Figure 4. Most notably, it outperforms methods such as AWAC [26]—the previous state-of-the-art—which attempts to eschew the problem of learning behavioral policies but instead uses an implicit constraint. Our approach, N-PPAC, exhibits an increase in stability and higher returns compared to comparable methods such as ABM and BRAC that explicitly regularize the online policy against a parametric behavioral policy and plateau at suboptimal performance levels as they are being forced to copy poor actions from the behavioral policy away from the expert data. In contrast, using a non-parametric behavioral policy allows us to avoid such undesirable behavior.

On dexterous hand manipulation environments, KL-regularized RL with a non-parametric behavioral policy performs on par or outperforms all related methods on both tasks, as shown in Figure 5. Most notably, on the door opening task, it achieves a stable success rate of 90% within only 100,000 environment interactions For comparison, AWAC requires 4× as many environment interactions to

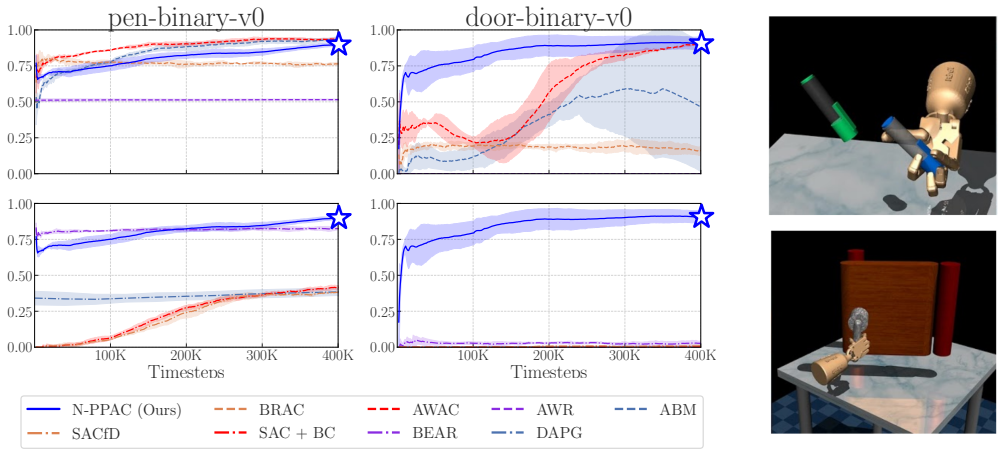

Figure 5: Left & Center: Comparison of N-PPAC (ours) vs. previous baselines on dexterous hand manipulation tasks. **Top**: KL-based methods (dashes), **Bottom**: Non-KL-based methods (dots and dashes). Both top and bottom plots include N-PPAC (**blue**). **Right**: The pen-binary-v0 (top) and door-binary-v0 (bottom) environments.

achieve the same performance and is significantly less stable, while most other methods fail to learn any meaningful behaviors.

Alternative divergence metrics underperform KL-regularization. KL-regularized RL with a non-parametric behavioral policy consistently outperforms methods that use alternative divergence metrics, as shown in the bottom plots of Figures 4 and 5.

5.3 Can the Pathology Be Fixed by Improved Parametric Uncertainty Quantification?

To assess whether the success of non-parametric behavioral policies is due to their predictive variance estimates—as suggested by Proposition 1—or due to better generalization from their predictive means, we perform an ablation study on the predictive variance of the behavioral policy. To isolate the effect of the predictive variance on optimization, we perform online training using behavioral policies with different predictive variance functions (parametric and non-parametric) and identical mean functions, which we set to be the predictive mean of the GP posterior (which achieves a success rate of ~80%). If the pathology identified in Proposition 1 can be remedied by commonly used parametric uncertainty quantification methods, we would expect the parametric and non-parametric behavioral policy variance functions to result in similar online policy success rates. We consider the challenging "door-binary-v0" environment for this ablation study.

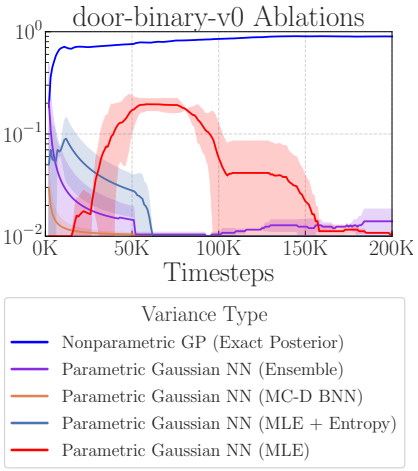

Figure 6: Post-online training success rates with different behavioral policy variance functions.

Parametric uncertainty quantification is insufficient. Figure 5 shows that using the parametric predictive variance functions results in online policies that only achieve success rates of up to 20% and eventually deteriorate, whereas the non-parametric variance yields an online policy that achieves a success rate of nearly 100%. This finding shows that commonly used uncertainty quantification methods, such as deep ensembles or BNNs with Monte Carlo dropout, do not generate sufficiently well-calibrated uncertainty estimates to remedy the pathology.

Lower-bounding the predictive variance does not remedy the pathology. The predictive variance of all MLE-based behavioral policies as well as of the ensemble behavioral policy are bounded away from zero at a minimum value of $\approx 10^{-9}$. Hence, setting a floor on the variance is not sufficient to avoid the pathology. This result further demonstrates the importance of accurate predictive variance estimation in allowing the online policy to choose expert actions with high likelihood and explore away from the data, where the behavioral policy assigns a low likelihood to any action.

5.4 Can a Single Expert Demonstration Be Sufficient to Accelerate Online Training?

To assess the usefulness of non-parametric behavioral policies in settings where only few expert demonstrations are available, we investigate whether the difference in performance between online policies trained with non-parametric and parametric behavioral policies, respectively, is exacerbated the fewer expert demonstrations are available. To answer this question, we consider the "HalfCheetah-v2" environment and compare online policies trained with different behavioral policies—non-parametric GPs, deep ensembles, and BNNs with Monte Carlo dropout—estimated either from 15 expert demonstrations (i.e., 15 state–action trajectories, containing 15,000 samples) or from a single expert demonstration (i.e., a single state–action trajectory, containing 1,000 samples).

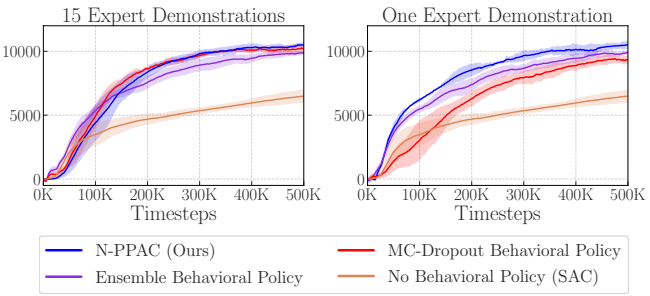

Figure 7: Returns during online training with different behavioral policies and varying amounts of expert demonstration data on "HalfCheetah-v2".

A single expert demonstration is sufficient for non-parametric behavioral policies. Figure 7 shows the returns for online policies trained with behavioral policies estimated from the full dataset (top plot) and from only a single expert state–action trajectory (bottom plot). On the full dataset, we find that all three methods are competitive and improve on the prior state-of-the-art but that the GP behavioral policy leads to the highest return. Remarkably, non-parametric GP behavioral policies perform just as well with only a single expert demonstration as with all 15 (i.e., with 1,000 data points, instead of 15,000 data points). These results further emphasizes the usefulness of non-parametric behavioral policies when accelerating online training with expert demonstrations—even when only very few expert demonstrations are available.

5.5 Are Non-Parametric GP Behavioral Policies Too Computationally Expensive?

Table 1 presents the time complexity of KL-regularized RL under non-parametric GP and parametric neural network behavioral policies, as measured by the average time elapsed per epoch on the "door-binary-v0" and "HalfCheetah-v2" environments. One epoch of online training on "door-binary-v0" and "HalfCheetah-v2" requires computing the KL divergence over 1,000 mini-batches of size 256 and 1,024, respectively. The time complexity of evaluating the log-density of a GP behavioral policy—needed for computing gradients of the KL divergence during online training—scales quadratically in the number of training data points and linearly in the dimensionality of the state and action space, respectively. As can be seen in Table 1, non-parametric GP behavioral policies only lead to a modest increase in the time needed to complete one epoch of training while resulting in significantly improved performance as shown in Figures 4 and 5.

Table 1: Time per epoch under different behavioral policies for expert demonstration data of varying size computed on a GeForce RTX 3080 GPU. The first and second value in each entry of the table give the time required when using a single parametric neural network and a Gaussian process behavioral policy, respectively.

Dataset	1,000 Data Points	5,000 Data Points	15,000 Data Points
HalfCheetah-v2	12.00s / 16.06s	11.59s / 18.31s	12.00s / 46.54s
door-binary-v0	19.62s / 23.78s	19.62s / 33.62s	-

6 Conclusion

We identified a previously unrecognized pathology in KL-regularized RL from expert demonstrations and showed that this pathology can significantly impede and even entirely prevent online learning. To remedy the pathology, we proposed the use of non-parametric behavioral policies, which we showed can significantly accelerate and improve online learning and yield online policies that (often greatly) outperform current state-of-the-art methods on challenging continuous control tasks. We hope that this work will encourage further research into better model classes for deep reinforcement learning algorithms, including and especially for RL from pixels.

Acknowledgments and Disclosure of Funding

We thank Ashvin Nair for sharing his code and results, as well as for providing helpful insights about the dexterous hand manipulation suite. We also thank Clare Lyle, Charline Le Lan, and Angelos Filos for detailed feedback on an early draft of this paper, Avi Singh for early discussions about behavioral cloning in entropy-regularized RL, and Tim Pearce for a useful discussion on the role of good models in RL. TGJR and CL are funded by the Engineering and Physical Sciences Research Council (EPSRC). TGJR is also funded by the Rhodes Trust and by a Qualcomm Innovation Fellowship. We gratefully acknowledge donations of computing resources by the Alan Turing Institute.

References

[1] Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pages 103–129, 1995.

[2] Philip J Ball, Cong Lu, Jack Parker-Holder, and Stephen Roberts. Augmented world models facilitate zero-shot dynamics generalization from a single offline environment. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 619–629. PMLR, 18–24 Jul 2021.

[3] Abdeslam Boularias, Jens Kober, and Jan Peters. Relative entropy inverse reinforcement learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 182–189. JMLR Workshop and Conference Proceedings, 2011.

[4] Ivan Bratko, Tanja Urbancic, and Claude Sammut. Behavioural cloning: phenomena, results and problems. *IFAC Proceedings Volumes*, 28(21):143–149, 1995.

[5] Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E Taylor, and Ann Nowé. Reinforcement learning from demonstrations through shaping. In *Twenty-fourth International Joint Conference on Artificial Intelligence*, 2015.

[6] Catherine Cang, Aravind Rajeswaran, Pieter Abbeel, and Michael Laskin. Behavioral priors and dynamics models: Improving performance and domain transfer in offline RL, 2021.

[7] Thomas Degris, Martha White, and Richard S. Sutton. Off-policy actor-critic. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 179–186, Madison, WI, USA, 2012. Omnipress.

[8] Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

[9] Alexandre Galashov, Siddhant M. Jayakumar, Leonard Hasenclever, Dhruva Tirumala, Jonathan Schwarz, Guillaume Desjardins, Wojciech M. Czarnecki, Yee Whye Teh, Razvan Pascanu, and Nicolas Heess. Information asymmetry in kl-regularized RL. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

[10] Yang Gao, Huazhe Xu, Ji Lin, Fisher Yu, Sergey Levine, and Trevor Darrell. Reinforcement learning from imperfect demonstrations. *arXiv preprint arXiv:1802.05313*, 2018.

[11] CW Groetsch. The theory of Tikhonov regularization for Fredholm equations. *Boston Pitman Publication*, 1984.

[12] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications, 2019.

[13] Hado V. Hasselt. Double Q-learning. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2613–2621, 2010.

[14] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[15] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21810–21823. Curran Associates, Inc., 2020.

[16] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014, 2000.

[17] George Konidaris, Scott Kuindersma, Roderic Grupen, and Andrew Barto. Robot learning from demonstration by constructing skill trees. *The International Journal of Robotics Research*, 31 (3):360–375, 2012.

[18] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems 32*, pages 11784–11794, 2019.

[19] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413, 2017.

[20] Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review, 2018.

[21] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *ICLR (Poster)*, 2016.

[22] Cong Lu, Philip J. Ball, Jack Parker-Holder, Michael A. Osborne, and Stephen J. Roberts. Revisiting design choices in model-based offline reinforcement learning, 2021.

[23] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.

[24] Kevin P. Murphy. *Machine learning: A Probabilistic Perspective*. MIT Press, 2013.

[25] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6292–6299, 2018.

[26] Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. Accelerating online reinforcement learning with offline datasets, 2020.

[27] Nicolás Navarro-Guerrero, Cornelius Weber, Pascal Schroeter, and Stefan Wermter. Real-world reinforcement learning for autonomous humanoid robot docking. *Robotics and Autonomous Systems*, 60(11):1400–1407, 2012.

[28] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[29] Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, Krzysztof Choromanski, Anna Choromanska, and Michael Jordan. Learning to score behaviors for guided policy optimization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7445–7454. PMLR, 13–18 Jul 2020.

[30] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2019.

[31] Jan Peters, Sethu Vijayakumar, and Stefan Schaal. Natural actor-critic. In *European Conference on Machine Learning*, pages 280–291. Springer, 2005.

[32] Joaquin Quiñonero Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.*, 6:1939–1959, December 2005. ISSN 1532-4435.

[33] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2018.

[34] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations, 2018.

[35] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[36] Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. *Proceedings of Robotics: Science and Systems VIII*, 2012.

[37] Michael T Rosenstein, Andrew G Barto, Jennie Si, Andy Barto, and Warren Powell. Supervised actor-critic reinforcement learning. *Learning and Approximate Dynamic Programming: Scaling Up to the Real World*, pages 359–380, 2004.

[38] Tim G. J. Rudner, Zonghao Chen, and Yarin Gal. Rethinking function-space variational inference in Bayesian neural networks. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021.

[39] Tim G. J. Rudner, Vitchyr H. Pong, Rowan McAllister, Yarin Gal, and Sergey Levine. Outcome-driven reinforcement learning via variational inference. In *Advances in Neural Information Processing Systems 34*. 2021.

[40] Stefan Schaal et al. Learning from demonstration. *Advances in neural information processing systems*, pages 1040–1046, 1997.

[41] John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft Q-learning. *arXiv preprint arXiv:1704.06440*, 2017.

[42] Noah Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, Nicolas Heess, and Martin Riedmiller. Keep doing what worked: Behavior modelling priors for offline reinforcement learning. In *International Conference on Learning Representations*, 2020.

[43] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, editors, *Algorithmic Learning Theory*, pages 13–31, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[44] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[45] Gerald Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

[46] Emanuel Todorov. Linearly-solvable markov decision problems. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19, pages 1369–1376. MIT Press, 2007.

[47] Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. Improving deterministic uncertainty estimation in deep learning for classification and regression, 2021.

[48] Ke Wang, Geoff Pleiss, Jacob Gardner, Stephen Tyree, Kilian Q Weinberger, and Andrew Gordon Wilson. Exact Gaussian processes on a million data points. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

[49] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

[50] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14129–14142. Curran Associates, Inc., 2020.

[51] Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. COMBO: Conservative offline model-based policy optimization, 2021.