

Deep Bayesian Quadrature Policy Optimization

Ravi Tej Akella¹, Kamyar Azizzadenesheli¹, Mohammad Ghavamzadeh²,
Anima Anandkumar³ and Yisong Yue³

¹Purdue University

²Google Research

³Caltech

{rakella,kamyar}@purdue.edu, ghavamza@google.com, {anima,yyue}@caltech.edu

Abstract

We study the problem of obtaining accurate policy gradient estimates using a finite number of samples. Monte-Carlo methods have been the default choice for policy gradient estimation, despite suffering from high variance in the gradient estimates. On the other hand, more sample efficient alternatives like Bayesian quadrature methods have received little attention due to their high computational complexity. In this work, we propose deep Bayesian quadrature policy gradient (DBQPG), a computationally efficient high-dimensional generalization of Bayesian quadrature, for policy gradient estimation. We show that DBQPG can substitute Monte-Carlo estimation in policy gradient methods, and demonstrate its effectiveness on a set of continuous control benchmarks. In comparison to Monte-Carlo estimation, DBQPG provides (i) more accurate gradient estimates with a significantly lower variance, (ii) a consistent improvement in the sample complexity and average return for several deep policy gradient algorithms, and, (iii) the uncertainty in gradient estimation that can be incorporated to further improve the performance.

1 Introduction

Policy gradient (PG) is a reinforcement learning (RL) approach that directly optimizes the agent’s policies by operating on the gradient of their expected return (Sutton et al. 2000; Baxter and Bartlett 2000). The use of deep neural networks for the policy class has recently demonstrated a series of success for PG methods (Lillicrap et al. 2015; Schulman et al. 2015) on high-dimensional continuous control benchmarks, such as MuJoCo (Todorov, Erez, and Tassa 2012). However, the derivation and analysis of the aforementioned methods mainly rely on access to the expected return and its true gradient. In general, RL agents do not have access to the true gradient of the expected return, i.e., the gradient of integration over returns; instead, they have access to its empirical estimate from sampled trajectories. Monte-Carlo (MC) sampling (Metropolis and Ulam 1949) is a widely used point estimation method for approximating this integration (Williams 1992). However, MC estimation returns high variance gradient estimates that are undesirably inaccurate, imposing a high sample complexity requirement for PG methods (Rubinstein 1969; Ilyas et al. 2018).

An alternate approach to approximate integrals in probabilistic numerics is Bayesian Quadrature (BQ) (O’Hagan 1991). Under mild regularity assumptions, BQ offers impressive empirical advances and strictly faster convergence rates (Kanagawa, Sriperumbudur, and Fukumizu 2016). Typically, the integrand in BQ is modeled using a Gaussian process (GP), such that the linearity of the integral operator provides a Gaussian posterior over the integral. In addition to a point estimate of the integral, BQ also quantifies the uncertainty in its estimation, a missing piece in MC methods.

In RL, the BQ machinery can be used to obtain a Gaussian approximation of the PG integral, by placing a GP prior over the action-value function. However, estimating the moments of this Gaussian approximation, i.e., the PG mean and covariance, requires the integration of GP’s kernel function, which, in general, does not have an analytical solution. Interestingly, Ghavamzadeh and Engel (2007) showed that the PG integral can be solved analytically when the GP kernel is an additive combination of an arbitrary state kernel and a fixed Fisher kernel. While the authors demonstrate a superior performance of BQ over MC using a small policy network on simple environments, their approach does not scale to large non-linear policies (> 1000 trainable parameters) and high-dimensional domains.

Contribution 1: We propose deep Bayesian quadrature policy gradient (DBQPG), a BQ-based PG framework that extends Ghavamzadeh and Engel (2007) to high-dimensional settings, thus placing it in the context of contemporary deep PG algorithms. The proposed framework uses a GP to implicitly model the action-value function, and without explicitly constructing the action-value function, returns a Gaussian approximation of the policy gradient, represented by a mean vector (gradient estimate) and covariance (gradient uncertainty). Consequently, this framework can be used with an explicit critic network (different from implicit GP critic¹) to leverage the orthogonal benefits of actor-critic frameworks (Schulman et al. 2016).

The statistical efficiency of BQ, relative to MC estimation, depends on the compatibility between the GP kernel and the MDP’s action-value function. To make DBQPG robust to a diverse range of target MDPs, we choose a base

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹The GP critic in BQ is implicit because it is used to solve PG integral analytically rather than explicitly predicting the Q -values.

kernel capable of universal approximation (e.g. RBF kernel), and augment its expressive power using deep kernel learning (Wilson et al. 2016). Empirically, DBQPG estimates gradients that are both much closer to the true gradient, and with much lower variance, when compared to MC estimation. Moreover, DBQPG is a linear-time program that leverages the recent advances in structured kernel interpolation (Wilson and Nickisch 2015) for computational efficiency and GPU acceleration for fast kernel learning (Gardner et al. 2018). Therefore, DBQPG can favorably substitute MC estimation subroutine in a variety of PG algorithms. Specifically, we show that replacing the MC estimation subroutine with DBQPG provides a significant improvement in the sample complexity and average return for vanilla PG (Sutton et al. 2000), natural policy gradient (NPG) (Kakade 2001) and trust region policy optimization (TRPO) (Schulman et al. 2015) algorithms on 7 diverse MuJoCo domains.

Contribution 2: We propose uncertainty aware PG (UAPG), a novel policy gradient method that utilizes the uncertainty in the gradient estimation for computing reliable policy updates. A majority of PG algorithms (Kakade 2001; Schulman et al. 2015) are derived assuming access to true gradients and therefore do not account the stochasticity in gradient estimation due to finite sample size. However, one can obtain more reliable policy updates by lowering the step-size along the directions of high gradient uncertainty. UAPG captures this intuition by utilizing DBQPG’s uncertainty to bring different components of a stochastic gradient estimate to the same scale. UAPG does this by normalizing the step-size of the gradient components by their respective uncertainties, returning a new gradient estimate with uniform uncertainty, i.e., with new gradient covariance as the identity matrix. UAPG further provides a superior sample complexity and average return in comparison to DBQPG. Check the supporting resources² for more details.

2 Preliminaries

Consider a Markov decision process (MDP) $\langle \mathcal{S}, \mathcal{A}, P, r, \rho_0, \gamma \rangle$, where \mathcal{S} is the state-space, \mathcal{A} is the action-space, $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta_{\mathcal{S}}$ is the transition kernel that maps each state-action pair to a distribution over the states $\Delta_{\mathcal{S}}$, $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward kernel, $\rho_0 : \Delta_{\mathcal{S}}$ is the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. We denote by $\pi_{\theta} : \mathcal{S} \rightarrow \Delta_{\mathcal{A}}$ a stochastic parameterized policy with parameters $\theta \in \Theta$. The MDP controlled by the policy π_{θ} induces a Markov chain over state-action pairs $z = (s, a) \in \mathcal{Z} = \mathcal{S} \times \mathcal{A}$, with an initial density $\rho_0^{\pi_{\theta}}(z_0) = \pi_{\theta}(a_0|s_0)\rho_0(s_0)$ and a transition probability distribution $P^{\pi_{\theta}}(z_t|z_{t-1}) = \pi_{\theta}(a_t|s_t)P(s_t|z_{t-1})$. We use the standard definitions for action-value function $Q_{\pi_{\theta}}$ and expected return $J(\theta)$ under π_{θ} :

$$Q_{\pi_{\theta}}(z_t) = \mathbb{E} \left[\sum_{\tau=0}^{\infty} \gamma^{\tau} r(z_{t+\tau}) \mid z_{t+\tau+1} \sim P^{\pi_{\theta}}(\cdot|z_{t+\tau}) \right],$$

$$J(\theta) = \mathbb{E}_{z_0 \sim \rho_0^{\pi_{\theta}}} [Q_{\pi_{\theta}}(z_0)]. \quad (1)$$

However, the gradient of $J(\theta)$ with respect to policy parameters θ cannot be directly computed from this formulation. The policy gradient theorem (Sutton et al. 2000; Konda and Tsitsiklis 2000) provides an analytical expression for the gradient of the expected return $J(\theta)$, as:

$$\nabla_{\theta} J(\theta) = \int_{\mathcal{Z}} dz \rho^{\pi_{\theta}}(z) \mathbf{u}(z) Q_{\pi_{\theta}}(z)$$

$$= \mathbb{E}_{z \sim \rho^{\pi_{\theta}}} [\mathbf{u}(z) Q_{\pi_{\theta}}(z)], \quad (2)$$

where $\mathbf{u}(z) = \nabla_{\theta} \log \pi_{\theta}(a|s)$ is the score function and $\rho^{\pi_{\theta}}$ is the discounted state-action visitation frequency defined as:

$$\rho^{\pi_{\theta}}(z) = \sum_{t=0}^{\infty} \gamma^t P_t^{\pi_{\theta}}(z), \text{ where} \quad (3)$$

$$P_t^{\pi_{\theta}}(z_t) = \int_{z^t} dz_0 \dots dz_{t-1} P_0^{\pi_{\theta}}(z_0) \prod_{\tau=1}^t P^{\pi_{\theta}}(z_{\tau}|z_{\tau-1}).$$

3 Policy Gradient Evaluation

The exact evaluation of the PG integral in Eq. 2 is often intractable for MDPs with a large (or continuous) state or action space. We discuss two prominent approaches to approximate this integral using a finite set of samples $\{z_i\}_{i=1}^n \sim \rho^{\pi_{\theta}}$: (i) Monte-Carlo method and (ii) Bayesian Quadrature.

Monte-Carlo (MC) method³ approximates the integral in Eq. 2 by the finite sum:

$$L_{\theta}^{MC} = \frac{1}{n} \sum_{i=1}^n Q_{\pi_{\theta}}(z_i) \mathbf{u}(z_i) = \frac{1}{n} \mathbf{U} \mathbf{Q}, \quad (4)$$

where $\mathbf{u}(z)$ is a $|\Theta|$ dimensional vector ($|\Theta|$ is the number of policy parameters). For conciseness, the function evaluations at sample locations $\{z_i\}_{i=1}^n \sim \rho^{\pi_{\theta}}$ are grouped into $\mathbf{U} = [\mathbf{u}(z_1), \dots, \mathbf{u}(z_n)]$, a $|\Theta| \times n$ dimensional matrix, and $\mathbf{Q} = [Q_{\pi_{\theta}}(z_1), \dots, Q_{\pi_{\theta}}(z_n)]^4$, an n dimensional vector. MC method returns the gradient mean evaluated at sample locations, which, according to the central limit theorem (CLT), is an unbiased estimate of the true gradient. However, CLT also suggests that the MC estimates suffer from a slow convergence rate ($n^{-1/2}$) and high variance, necessitating a large sample size n for reliable PG estimation (Ilyas et al. 2018). Yet, MC methods are more computationally efficient than their sample-efficient alternatives (e.g. BQ), making them ubiquitous in PG algorithms.

Bayesian quadrature (BQ) (O’Hagan 1991) is an approach from *probabilistic numerics* (Hennig, Osborne, and Girolami 2015) that converts numerical integration into a

³Here, MC refers to the Monte-Carlo numerical integration method, and not the TD(1) (a.k.a Monte-Carlo) Q -estimates.

⁴ $Q_{\pi_{\theta}}(z)$ is computed using TD(1) (a.k.a Monte-Carlo) method or an explicit critic module (different from BQ’s implicit GP critic)

²Full-text (with supplement): <https://arxiv.org/abs/2006.15637>
Code: <https://github.com/Akella17/Deep-Bayesian-Quadrature-Policy-Optimization>

Bayesian inference problem. The first step in BQ is to formulate a prior stochastic model over the integrand. This is done by placing a Gaussian process (GP) prior on the Q_{π_θ} function, i.e., a mean zero GP $\mathbb{E}[Q_{\pi_\theta}(z)] = 0$, with a covariance function $k(z_p, z_q) = \text{Cov}[Q_{\pi_\theta}(z_p), Q_{\pi_\theta}(z_q)]$, and observation noise σ . Next, the GP prior on Q_{π_θ} is conditioned (Bayes rule) on the sampled data $\mathcal{D} = \{z_i\}_{i=1}^n$ to obtain the posterior moments $\mathbb{E}[Q_{\pi_\theta}(z)|\mathcal{D}]$ and $C_Q(z_1, z_2) = \text{Cov}[Q_{\pi_\theta}(z_1), Q_{\pi_\theta}(z_2)|\mathcal{D}]$. Since the transformation from $Q_{\pi_\theta}(z)$ to $\nabla_\theta J(\theta)$ happens through a linear integral operator (in Eq. 2), $\nabla_\theta J(\theta)$ also follows a Gaussian distribution:

$$\begin{aligned} L_\theta^{BQ} &= \mathbb{E}[\nabla_\theta J(\theta)|\mathcal{D}] = \int_{\mathcal{Z}} dz \rho^{\pi_\theta}(z) \mathbf{u}(z) \mathbb{E}[Q_{\pi_\theta}(z)|\mathcal{D}] \\ C_\theta^{BQ} &= \text{Cov}[\nabla_\theta J(\theta)|\mathcal{D}] \\ &= \int_{\mathcal{Z}^2} dz_1 dz_2 \rho^{\pi_\theta}(z_1) \rho^{\pi_\theta}(z_2) \mathbf{u}(z_1) C_Q(z_1, z_2) \mathbf{u}(z_2)^\top, \end{aligned} \quad (5)$$

where the mean vector L_θ^{BQ} is the PG estimate and the covariance C_θ^{BQ} is its uncertainty estimation. While the integrals in Eq. 5 are still intractable for an arbitrary GP kernel k , they have closed-form solutions when k is the additive composition of a state kernel k_s (arbitrary) and the Fisher kernel k_f (indispensable) (Ghavamzadeh and Engel 2007):

$$\begin{aligned} k(z_1, z_2) &= c_1 k_s(s_1, s_2) + c_2 k_f(z_1, z_2), \\ \text{with } k_f(z_1, z_2) &= \mathbf{u}(z_1)^\top \mathbf{G}^{-1} \mathbf{u}(z_2), \end{aligned} \quad (6)$$

where c_1, c_2 are hyperparameters⁵ and \mathbf{G} is the Fisher information matrix of π_θ . Using the matrices,

$$\begin{aligned} \mathbf{K}_f &= \mathbf{U}^\top \mathbf{G}^{-1} \mathbf{U}, \quad \mathbf{K} = c_1 \mathbf{K}_s + c_2 \mathbf{K}_f, \\ \mathbf{G} &= \mathbb{E}_{z \sim \rho^{\pi_\theta}}[\mathbf{u}(z) \mathbf{u}(z)^\top] \approx \frac{1}{n} \mathbf{U} \mathbf{U}^\top, \end{aligned} \quad (7)$$

the PG mean L_θ^{BQ} and covariance C_θ^{BQ} can be computed analytically (proof is deferred to the supplement Sec. B),

$$\begin{aligned} L_\theta^{BQ} &= c_2 \mathbf{U} (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{Q}, \\ C_\theta^{BQ} &= c_2 \mathbf{G} - c_2^2 \mathbf{U} (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{U}^\top. \end{aligned} \quad (8)$$

Here, \mathbf{Q} is same as in the MC method. Note that removing the Fisher kernel k_f (setting $c_2 = 0$ in Eq. 8) causes $L_\theta^{BQ} = 0$ and $C_\theta^{BQ} = 0$. Further, Ghavamzadeh and Engel (2007) showed that BQ-PG L_θ^{BQ} reduces to MC-PG L_θ^{MC} when the state kernel k_s is removed ($c_1 = 0$).

To summarize, (i) the presence of Fisher kernel ($c_2 \neq 0$) is essential for obtaining an analytical solution for BQ-PG, and (ii) with the Fisher kernel fixed, the choice of state kernel alone determines the convergence rate of BQ-PG relative to the MC baseline (equivalently BQ-PG with $c_1 = 0$).

4 Deep Bayesian Quadrature Policy Gradient

Here, we introduce the steps needed for obtaining a practical BQ-PG method, that (i) is more sample-efficient than MC baseline, and (ii) easily scales to high-dimensional settings.

⁵ c_1, c_2 are redundant hyperparameters that are simply introduced to offer a better explanation of BQ-PG.

Kernel Selection: In the previous section, it was highlighted that the flexibility of kernel selection is limited to the choice of the state kernel k_s . To understand the role of k_s , we first highlight the implication of the kernel composition proposed in Ghavamzadeh and Engel (2007). Specifically, the additive composition of the state kernel k_s and the Fisher kernel k_f implicitly divides the Q_{π_θ} function into state-value function and advantage function, separately modeled by k_s and k_f respectively (supplement, Sec. C.1). Thus, removing the Fisher kernel k_f ($c_2 = 0$) results in a non-existent advantage function, which explains $L_\theta^{BQ} = 0$ and $C_\theta^{BQ} = 0$. More interestingly, removing the state kernel k_s ($c_1 = 0$) results in a non-existent state-value function, which also reduces L_θ^{BQ} to L_θ^{MC} . In other words, MC-PG is a limiting case of BQ-PG where the state-value function is suppressed to 0 (more details in the supplement Sec. C.2).

Thus, BQ-PG can offer more accurate gradient estimates than MC-PG, along with well-calibrated uncertainty estimates, when the state kernel k_s is a better prior than the trivial $k_s = 0$, for the MDP’s state-value function. To make the choice of k_s robust to a diverse range of target MDPs, we suggest (i) a base kernel capable of universal approximation (e.g. RBF kernel), followed by (ii) increasing the base kernel’s expressive power by using a deep neural network (DNN) to transform its inputs (Wilson et al. 2016). Deep kernels combine the non-parametric flexibility of GPs with the structural properties of NNs to obtain more expressive kernel functions when compared to their base kernels. The kernel parameters ϕ (DNN parameters + base kernel hyperparameters) are tuned using the gradient of GP’s log marginal likelihood J_{GP} (Rasmussen and Williams 2005),

$$J_{GP}(\phi|\mathcal{D}) \propto \log |\mathbf{K}| - \mathbf{Q}^\top \mathbf{K}^{-1} \mathbf{Q}, \quad (9)$$

$$\nabla_\phi J_{GP} = \mathbf{Q}^\top \mathbf{K}^{-1} (\nabla_\phi \mathbf{K}) \mathbf{K}^{-1} \mathbf{Q} + \text{Tr}(\mathbf{K}^{-1} \nabla_\phi \mathbf{K}).$$

Scaling BQ to large sample sizes n : The complexity of estimating L_θ^{BQ} (Eq. 8) is largely influenced by the matrix-inversion operation $(\mathbf{K} + \sigma^2 \mathbf{I})^{-1}$, whose exact computation scales with an $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ space complexity. Our first step is to shift the focus from an expensive matrix-inversion operation to an approximate inverse matrix-vector multiplication (i-MVM). In particular, we use the conjugate gradient (CG) method to compute the i-MVM $(\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{v}$ within machine precision using $p \ll n$ iterations of MVM operations $\mathbf{K} \mathbf{v}' = c_1 \mathbf{K}_s \mathbf{v}' + c_2 \mathbf{K}_f \mathbf{v}'$.

While the CG method nearly reduces the time complexity by an order of n , the MVM operation with a dense matrix still incurs a prohibitive $\mathcal{O}(n^2)$ computational cost. Fortunately, the matrix factorization of \mathbf{K}_f (Eq. 7) allows for computing the Fisher kernel MVM $\mathbf{K}_f \mathbf{v}$ in linear-time through automatic differentiation, without the explicit creation or storage of the \mathbf{K}_f matrix (more details in Sec. D.1 of the supplement). On the other hand, since the choice of k_s is arbitrary, we deploy *structured kernel interpolation* (SKI) (Wilson and Nickisch 2015), a general kernel sparsification strategy to efficiently compute the state kernel MVM $\mathbf{K}_s \mathbf{v}$. SKI uses a set of $m \leq n$ “inducing points” $\{\hat{s}_i\}_{i=1}^m$ to approximate \mathbf{K}_s with a rank m matrix $\hat{\mathbf{K}}_s =$

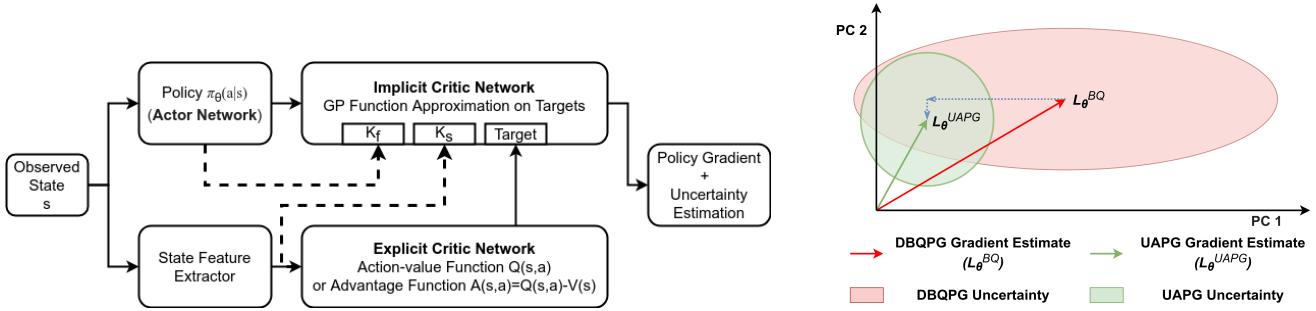


Figure 1: (Left) Overview of the DBQPG method, (Right) Illustration of DBQPG and UAPG updates along the first two principal components (PCs) of the gradient covariance matrix.

$\mathbf{W}\mathbf{K}_s^m\mathbf{W}^\top$, where \mathbf{K}_s^m is an $m \times m$ Gram matrix with entries $\mathbf{K}_s^m(p,q) = k_s(\hat{s}_p, \hat{s}_q)$, and \mathbf{W} is an $n \times m$ interpolation matrix whose entries depend on the relative placement of sample points $\{s_i\}_{i=1}^n$ and inducing points $\{\hat{s}_i\}_{i=1}^m$. In practice, a sparse \mathbf{W} matrix that follows local cubic interpolation (only 4 non-zero entries per row) provides a good approximation $\hat{\mathbf{K}}_s$, and more importantly, offers $\hat{\mathbf{K}}_s\mathbf{v}$ MVM in $\mathcal{O}(n+m^2)$ time and storage complexity.

Further, the SKI framework also provides the flexibility to select the inducing point locations for exploiting the structure of specialized GP kernels. For instance, one-dimensional stationary kernels, i.e., $k_s(x,y) = k_s(x-y)$, can additionally leverage the Toeplitz method (Turner 2010) by picking evenly-spaced inducing points. Since these matrices are constant along the diagonal, i.e. $\mathbf{K}_s(x,y) = \mathbf{K}_s(x+1,y+1)$, the Toeplitz method utilizes fast Fourier transform to attain an $\mathcal{O}(n+m \log m)$ time and $\mathcal{O}(n+m)$ storage for the MVM operation. Further, Toeplitz methods can be extended to multiple dimensions by assuming that the kernel decomposes as a sum of one-dimensional stationary kernels along each of the input dimensions. Compared to conventional inducing point methods that operate with $m \ll n$ inducing points, choosing a base kernel that conforms with the Toeplitz method enables the realization of larger m values, thereby providing a more accurate approximation of \mathbf{K}_s .

Practical DBQPG Algorithm: The DBQPG algorithm is designed to compute the gradient from a batch of samples (parallelly) leveraging the automatic differentiation framework and fast kernel computation methods, thus placing BQ-PG in the context of contemporary PG algorithms (see Fig. 1(left)). In contrast, the original BQ-PG method (Ghavamzadeh and Engel 2007) was designed to process the samples sequentially (slow). Other major improvements in DBQPG include (i) replacing a traditional inducing points method ($\mathcal{O}(m^2n+m^3)$ time and $\mathcal{O}(mn+m^2)$ storage) with SKI, a more efficient alternative ($\mathcal{O}(n+m^2)$ time and storage), and (ii) replacing a fixed base kernel for k_s with the more expressive deep kernel, followed by learning the kernel parameters (Eq. 9). The performance gains from deep kernel learning and SKI are documented in supplement Sec. G.

For DBQPG, our default choice for the prior state covariance function k_s is a deep RBF kernel which comprises of an RBF base kernel on top of a DNN feature extractor. Our

choice of RBF as the base kernel is based on: (i) its compelling theoretical properties such as infinite basis expansion and universal function approximation (Micchelli, Xu, and Zhang 2006) and (ii) its compatibility with the Toeplitz method. Thus, the overall computational complexity of estimating \mathbf{L}_θ^{BQ} (Eq. 8) is $\mathcal{O}(p(n+Ym \log m))$ time and $\mathcal{O}(n+Ym)$ storage (Y : state dimensionality; p : CG iterations; m : number of inducing points in SKI+Toeplitz approximation for a deep RBF kernel; automatic differentiation for Fisher kernel).

Note that we intentionally left out kernel learning from the complexity analysis since a naive implementation of the gradient-based optimization step (Eq. 9) incurs a cubic time complexity in sample size. Our implementation relies on the black-box matrix-matrix multiplication (BBMM) feature (a batched version of CG algorithm that effectively uses GPU hardware) offered by the *GPyTorch library* (Gardner et al. 2018), coupled with a SKI approximation over \mathbf{K}_s . This combination offers a linear-time routine for kernel learning. In other words, DBQPG with kernel learning is a linear-time program that leverages GPU acceleration to efficiently estimate the gradient of a large policy network, with a few thousands of parameters, on high-dimensional domains.

5 Uncertainty Aware Policy Gradient

We propose UAPG, a novel uncertainty aware PG method that utilizes the gradient uncertainty \mathbf{C}_θ^{BQ} from DBQPG to provide more reliable policy updates. Most classical PG methods consider stochastic gradient estimates as the true gradient, without accounting the uncertainty in their gradient components, thus, occasionally taking large steps along the directions of high uncertainty. UAPG provides more reliable policy updates by lowering the stepsize along the directions with high uncertainty. In particular, UAPG uses \mathbf{C}_θ^{BQ} to normalize the components of \mathbf{L}_θ^{BQ} with their respective uncertainties, bringing them all to the same scale. Thus, UAPG offers PG estimates with a uniform uncertainty, i.e. their gradient covariance is the identity matrix. See Fig. 1 (right). In theory, the UAPG update can be formulated as $(\mathbf{C}_\theta^{BQ})^{-\frac{1}{2}}\mathbf{L}_\theta^{BQ}$.

Practical UAPG Algorithm: Empirical \mathbf{C}_θ^{BQ} estimates are often ill-conditioned matrices (spectrum decays quickly)

with a numerically unstable inversion. Since C_θ^{BQ} only provides a faithful estimate of the top few principal directions of uncertainty, the UAPG update is computed from a rank- δ singular value decomposition (SVD) approximation of $C_\theta^{BQ} \approx \nu_\delta \mathbf{I} + \sum_{i=1}^{\delta} \mathbf{h}_i (\nu_i - \nu_\delta) \mathbf{h}_i^\top$ as follows:

$$\mathbf{L}_\theta^{UAPG} = \nu_\delta^{-\frac{1}{2}} \left(\mathbf{I} + \sum_{i=1}^{\delta} \mathbf{h}_i (\sqrt{\nu_\delta / \nu_i} - 1) \mathbf{h}_i^\top \right) \mathbf{L}_\theta^{BQ}. \quad (10)$$

The principal components (PCs) $\{\mathbf{h}_i\}_{i=1}^{\delta}$ denote the top δ directions of uncertainty and the singular values $\{\nu_i\}_{i=1}^{\delta}$ denote their corresponding magnitudes of uncertainty. The rank- δ decomposition of C_θ^{BQ} can be computed in linear-time using the randomized SVD algorithm (Halko, Martinson, and Tropp 2011). The UAPG update in Eq. 10 dampens the stepsize of the top δ directions of uncertainty, relative to the stepsize of remaining gradient components. Thus, in comparison to DBQPG, UAPG lowers the risk of taking large steps along the directions of high uncertainty, providing reliable policy updates.

For natural gradient $\mathbf{L}_\theta^{NBQ} = \mathbf{G}^{-1} \mathbf{L}_\theta^{BQ}$, the gradient uncertainty can be computed as follows:

$$\begin{aligned} \mathbf{C}_\theta^{NBQ} &= \mathbf{G}^{-1} \mathbf{C}_\theta^{BQ} \mathbf{G}^{-1} \\ &= c_2 \mathbf{G}^{-1} - c_2^2 \mathbf{G}^{-1} \mathbf{U} (c_1 \mathbf{K}_s + c_2 \mathbf{K}_f + \sigma^2 \mathbf{I})^{-1} \mathbf{U}^\top \mathbf{G}^{-1} \\ &= c_2 (\mathbf{G} + c_2 \mathbf{U} (c_1 \mathbf{K}_s + \sigma^2 \mathbf{I})^{-1} \mathbf{U}^\top)^{-1}. \end{aligned} \quad (11)$$

Since \mathbf{C}_θ^{NBQ} is the inverse of an ill-conditioned matrix, we instead apply a low-rank approximation on $\mathbf{C}_\theta^{NBQ^{-1}} \approx \nu_\delta \mathbf{I} + \sum_{i=1}^{\delta} \mathbf{h}_i (\nu_i - \nu_\delta) \mathbf{h}_i^\top$ for the UAPG update of NPG:

$$\mathbf{L}_\theta^{NUAPG} = \nu_\delta^{\frac{1}{2}} \left(\mathbf{I} + \sum_{i=1}^{\delta} \mathbf{h}_i \left(\min \left(\sqrt{\frac{\nu_i}{\nu_\delta}}, \epsilon \right) - 1 \right) \mathbf{h}_i^\top \right) \mathbf{G}^{-1} \mathbf{L}_\theta^{BQ}, \quad (12)$$

where, $\{\mathbf{h}_i, \nu_i\}_{i=1}^{\delta}$ correspond to the top δ PCs of $\mathbf{C}_\theta^{NBQ^{-1}}$ (equivalently the bottom δ PCs of \mathbf{C}_θ^{NBQ}), and $\epsilon > 1$ is a hyperparameter. We replace $\sqrt{\nu_i / \nu_\delta}$ with $\min(\sqrt{\nu_i / \nu_\delta}, \epsilon)$ to avoid taking large steps along these directions, solely on the basis of their uncertainty estimates. For $c_2 \ll 1$, it is interesting to note that (i) $\mathbf{C}_\theta^{BQ} \approx c_2 \mathbf{G}$ and $\mathbf{C}_\theta^{NBQ} \approx c_2 \mathbf{G}^{-1}$, which implies that the most uncertain gradient directions for vanilla PG approximately correspond to the most confident directions for NPG, and (ii) the ideal UAPG update for both vanilla PG and NPG converges along the $\mathbf{G}^{-\frac{1}{2}} \mathbf{L}_\theta^{BQ}$ direction. A more rigorous discussion on the relations between Vanilla PG, NPG and their UAPG updates can be found in the supplement Sec. E.1.

6 Experiments

We study the behaviour of BQ-PG methods (Algorithm 1) on MuJoCo environments, using the *mujoco-py* library of OpenAI Gym (Brockman et al. 2016). In our experiments, we replace Q with generalized advantage estimates (Schulman et al. 2016) computed using an explicit state-value critic

Algorithm 1 BQ-PG Estimator Subroutine

- 1: **BQ-PG**(θ, n)
 - θ : policy parameters
 - n : sample size for PG estimation
 - 2: Collect n state-action pairs (samples) from running the policy π_θ in the environment.
 - 3: Compute action-values Q for these n samples using MC rollouts (TD(1) estimate) or an explicit critic network.
 - 4: Update kernel parameters and explicit critic’s parameters using GP MLL (Eq. 9) and TD error respectively.
 - 5: Policy gradient estimation (DBQPG):

$$\mathbf{L}_\theta = \begin{cases} \mathbf{L}_\theta^{BQ} & \text{(Vanilla PG)} \\ \mathbf{G}^{-1} \mathbf{L}_\theta^{BQ} & \text{(Natural PG)} \end{cases}$$
 - 6: **(Optional)** Compute $\{\mathbf{h}_i, \nu_i\}_{i=1}^{\delta}$ using fast SVD over \mathbf{C}_θ^{BQ} (Eq. 8, vanilla PG) or $\mathbf{C}_\theta^{NBQ^{-1}}$ (Eq. 11, NPG).
 - 7: **(Optional)** Uncertainty-based adjustment (UAPG):

$$\mathbf{L}_\theta = \begin{cases} \nu_\delta^{-\frac{1}{2}} \left(\mathbf{I} + \sum_{i=1}^{\delta} \mathbf{h}_i \left(\sqrt{\frac{\nu_i}{\nu_\delta}} - 1 \right) \mathbf{h}_i^\top \right) \mathbf{L}_\theta & \text{(Vanilla PG)} \\ \nu_\delta^{\frac{1}{2}} \left(\mathbf{I} + \sum_{i=1}^{\delta} \mathbf{h}_i \left(\min \left(\sqrt{\frac{\nu_i}{\nu_\delta}}, \epsilon \right) - 1 \right) \mathbf{h}_i^\top \right) \mathbf{L}_\theta & \text{(Natural PG)} \end{cases}$$
 - 8: **return** \mathbf{L}_θ
-

network, i.e., $V_{\pi_\theta}(s)$ is approximated using a linear layer on top of the state feature extractor in Fig. 1(left).

Quality of Gradient Estimation: Inspired from the experimental setup of Ilyas et al. (2018), we evaluate the quality of PG estimates obtained via DBQPG and MC estimation using two metrics: (i) **gradient accuracy** or the average cosine similarity of the obtained gradient estimates with respect to the true gradient estimates (estimated from 10^6 state-action pairs) and (ii) **variance** in the gradient estimates (normalized by the norm of the mean gradient for scale invariance). See Fig. 2. We observe that DBQPG provides more accurate gradient estimates with a considerably lower variance. Interestingly, DBQPG and MC estimates offer nearly the same quality gradients at the start of training. However, as the training progresses, and DBQPG learns kernel bases, we observe that DBQPG returns superior quality gradient estimates. Moreover, as training progress from 0 to 150 iterations, the gradient norms of both DBQPG and MC estimates drop by a factor of 3, while the “unnormalized” gradient variances increase by 5 folds. The drop in the signal-to-noise ratio for gradient estimation explains the fall in accuracy over training time. Further, the wall-clock time of DBQPG and UAPG updates is comparable to MC-PG (Fig. 6 in supplement). These results motivate substituting MC with BQ-based PG estimates in deep PG algorithms.

Compatibility with Deep PG Algorithms: We examine the compatibility of BQ-based methods with the following on-policy deep policy gradient algorithms: (i) Vanilla policy gradient, (ii) natural policy gradient (NPG), and (iii) trust region policy optimization (TRPO), as shown in Fig. 3. In these experiments, only the MC estimation subroutine is replaced with BQ-based methods, keeping the rest of the al-

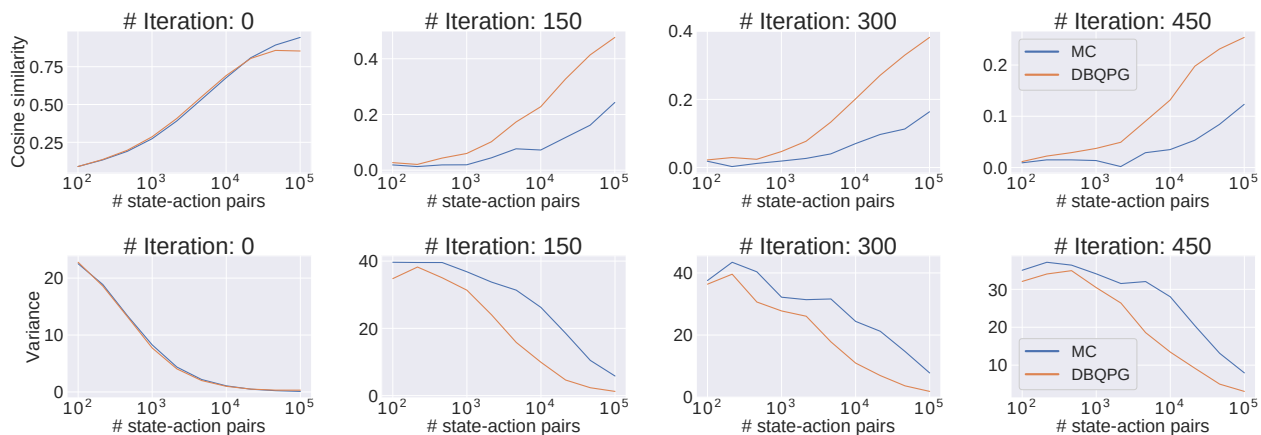


Figure 2: An empirical analysis of the quality of policy gradient estimates as a function of the state-action sample size. The experiments are conducted for 0^{th} , 150^{th} , 300^{th} , and 450^{th} iteration along the training phase of DBQPG (vanilla PG) algorithm in MuJoCo Swimmer-v2 environment. All the results have been averaged over 25 repeated gradient measurements across 100 random runs. (a) The accuracy plot results are obtained w.r.t the “true gradient”, which is computed using MC estimates of 10^6 state-action pairs. (b) The normalized variance is computed using the ratio of trace of empirical gradient covariance matrix (like Zhao et al. (2011)) and squared norm of gradient mean.

gorithm unchanged. Note that for TRPO, we use the UAPG update of NPG to compute the step direction. We observe that DBQPG consistently outperforms MC estimation, both in final performance and sample complexity across all the deep PG algorithms. This observation resonates with our previous finding of the superior gradient quality of DBQPG estimates, and strongly advocates the use of DBQPG over MC for PG estimation.

For UAPG, we observe a similar trend as DBQPG. The advantage of UAPG estimates is more pronounced in the vanilla PG, and NPG experiments since the performance on these algorithms is highly sensitive to the choice of learning rates. UAPG adjusts the stepsize of each gradient component based on its uncertainty, resulting in a robust update in the face of uncertainty, a better sample complexity, and average return. We observe that UAPG performs at least as good as, if not considerably better than, DBQPG on most experiments. Since TRPO updates are less sensitive to the learning rate, UAPG adjustment does not provide a significant improvement over DBQPG.

7 Related Work

The high sample complexity of MC methods has been a long-standing problem in the PG literature (Rubinstein 1969). Previous approaches that address this issue broadly focus on two aspects: (i) improving the quality of PG estimation using a value function approximation (Konda and Tsitsiklis 2003), or (ii) attaining faster convergence by robustly taking larger steps in the right direction. The former class of approaches trade a tolerable level of bias for designing a lower variance PG estimator (Schulman et al. 2016; Reisinger, Stone, and Miikkulainen 2008). Following the latter research direction, Kakade (2001) and Kakade and Langford (2002) suggest replacing the vanilla PG with natural policy gradient (NPG), the steepest descent direc-

tion in the policy distribution space. While NPG improves over vanilla PG methods in terms of sample complexity (Peters and Schaal 2008; Agarwal et al. 2019), it is just as vulnerable to catastrophic policy updates. Trust region policy optimization (TRPO) (Schulman et al. 2015) extends the NPG algorithm with a robust stepsize selection mechanism that guarantees monotonic improvements for expected (true) policy updates. However, the practical TRPO algorithm loses its improvement guarantees for stochastic PG estimates, thereby necessitating a large sample size for computing reliable policy updates. The advantages of these approaches, in terms of sample efficiency, are orthogonal to the benefits of DBQPG and UAPG methods.

Another line of research focuses on using Gaussian processes (GP) to directly approximate the PG integral (Ghavamzadeh and Engel 2006). This work was followed by the Bayesian Actor-Critic (BAC) algorithm (Ghavamzadeh and Engel 2007), which exploits the MDP framework for improving the statistical efficiency of PG estimation. Like DBQPG, BAC is a BQ-based PG method that uses a GP to approximate the action-value function. However, BAC is an online algorithm that uses Gaussian process temporal difference (GPTD) (Engel, Mannor, and Meir 2005), a sequential kernel sparsification method, for learning the value function. BAC’s sequential nature and a prohibitive $\mathcal{O}(m^2n + m^3)$ time and $\mathcal{O}(mn + m^2)$ storage complexity (m is the dictionary size, i.e., the number of inducing points) prevents it from scaling to large non-linear policies and high-dimensional continuous domains. Further, a recent extension of BAC (Ghavamzadeh, Engel, and Valko 2016) uses the uncertainty to adjust the learning rate of policy parameters, $(\mathbf{I} - \frac{1}{\nu} \mathbf{C}_\theta^{BQ}) \mathbf{L}_\theta^{BQ}$ (ν is an upper bound for \mathbf{C}_θ^{BQ}), much like the UAPG method. While this update reduces the step-size of more uncertain directions, it does not provide gradient estimates with uniform uncertainty, making it less

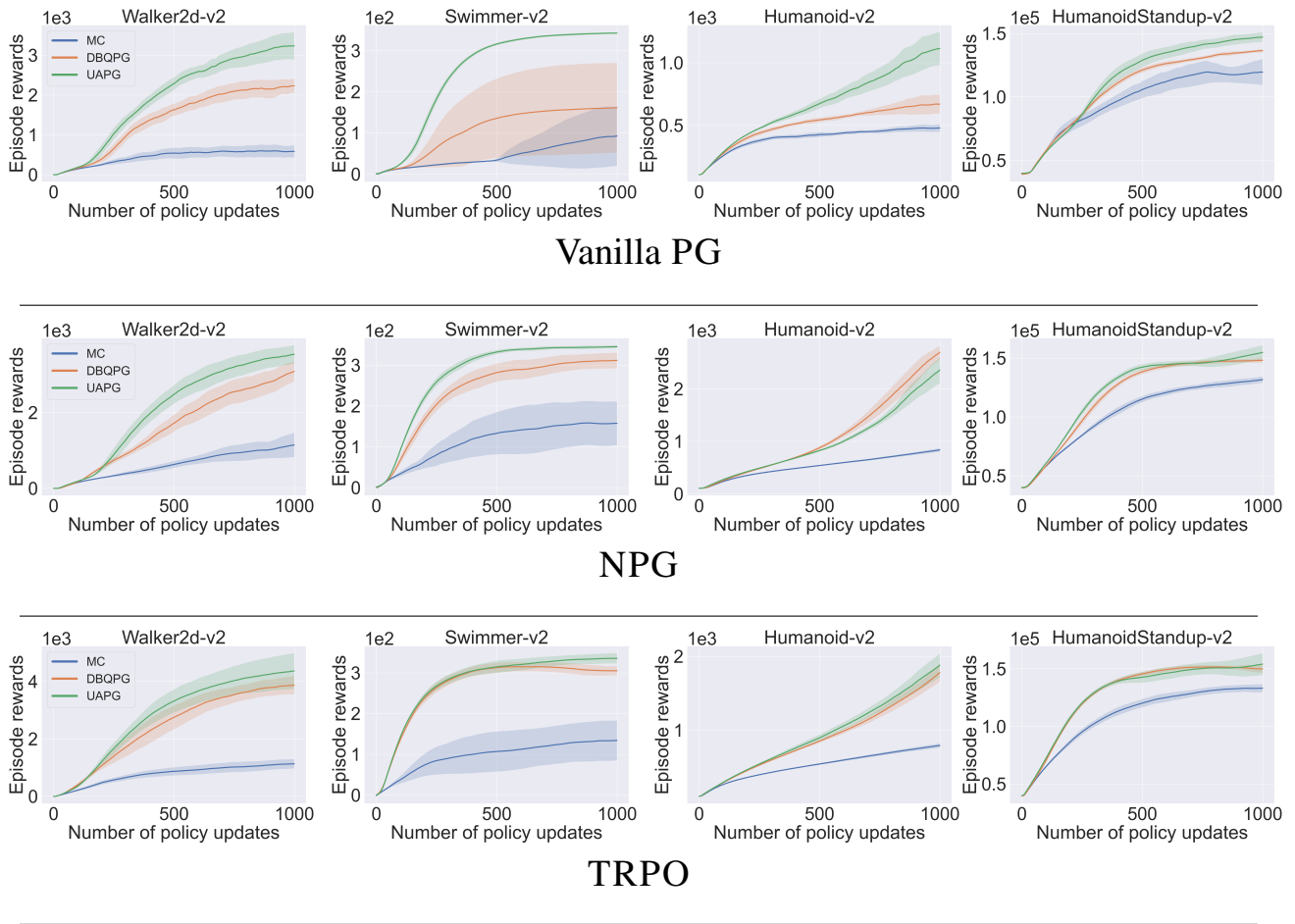


Figure 3: The average performance (10 runs) of BQ-based methods and MC estimation in vanilla PG, NPG, and TRPO frameworks across 4 MuJoCo environments (refer Fig. 9 in the supplement for comparison across all 7 MuJoCo environments).

robust to bad policy updates relative to UAPG. Moreover, this method would not work for NPG and TRPO algorithms as their covariance is the inverse of an ill-conditioned matrix. Refer supplement Sec. G for a more detailed comparison of DBQPG and UAPG with prior BQ-PG works.

8 Discussion

We study the problem of estimating accurate policy gradient (PG) estimates from a finite number of samples. This problem becomes relevant in numerous RL applications where an agent needs to estimate the PG using samples gathered through interaction with the environment. Monte-Carlo (MC) methods are widely used for PG estimation despite offering high-variance gradient estimates and a slow convergence rate. We propose DBQPG, a high-dimensional generalization of Bayesian quadrature that, like MC method, estimates PG in linear-time. We empirically study DBQPG and demonstrate its effectiveness over MC methods. We show that DBQPG provides more accurate gradient estimates, along with a significantly smaller variance in the gradient estimation. Next, we show that replacing the MC method

with DBQPG in the gradient estimation subroutine of three deep PG algorithms, viz., Vanilla PG, NPG, and TRPO, offers consistent gains in sample complexity and average return. These results strongly recommend the substitution of MC estimator with DBQPG in deep PG algorithms.

To obtain reliable policy updates from stochastic gradient estimates, one needs to estimate the uncertainty in gradient estimation, in addition to the gradient estimation itself. The proposed DBQPG method additionally provides this uncertainty along with the PG estimate. We propose UAPG, a PG method that uses DBQPG’s uncertainty to normalize the gradient components by their uncertainties, returning a uniformly uncertain gradient estimate. Such a normalization will lower the step size of gradient components with high uncertainties, resulting in reliable updates with robust step sizes. We show that UAPG further improves the sample complexity over DBQPG on Vanilla PG, NPG, and TRPO algorithms. Overall, our study shows that it is possible to scale Bayesian quadrature to high-dimensional settings, while its better gradient estimation and well-calibrated gradient uncertainty can significantly boost the performance of deep PG algorithms.

Ethics Statement

When deploying deep policy gradient (PG) algorithms for learning control policies in physical systems, sample efficiency becomes an important design criteria. In the past, numerous works have focused on improving the sample efficiency of PG estimation through variance reduction, robust stepsize selection, etc. In this paper, we propose deep Bayesian quadrature policy gradient (DBQPG), a statistically efficient policy gradient estimator that offers orthogonal benefits for improving the sample efficiency. In comparison to Monte-Carlo estimation, the default choice for PG estimation, DBQPG returns more accurate gradient estimates with much lower empirical variance. Since DBQPG is a general gradient estimation subroutine, it can directly replace Monte-Carlo estimation in most policy gradient algorithms, as already demonstrated in our paper. Therefore, we think that the DBQPG method directly benefits most policy gradient algorithms and is indirectly beneficial for several downstream reinforcement learning applications.

We also propose uncertainty aware policy gradient (UAPG), a principled approach for incorporating the uncertainty in gradient estimation (also quantified by the DBQPG method) to obtain reliable PG estimates. UAPG lowers the risk of catastrophic performance degradation with stochastic policy updates, and empirically performs at least as good as, if not better than, the DBQPG method. Hence, we believe that the UAPG method is more relevant to reinforcement learning applications with safety considerations, such as robotics.

References

- Agarwal, A.; Kakade, S. M.; Lee, J. D.; and Mahajan, G. 2019. Optimality and approximation with policy gradient methods in markov decision processes. *arXiv preprint arXiv:1908.00261*.
- Azizzadenesheli, K.; and Anandkumar, A. 2018. Efficient Exploration through Bayesian Deep Q-Networks. *arXiv preprint arXiv:1802.04412*.
- Baxter, J.; and Bartlett, P. L. 2000. Direct gradient-based reinforcement learning. In *2000 IEEE International Symposium on Circuits and Systems. Emerging Technologies for the 21st Century. Proceedings (IEEE Cat No. 00CH36353)*, volume 3, 271–274. IEEE.
- Briol, F.-X.; Oates, C.; Girolami, M.; and Osborne, M. A. 2015. Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. In *Advances in Neural Information Processing Systems*, 1162–1170.
- Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. OpenAI Gym.
- Duan, Y.; Chen, X.; Houthoofd, R.; Schulman, J.; and Abbeel, P. 2016. Benchmarking Deep Reinforcement Learning for Continuous Control. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16*, 1329–1338. JMLR.org. URL <http://dl.acm.org/citation.cfm?id=3045390.3045531>.
- Engel, Y.; Mannor, S.; and Meir, R. 2003. Bayes Meets Bellman: The Gaussian Process Approach to Temporal Difference Learning. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning, ICML'03*, 154–161. AAAI Press. ISBN 1-57735-189-4. URL <http://dl.acm.org/citation.cfm?id=3041838.3041858>.
- Engel, Y.; Mannor, S.; and Meir, R. 2005. Reinforcement Learning with Gaussian Processes. In *Proceedings of the 22Nd International Conference on Machine Learning, ICML '05*, 201–208. New York, NY, USA: ACM. ISBN 1-59593-180-5. doi:10.1145/1102351.1102377. URL <http://doi.acm.org/10.1145/1102351.1102377>.
- Gardner, J. R.; Pleiss, G.; Bindel, D.; Weinberger, K. Q.; and Wilson, A. G. 2018. GPyTorch: Blackbox Matrix-matrix Gaussian Process Inference with GPU Acceleration. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems, NIPS'18*, 7587–7597. USA: Curran Associates Inc. URL <http://dl.acm.org/citation.cfm?id=3327757.3327857>.
- Ghavamzadeh, M.; and Engel, Y. 2006. Bayesian Policy Gradient Algorithms. In *Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS'06*, 457–464. Cambridge, MA, USA: MIT Press.
- Ghavamzadeh, M.; and Engel, Y. 2007. Bayesian Actor-critic Algorithms. In *Proceedings of the 24th International Conference on Machine Learning, ICML '07*, 297–304. New York, NY, USA: ACM. ISBN 978-1-59593-793-3. doi:10.1145/1273496.1273534. URL <http://doi.acm.org/10.1145/1273496.1273534>.
- Ghavamzadeh, M.; Engel, Y.; and Valko, M. 2016. Bayesian Policy Gradient and Actor-Critic Algorithms. *Journal of Machine Learning Research* 17(66): 1–53. URL <http://jmlr.org/papers/v17/10-245.html>.
- Halko, N.; Martinsson, P.-G.; and Tropp, J. A. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review* 53(2): 217–288.
- Hennig, P.; Osborne, M. A.; and Girolami, M. 2015. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 471(2179).
- Ilyas, A.; Engstrom, L.; Santurkar, S.; Tsipras, D.; Janoos, F.; Rudolph, L.; and Madry, A. 2018. Are Deep Policy Gradient Algorithms Truly Policy Gradient Algorithms? *ArXiv abs/1811.02553*.
- Kakade, S. 2001. A Natural Policy Gradient. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, NIPS'01*, 1531–1538. Cambridge, MA, USA: MIT Press. URL <http://dl.acm.org/citation.cfm?id=2980539.2980738>.
- Kakade, S.; and Langford, J. 2002. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, 267–274.

- Kanagawa, M.; Sriperumbudur, B. K.; and Fukumizu, K. 2016. Convergence guarantees for kernel-based quadrature rules in misspecified settings. In *Advances in Neural Information Processing Systems*, 3288–3296.
- Kanagawa, M.; Sriperumbudur, B. K.; and Fukumizu, K. 2020. Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. *Foundations of Computational Mathematics* 20(1): 155–194.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. URL <http://arxiv.org/abs/1412.6980>.
- Konda, V. R.; and Tsitsiklis, J. N. 2000. Actor-Critic Algorithms. In Solla, S. A.; Leen, T. K.; and Müller, K., eds., *Advances in Neural Information Processing Systems 12*, 1008–1014. MIT Press. URL <http://papers.nips.cc/paper/1786-actor-critic-algorithms.pdf>.
- Konda, V. R.; and Tsitsiklis, J. N. 2003. On Actor-Critic Algorithms. *SIAM J. Control Optim.* 42(4): 1143–1166. ISSN 0363-0129. doi:10.1137/S0363012901385691. URL <https://doi.org/10.1137/S0363012901385691>.
- Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.
- Metropolis, N.; and Ulam, S. 1949. The monte carlo method. *Journal of the American statistical association* 44(247): 335–341.
- Micchelli, C. A.; Xu, Y.; and Zhang, H. 2006. Universal Kernels. *J. Mach. Learn. Res.* 7: 2651–2667. ISSN 1532-4435.
- O’Hagan, A. 1991. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference* 29(3): 245–260. URL <http://www.sciencedirect.com/science/article/B6V0M-45F5GDM-53/1/6e05220bfd4a6174e890f60bb391107c>.
- Peters, J.; and Schaal, S. 2008. Reinforcement Learning of Motor Skills with Policy Gradients. *Neural Networks* 21(4): 682–697.
- Rasmussen, C. E.; and Williams, C. K. I. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press. ISBN 026218253X.
- Reisinger, J.; Stone, P.; and Miiikkulainen, R. 2008. Online Kernel Selection for Bayesian Reinforcement Learning. In *Proceedings of the 25th International Conference on Machine Learning, ICML ’08*, 816–823. New York, NY, USA: Association for Computing Machinery. ISBN 9781605582054.
- Rubinstein, R. Y. 1969. Some Problems in Monte Carlo Optimization. *Ph.D. thesis*.
- Saatci, Y. 2012. *Scalable Inference for Structured Gaussian Process Models*. University of Cambridge. URL <https://books.google.co.in/books?id=9pC3oQEACAAJ>.
- Schulman, J.; Levine, S.; Moritz, P.; Jordan, M. I.; and Abbeel, P. 2015. Trust Region Policy Optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*.
- Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; and Abbeel, P. 2016. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Silverman, B. W. 1985. Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting. *Journal of the Royal Statistical Society. Series B (Methodological)* 47(1): 1–52. ISSN 00359246. URL <http://www.jstor.org/stable/2345542>.
- Sutton, R. S.; McAllester, D.; Singh, S.; and Mansour, Y. 2000. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems, NIPS’99*, 1057–1063. Cambridge, MA, USA: MIT Press. URL <http://dl.acm.org/citation.cfm?id=3009657.3009806>.
- Todorov, E.; Erez, T.; and Tassa, Y. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026–5033. IEEE.
- Townsend, J. 2017. A new trick for calculating Jacobian vector products. URL <https://j-towns.github.io/2017/06/12/A-new-trick.html>.
- Turner, R. E. 2010. *Statistical Models for Natural Sounds*. Ph.D. thesis, Gatsby Computational Neuroscience Unit, UCL.
- Williams, R. J. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Mach. Learn.* 8(3-4): 229–256. ISSN 0885-6125. doi:10.1007/BF00992696. URL <https://doi.org/10.1007/BF00992696>.
- Wilson, A. G.; Hu, Z.; Salakhutdinov, R.; and Xing, E. P. 2016. Deep Kernel Learning. In Gretton, A.; and Robert, C. C., eds., *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, 370–378. Cadiz, Spain: PMLR. URL <http://proceedings.mlr.press/v51/wilson16.html>.
- Wilson, A. G.; and Nickisch, H. 2015. Kernel Interpolation for Scalable Structured Gaussian Processes (KISS-GP). In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML’15*, 1775–1784. JMLR.org. URL <http://dl.acm.org/citation.cfm?id=3045118.3045307>.
- Woodbury, M. A. 1950. Inverting modified matrices. *Memorandum report* 42(106): 336.
- Zhao, T.; Hachiya, H.; Niu, G.; and Sugiyama, M. 2011. Analysis and Improvement of Policy Gradient Estimation. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.