All-Action Policy Gradient Methods: A Numerical Integration Approach

Benjamin Petit
Stanford University
benpetit@stanford.edu

Loren Amdahl-Culleton
Stanford University
lkac@stanford.edu

Yao Liu
Stanford University
yaoliu@stanford.edu

Jimmy Smith
Stanford University
jsmith14@stanford.edu

Pierre-Luc Bacon
Stanford University
plbacon@cs.stanford.edu

Abstract

While often stated as an instance of the likelihood ratio trick [Rubinstein, 1989], the original policy gradient theorem [Sutton et al., 1999] involves an integral over the action space. When this integral can be computed, the resulting “all-action" estimator [Sutton et al., 2001] provides a conditioning effect [Bratley et al., 1987] reducing the variance significantly compared to the REINFORCE estimator [Williams, 1992]. In this paper, we adopt a numerical integration perspective to broaden the applicability of the all-action estimator to general spaces and to any function class for the policy or critic components, beyond the Gaussian case considered by Ciosek and Whiteson [2018]. In addition, we provide a new theoretical result on the effect of using a biased critic which offers more guidance than the previous “compatible features" condition of Sutton et al. [1999]. We demonstrate the benefit of our approach in continuous control tasks with nonlinear function approximation. Our results show improved performance and sample efficiency.

1 Introduction

Likelihood ratio (LR) gradient estimators [Aleksandrov et al., 1968, Glynn, 1987, Reiman and Weiss, 1989, Rubinstein, 1989] have been widely used in reinforcement learning [Sutton and Barto, 2018] since the seminal work of Williams [1992] in the class of policy gradient methods [Kimura et al., 1995, 1997, Kimura and Kobayashi, 1998, Marbach and Tsitsiklis, 1998, Sutton et al., 1999, Konda and Tsitsiklis, 2000, Baxter and Bartlett, 2001]. The popularity of LR methods stems from its ease of implementation [Schulman et al., 2015] and its applicability to both discrete and continuous actions spaces [Konda and Tsitsiklis, 2000], in the batch or online settings [Baxter and Bartlett, 2001].

However, likelihood ratio methods may also suffer [Mania et al., 2018] from high variance in the long horizon setting [L’Ecuyer, 1991] or when the change of measure fails to have full support [L’Ecuyer, 1990]. Hence, variance reduction techniques [Bratley et al., 1987, L’Ecuyer, 1994] must usually be used in conjunction with the LR method [L’Ecuyer, 1991]. Three main strategies are employed in practice, namely: 1) the *baseline* approach [Williams, 1992] 2) by leveraging the Markovian structure [Williams, 1992, Glynn and L’Ecuyer, 1995] 3) by chunking a long trajectory into smaller *replications* via a *regenerative* state [L’Ecuyer, 1991, Baxter and Bartlett, 2001, Konda and Tsitsiklis, 2000] or by truncating the horizon with a smaller discount factor [Baxter and Bartlett, 2001].

An intuitive but lesser known variance reduction technique is that of *conditioning*, stemming from *conditional Monte Carlo* methods [Hammersley and Handscomb, 1964, Bratley et al., 1987]. At a high level: if there is a Y such that $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$ then by the law of iterated expectation the

Optimization Foundations for Reinforcement Learning Workshop at NeurIPS 2019, Vancouver, Canada.

variance may be reduced by computing $\mathbb{E}[X|Y]$ separately. Conditioning is exactly the principle at play behind the so-called *expected* methods [van Seijen et al., 2009, Ciosek and Whiteson, 2018, Sutton and Barto, 2018] in RL (although this connection had never been stated explicitly before).

Interestingly, the very statement of the *policy gradient theorem* by [Sutton et al., 1999, theorem 1] hints clearly at an application of the conditioning technique, yet this approach has not been widely used in practice, nor analyzed properly. In an unfinished paper, Sutton et al. [2001], posit the superiority of this approach in what they call the *all-action* policy gradient estimator and lay out an agenda to show this formally and in practice (but never provided those results). For more than a decade, the *all-action* form has been mostly forgotten in favor of the *single-action* LR variant and was only re-discovered recently in [Allen et al., 2017, Ciosek and Whiteson, 2018, Fellows et al., 2018].

In this paper, we provide a first explanation as to why all-action methods may improve the variance of policy gradient estimators by establishing a connection to conditional Monte Carlo methods. Using ideas from the numerical integration literature [Forsythe et al., 1977], we then propose general policy gradient methods capable of implementing all-action estimators in continuous action spaces. More precisely, we investigate the use of quadrature formula and Monte Carlo integration methods for approximating the conditional expectation over actions involved in the policy gradient theorem. This approach is flexible and does not require a specific parameterization of the *actor* or *critic* components as in Ciosek and Whiteson [2018]. We show that our perspective applies readily to nonlinear function approximators and can scale to the challenging environments in MuJoCo [Todorov et al., 2012].

We also provide a number of new theoretical results pertaining to the use of a biased critic in policy gradient methods. These new insights contribute to a better understanding of policy gradient methods with general function approximation, beyond the limited scope of the *compatible features* theorem [Sutton et al., 1999, theorem 2] for linear critics. In particular, theorem 3 is a general result on the expected dynamics of stochastic gradient ascent for a biased critic. It shows that if the bias term can be controlled, then a good solution may still be obtained. This result mirrors a similar condition for SGD requiring the noise to vanish in the limit [Bertsekas, 2016]. In the case of Monte-Carlo integration, theorem 1 provides a bound on the mean squared error of the all-action estimator with a learned critic and is shown to decrease as a function of the number of sampled actions. This error is finally compared with that of the LR estimator based on rollouts in theorem 2.

2 Background and Notation

The following presentation is based on the the Markov Decision Process (MDP) framework in the infinite horizon discounted setting [Puterman, 1994]. We assume a continuous discounted MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ (more details can be found in the supplementary material).

Policy gradient methods seek to identify an optimal policy by searching within a designated parameterized set of policies by gradient ascent. The concept of optimality in this case is defined with respect to the expected discounted return from a given initial state distribution. We write $\pi^* = \arg\max_\pi J(\pi)$ to denote an optimal policy where $J(\pi) \equiv \mathbb{E}_\tau[G(\tau)] \equiv \mathbb{E}_\tau\left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t)\right]$. The policy gradient theorem [Sutton et al., 1999] provides an expression for the gradient of expected return with respect to the parameters of a policy. Let $d_{\pi_\theta} \in \mathcal{M}(\mathcal{S})$ [Konda and Tsitsiklis, 2000] be the *discounted stationary distribution* of state under policy π, and $J(\theta) \equiv J(\pi)$ when π is parameterized by θ. The policy gradient theorem [Sutton et al., 1999, theorem 1] (henceforth referred to as PGT) states that:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s \sim d_{\pi_\theta}(\cdot)}\left[\int_{a \in A} \frac{\partial \pi_\theta(a|s)}{\partial \theta} Q^{\pi_\theta}(s,a) da\right] \equiv \mathbb{E}_{s \sim d_{\pi_\theta}(\cdot)}[Z(s,\theta)] \quad . \tag{1}$$

Because the term within the expectation involves an integral over the action space, it is often more convenient to proceed to a *change of measure* [L'Ecuyer, 1990] by the likelihood ratio approach:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s \sim d_{\pi_\theta}(\cdot)}\left[\mathbb{E}_{a \sim \pi_\theta(\cdot|s)}\left[\frac{\partial \log \pi_\theta(a|s)}{\partial \theta} Q^{\pi_\theta}(s,a)\right]\right] \equiv \mathbb{E}_s\left[\mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[z(s,a,\theta)]\right] \quad . \tag{2}$$

In this paper, we assume that $\nabla \pi_\theta / \pi_\theta$ is uniformly bounded to ensure that the change of measure are well defined. Policy search methods [Szepesvari, 2010] – such as REINFORCE [Williams, 1992] –

use sample future returns $G_t = \sum_{s=t}^{T} \gamma^{s-t} r_s$ (rollouts) in an estimator of the form:

$$\nabla_\theta J(\theta) \simeq \frac{1}{T} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) G_t = \frac{1}{T} \sum_{t=0}^{T-1} \hat{z}(s_t, a_t, \theta) \ \ , \tag{3}$$

whereas actor-critic methods [Sutton, 1984] use an estimator of Q^{π_θ} [Konda and Tsitsiklis, 2000].

3 All-Action Estimators and Conditional Monte Carlo Methods

By a change of measure via the log trick, we have seen that the inner integral in (1) can be transformed into an expectation which can be sampled from along the stationary distribution. While sampling a single action for every state visited along a trajectory would suffice to obtain an unbiased estimator, a better approximation of the inner expectation term is obtained by sampling a larger number of actions. Sutton et al. [2001] refer to the former type methods as the *single-action* ones whereas the latter are called *all-action methods*. As shown below, all-action methods are preferable due to their variance reduction effect. However, when the action space is large (or continuous) it may become intractable to compute the inner expectation exactly unless an analytical expression is known apriori as in [Ciosek and Whiteson, 2018]. We address this issue using numerical integration methods.

The variance reduction brought by all-action estimators can be simply understood using the law of total variance [Bratley et al., 1987]. Let $\hat{Z}(s)$ be a random variable such that $\mathbb{E}\left[\hat{Z}(s)|s\right] = Z(s, \theta)$ (an unbiased estimator of the policy gradient) then the variance of $\hat{Z}(s)$ is:

$$\text{Var}\left[\hat{Z}(s)\right] = \text{Var}_s\left[Z(s,\theta)\right] + \mathbb{E}_s\left[\text{Var}\left[\hat{Z}(s)|s\right]\right] \ \ . \tag{4}$$

The first term corresponds to the variance due to the sampling of actions when computing $Z(s)$ for a given state s while the second one is attributed to the sampling of states. Because the variance is nonnegative for every s in the second term, we have $\text{Var}\left[\hat{Z}(s)\right] \geq \text{Var}_s\left[Z(s,\theta)\right]$. This is the core idea behind the so-called *conditional Monte-Carlo* methods [Hammersley and Handscomb, 1964] and the method of *conditioning* as a variance reduction technique [Bratley et al., 1987].

Because the inner conditional expectation in (2) involves the action-value function Q^{π_θ} (unknown in the model-free setting), some algorithmic considerations are needed to implement this idea. When an arbitrarily resettable simulator is available, a potential solution consists in sampling a rollout for every given (s, a) pair. While unbiased, we dismiss this approach due to its high computational cost and lack of generality in domains where offline batch data is available. The solution put forward in this paper consists in using function approximation methods to estimate Q^{π_θ} separately in an actor-critic fashion [Sutton, 1984] and then numerically integrate the resulting approximate quantity.

4 Problem Formulation

In the absence of special structure [Ciosek and Whiteson, 2018], the main challenge in implementing all-action estimators lies in the intractability of computing the inner expectation in (2) for general action spaces. In this paper, we propose two approaches to tackle this problem: numerical quadrature rules and Monte Carlo integration. We aim at deriving efficient estimators of the form:

$$\nabla J(\theta) \simeq \frac{1}{T} \sum_{t=0}^{T-1} \hat{Z}^{\hat{A}_\theta}(s_t) \ \ , \tag{5}$$

for a given trajectory τ and approximate advantage function $\hat{A}_\theta$ [Baird, 1993].

Quadrature Formula: With fixed-grid quadrature methods [Forsythe et al., 1977], N evenly-spaced points are chosen apriori over the range of action space $\{a_1, ..., a_N\}$ and we define:

$$\hat{Z}_{FG}^{\hat{A}_\theta}(s, \theta) = \frac{\epsilon}{N} \sum_{i=1}^{N} \frac{\partial \pi_\theta(a_i|s)}{\partial \theta} \hat{A}_\theta(s, a_i) = \frac{\epsilon}{N} \sum_{i=1}^{N} \hat{z}^{\hat{A}_\theta}(s, a_i, \theta) \ \ , \tag{6}$$

where ϵ is the step size and $\hat{z}^{\hat{A}_\theta}(s, a, \theta) = \nabla_\theta \log \pi_\theta(a|s) \hat{A}_\theta(s, a)$. These methods generalize to higher dimensions but the number of actions required increases exponentially with the dimension of $\mathcal{A}$. We present a policy update subroutine that uses trapezoidal integration in appendix.

Monte Carlo Integration: While numerical quadrature methods are efficient in low dimensional settings, they display exponential sample complexity as the dimension increases. In higher dimensions, Monte Carlo integration methods offer a performance advantage [Evans and Swartz, 2000] while being easier to analyze [Glynn and Iglehart, 1988] and implement. Using the change of measure approach in (2), we derive the all-action Monte Carlo integration estimator from:

$$Z(s, \theta) = \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[z^{A_\theta}(s, a, \theta) \right] \simeq \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[\hat{z}^{\hat{A}_\theta}(s, a, \theta) \right] \ , \tag{7}$$

which we can then approximate from samples using N_S iid samples from $\pi_\theta(\cdot|s)$:

$$\hat{Z}^{\hat{A}_\theta}_{N_S}(s) = \frac{1}{N_S} \sum_{k=1}^{N_S} \nabla_\theta \log \pi_\theta(a_k|s) \hat{A}_\theta(s, a_k) \ . \tag{8}$$

5 Mean Squared Error Analysis

The all-action estimators presented in section 4 are cast under the actor-critic framework [Sutton, 1984] in which an approximation of Q^{π_θ} and the corresponding advantage term are maintained separately. In this section, we investigate how the use of an approximate critic in the all-action estimator affects our ability to estimate the true policy gradient. More specifically, we provide a bound in theorem 1 on the mean squared error of the gradient estimator computed by Monte-Carlo integration, compared to the unbiased LR (REINFORCE) estimator.

Theorem 1. *Assume that for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, $\log \pi_\theta(s, a)$ is a $\sqrt{M}$-Lipschitz function of $\theta \in \mathbb{R}^d$, i.e. $\forall a \in A$, $\forall s \in S$, $||\nabla_\theta \log \pi_\theta\,(a|s)\,||_2^2 \leq M \in \mathbb{R}$. Furthermore, let $L^{MC}_{N_S} = \mathbb{E}\left[||\hat{Z}^{\hat{A}_\theta}_{N_S}(s) - \nabla J(\theta)||^2 \right]$ be the MSE (taken w.r.t. both the state and the sampled actions) of the Monte Carlo integration estimate $\hat{Z}^{\hat{A}_\theta}_{N_S}(s)$ estimator (Eq. 8) to the true policy gradient. We have:*

$$L^{MC}_{N_S} \leq M L_{\hat{A}_\theta} + \left(L + M L_{\hat{A}_\theta} d \right) / N_S \ , \tag{9}$$

where $L_{\hat{A}_\theta} = \mathbb{E}\left[\left(\hat{A}_\theta(s, a) - A_\theta(s, a) \right)^2 \right]$ is the MSE of the advantage estimate and $L = MSE\left[\nabla_\theta \log \pi_\theta(a|s) A_\theta(s, a) \right] \in \mathbb{R}$.

Assuming that the critic error remains small, and with sufficient representation power in the function approximator class, the term $L_{\hat{A}_\theta}$ in theorem 1 can be made relatively small. We then see that most of the variance can then be annihilated by increasing the number of sampled actions N_s.

In order to better understand the effect of using an approximate critic in AAPG, we provide a similar bound in theorem 2 on the MSE for the classical LR (REINFORCE) based on rollouts only.

Theorem 2. *Let the MSE of the REINFORCE estimator to the true policy gradient be:*

$$L_R = MSE_{\pi_\theta} \left[\nabla_\theta \log \pi_\theta(a|s) \left(\sum_{k=0}^{T} \gamma^k r(s_k, a_k, s_{k+1}) - \hat{V}_\theta(s_k) \right) \right] \ .$$

In addition to the regularity assumption of Theorem 1, assume that $\mathbb{E}_{\pi_\theta}\left[\left(G_t - \hat{V}_\theta(s_0) - A_\theta(s_0, a_0) \right)^2 \right] \leq \xi \in \mathbb{R}$. Then L_R satisfies $L_R \leq L + M\xi$.

Remarks The constant ξ in theorem 2 can be understood as the error in the observed rewards, which is a function of the randomness of the reward distribution and that of the policy. Overall, the variance of the REINFORCE estimator is a function of 1) The randomness due to sampling states and actions, even when Q_θ is perfectly known (L); 2) The sensitivity of the policy to variations of θ (M); 3) The stochasticity of the rewards that the policy samples from the environment (ξ). As N_S increases, the estimator $\hat{Z}^{\hat{A}_\theta}_{N_S}$ becomes more efficient at reducing the mean squared error of the policy gradient compared to the *single-action* REINFORCE estimator $\hat{Z}_R$. This property is observed empirically in section 7.2 where the error is shown to decrease almost perfectly as $1/N_s$.

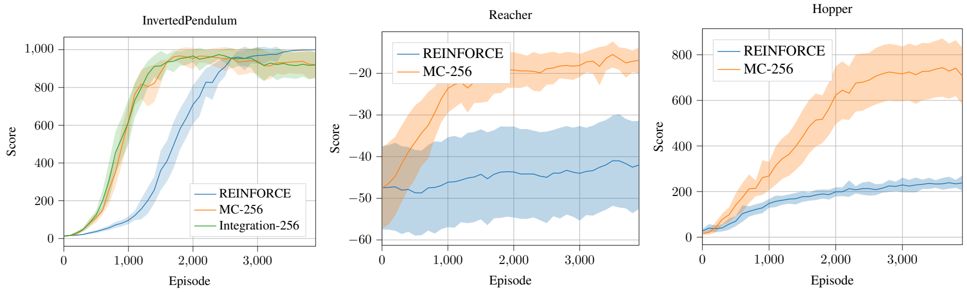

Figure 1: Learning curves for InvertedPendulum-v1 (left), Reacher-v1 (center), Hopper-v1 (right)

6 Convergence of Policy Gradient Methods with a Biased Critic

Policy gradients belong to the class of stochastic gradient methods for which convergence results are often [Bertsekas, 2016] stated only in the unbiased case. Hence, for likelihood ratio methods of the style of REINFORCE, classical convergence results apply readily [Glynn, 1987, L'Ecuyer, 1991]. In order to address the function approximation case, [Sutton et al., 1999] show that if the critic is linear in the so-called *compatible features* then the resulting policy gradient estimator is unbiased. However, the compatible features condition is rather stringent and cannot be satisfied easily beyond the linear case. Because we consider all-action estimators based on approximate critic which can be linear or nonlinear, we need a more general result. Theorem 3 characterizes the expected dynamics of the stochastic ascent procedure arising from policy gradient estimators using an approximate critic.

Theorem 3. *Assume that $\theta \mapsto J(\theta)$ is a μ-Lipschitz smooth concave function ($||\nabla J(x) - \nabla J(y)||_2 \leq \mu||x-y||_2$ for all x, y). If the step size at step k, δ_k, satisfies $\delta_k \leq \frac{1}{\mu}$, then in expectation:*

$$J(\theta_{k+1}) \geq J(\theta_k) + \frac{1}{2}||\nabla J(\theta_k)||^2 + \frac{1}{\mu}\left[\left(\nabla J(\theta_k) - \frac{1}{2}\mathit{Bias}\left(\hat{\nabla} J(\theta_k)\right)\right) \cdot \mathit{Bias}\left(\hat{\nabla} J(\theta_k)\right) - \mathit{Tr}\left(\mathit{Var}\left(\hat{\nabla} J(\theta_k)\right)\right)\right]$$

Where $\hat{\nabla} J(\theta_k)$ is any (potentially biased) estimator of $\nabla J(\theta_k)$ and $\theta_{k+1} = \theta_k + \delta_k \hat{\nabla} J(\theta_k)$.

Remarks This result highlights several central aspects of our methods: 1) Reducing the variance of the gradient estimator is beneficial in terms of sample efficiency; 2) Having a biased gradient estimator can either increase or decrease the convergence speed, depending on the sign of the dot product of the bias and the true gradient; 3) Assuming that the bias of the gradient estimator remains small compared to the true gradient, monotonic improvements can still be guaranteed in expectation.

7 Experimental Results

In this section, we demonstrate the use of the all-action policy gradient (AAPG) algorithms on a variety of MuJoCo tasks [Todorov et al., 2012]. We used a Gaussian parameterization for the policy whose mean is the output of a neural network, and whose covariance matrix is diagonal and fixed (any other distribution could have been used, such as Beta or χ^2). Both the mean and the actions sampled from the policy are clipped to the admissible range of the action space. We learn both $\hat{Q}$ and $\hat{V}$ using neural networks. The $\hat{V}$ function is learned by least-squares regression on the discounted returns while $\hat{Q}$ is learned by expected SARSA [van Seijen et al., 2009]. It is essential to note that our experiment setup makes REINFORCE (with baseline) and our methods fully comparable since the only difference is in how the policy gradients are computed. All hyperparameters are the same for all algorithms on a given task. More implementation details are given in the supplementary material.

7.1 Performance of All-Action Monte Carlo Integration Estimator

In Figure 1, we give learning curves for 3 MuJoCo continuous control tasks with action spaces of diverse dimensions, averaged across 25 random initializations. Error bars use Student 90% confidence

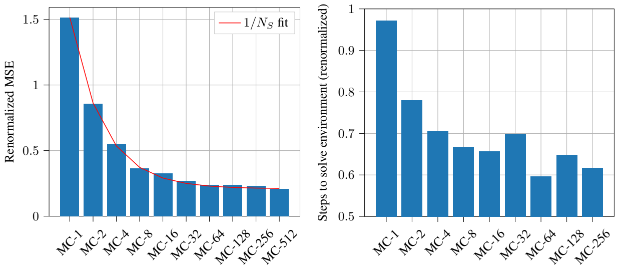

Figure 2: Empirical MSE of $\hat{Z}_{N_S}^{\hat{A}_\theta}$ for $0 \leq \log_2(N_S) \leq 9$ in a given training step on InvertedPendulum-v1 (left) and average number of steps (across 25 random initializations) to solve the environment. The rightmost figure shows that reducing the MSE of the gradient estimator increases sample efficiency.

intervals for the empirical mean at every episode. As an illustration of numerical quadrature methods, we provide runs with trapezoidal integration AAPG for InvertedPendulum-v1, but chose to focus on Monte-Carlo methods in the other experiments since quadrature methods scale poorly with dimension and did not yield interesting improvements over Monte-Carlo even in a simple setting.

7.2 Gradient MSE and sample efficiency

As an empirical verification of Theorem 1, we can use the following approach to compute an approximation of the MSE of the $\hat{Z}_{N_S}^{\hat{A}_\theta}$ estimators for various values of N_S: We first train the policy, critic, and baseline for 1000 episodes on a given environment. Since the REINFORCE estimator is an unbiased estimate of the policy gradient, we use an additional 1000 rollouts to compute an accurate estimate of the true gradient. We then use this proxy gradient to compute the MSE of the $\hat{Z}_{N_S}^{\hat{A}_\theta}$ estimator for various values of N_S by computing the empirical MSE across $1,000$ gradient estimates for each value of N_S. Our results on *InvertedPendulum-v1* are shown in Figure 2, and highlight the benefits of sampling more actions for each state when computing the policy gradient estimate. As N_S increases, the MSE converges (with a clearly visible $1/N_S$ progression) to a positive value, which corresponds to the MSE of the learned advantage estimate (first term of Eq. 9). Figure 2 also draws a parallel between the MSE reduction and the sample efficiency of the algorithm (here defined as the number of steps needed to reach an average score of 950 over a window of 100 consecutive episodes).

8 Conclusion

We have shown, both in theory and practice, that the all-action policy gradient estimator of [Sutton et al., 2001] can yield performance gains and better sample efficiency by acting as a variance reduction technique. We have established that the mechanism at play in this approach is that of *conditioning*. We also provided practical algorithms for implementing all-actions estimators in large or continuous action spaces which are compatible with linear or nonlinear function approximation. We derived novel bounds on the MSE of our Monte-Carlo integration estimator and compared it both theoretically and experimentally to that of the REINFORCE estimator. Our theoretical analysis also provides the first characterization of the dynamics of policy gradient methods with a biased critic.

Our methods could be extended in many ways, notably by combining them with other variance reduction techniques such as control variates. As pointed out by [Glynn and Szechtman, 2002], the conditional expectation term used in the conditioning technique can itself be used as a control variate. In other words, efforts spent computing the integrand by numerical integration can then be re-used to form a complementary control variate. Note finally that given the basis of our Monte Carlo integration all-action estimator in importance sampling, it is natural to think of this approach as a form of off-policy policy gradient method [Degris et al., 2012] where the *off-policiness* is only inside the inner expectation (2) rather than the outer expectation over the stationary distribution [Liu et al., 2019]. Hence, additional variance reductions could be obtained by choosing a better sampling distribution beyond π_θ itself [L'Ecuyer, 1994].

References

Reuven Y. Rubinstein. Sensitivity analysis and performance extrapolation for computer simulation models. *Operations Research*, 37(1):72–81, feb 1989.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 1057–1063, Cambridge, MA, USA, 1999. MIT Press.

Richard S. Sutton, Satinder Singh, and David McAllester. Comparing policy-gradient algorithms. In *unfinished paper*, 2001. URL `http://incompleteideas.net/papers/SSM-unpublished.pdf`.

Paul Bratley, Bennett L. Fox, and Linus E. Schrage. *A Guide to Simulation*. Springer New York, 1987.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992.

Kamil Ciosek and Shimon Whiteson. Expected policy gradients. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

V.M. Aleksandrov, V.I. Sysoyev, and V.V. Shemeneva. Stochastic optimization. *Engineering Cybernetics*, 5(5):11–16, 1968.

Peter W. Glynn. Likelilood ratio gradient estimation: An overview. In *Proceedings of the 19th Conference on Winter Simulation*, WSC 1987, pages 366–375, New York, NY, USA, 1987. ACM. ISBN 0-911801-32-4.

Martin I. Reiman and Alan Weiss. Sensitivity analysis for simulations via likelihood ratios. *Operations Research*, 37(5):830–844, 1989.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning series. MIT Press, 2018. ISBN 9780262039246.

Hajime Kimura, Masayuki Yamamura, and Shigenobu Kobayashi. Reinforcement learning by stochastic hill climbing on discounted reward. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, pages 295–303, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-377-8.

Hajime Kimura, Kazuteru Miyazaki, and Shigenobu Kobayashi. Reinforcement learning in pomdps with function approximation. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 152–160, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. ISBN 1-55860-486-3.

Hajime Kimura and Shigenobu Kobayashi. Reinforcement learning for continuous action using stochastic gradient ascent. In *Intelligent Autonomous Systems (IAS-5)*, pages 288–295, 1998.

Peter Marbach and John N. Tsitsiklis. Simulation-based optimization of markov reward processes. In *Proceedings of the 37th IEEE Conference on Decision and Control (Cat. No.98CH36171)*, volume 3, pages 2698–2703 vol.3, Dec 1998.

Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1008–1014. MIT Press, 2000.

Jonathan Baxter and Peter L. Bartlett. Infinite-horizon policy-gradient estimation. *J. Artif. Int. Res.*, 15(1):319–350, November 2001.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3528–3536. Curran Associates, Inc., 2015.

Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search provides a competitive approach to reinforcement learning. *CoRR*, abs/1803.07055, 2018.

Pierre L'Ecuyer. An overview of derivative estimation. In *1991 Winter Simulation Conference Proceedings.*, pages 207–217, Dec 1991.

Pierre L'Ecuyer. A unified view of the ipa, sf and lr gradient estimation techniques. *Management Science*, 36(11):1364–1383, nov 1990.

Pierre L'Ecuyer. Efficiency improvement and variance reduction. In *Proceedings of the 26th Conference on Winter Simulation*, WSC '94, pages 122–132, San Diego, CA, USA, 1994. Society for Computer Simulation International.

Peter W. Glynn and Pierre L'Ecuyer. Likelihood ratio gradient estimation for stochastic recursions. *Advances in Applied Probability*, 27(4):1019–1053, December 1995.

J. M. Hammersley and D. C. Handscomb. Conditional monte carlo. In *Monte Carlo Methods*, pages 76–84. Springer Netherlands, 1964.

Harm van Seijen, Hado van Hasselt, Shimon Whiteson, and Marco Wiering. A theoretical and empirical analysis of expected sarsa. In *2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*. IEEE, March 2009.

Cameron Allen, Kavosh Asadi, Melrose Roderick, Abdel-rahman Mohamed, George Konidaris, and Michael Littman. Mean Actor Critic. *arXiv e-prints*, abs/1709.00503, Sep 2017.

Matthew Fellows, Kamil Ciosek, and Shimon Whiteson. Fourier policy gradients. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1485–1494, 2018.

George E. Forsythe, Michael A. Malcolm, and Cleve B. Moler. *Computer Methods for Mathematical Computations*. Prentice Hall Professional Technical Reference, 1977. ISBN 0131653326.

E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, Oct 2012.

D.P. Bertsekas. *Nonlinear Programming*. Athena scientific optimization and computation series. Athena Scientific, 2016. ISBN 9781886529052.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994. ISBN 0471619779.

Csaba Szepesvari. *Algorithms for Reinforcement Learning*. Morgan and Claypool Publishers, 2010. ISBN 1608454924, 9781608454921.

Richard Stuart Sutton. *Temporal Credit Assignment in Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 1984.

Leemon C. Baird. Advantage updating. Technical Report WL–TR-93-1146, Wright Laboratory, 1993.

Michael Evans and Timothy Swartz. *Approximating integrals via Monte Carlo and deterministic methods*, volume 20. OUP Oxford, 2000.

Peter W. Glynn and Donald L. Iglehart. Simulation methods for queues: An overview. *Queueing Systems*, 3(3):221–255, Sep 1988. ISSN 1572-9443.

Peter W. Glynn and Roberto Szechtman. Some new perspectives on the method of control variates. In Kai-Tai Fang, Harald Niederreiter, and Fred J. Hickernell, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2000*, pages 27–49, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-642-56046-0.

Thomas Degris, Martha White, and Richard S. Sutton. Off-policy actor-critic. In *International Conference on Machine Learning (ICML)*, 2012.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. In *Uncertainty in Artificial Intelligence (UAI)*, 2019.