From Importance Sampling to Doubly Robust Policy Gradient

Jiawei Huang [1] Nan Jiang [1]

Abstract

We show that on-policy policy gradient (PG) and its variance reduction variants can be derived by taking finite difference of function evaluations supplied by estimators from the importance sampling (IS) family for off-policy evaluation (OPE). Starting from the doubly robust (DR) estimator (Jiang & Li, 2016), we provide a simple derivation of a very general and flexible form of PG, which subsumes the state-of-the-art variance reduction technique (Cheng et al., 2019) as its special case and immediately hints at further variance reduction opportunities overlooked by existing literature. We analyze the variance of the new DR-PG estimator, compare it to existing methods as well as the Cramer-Rao lower bound of policy gradient, and empirically show its effectiveness.

1. Introduction

In reinforcement learning, policy gradient (PG) refers to the family of algorithms that estimate the gradient of the expected return w.r.t. the policy parameters, often from on-policy Monte-Carlo trajectories. Off-policy evaluation (OPE) refers to the problem of evaluating a policy that is different from the data generating policy, often by *importance sampling* (IS) techniques.

Despite the superficial difference that standard PG is on-policy while IS for OPE is off-policy by definition, they share many similarities: both PG and IS are arguably based on the Monte-Carlo principle (as opposed to the dynamic programming principle); both of them often suffer from high variance, and variance reduction techniques have been studied extensively for PG and IS separately in the literature. Given these similarities, one may naturally wonder: *is there a deeper connection between the two topics?*

Summary of the Paper We provide a simple and positive answer to the above question in the episodic RL setting. In particular, one can write down the policy gradient as (we informally illustrate the idea with scalar θ for now)

$$\lim_{\Delta\theta \to 0} \frac{J(\pi_{\theta+\Delta\theta}) - J(\pi_\theta)}{\Delta\theta}, \tag{1}$$

where θ is the current policy parameter, and $J(\cdot)$ is the expected return of a policy w.r.t. some initial state distribution. The connection between IS and PG is extremely simple: using any method in the IS family to estimate $J(\cdot)$ in Eq.(1) will lead to a version of PG, and most unbiased PG estimators—with different variance reduction techniques—can be recovered in this way. Furthermore, by deriving PG from the doubly robust (DR) estimator for OPE (Jiang & Li, 2016), we obtain a very general and flexible form of PG with variance reduction, which immediately subsumes the state-of-the-art technique by Cheng et al. (2019) as its special case. In fact, the resulting estimator can achieve more variance reduction than Cheng et al. (2019) given additional side information. See Table 1 for some highlighted results.

2. Related Work

To the best of our knowledge, Tang & Abbeel (2010) was the first to explicitly mention the connection between (per-trajectory) IS and PG, which corresponds to the first row of our Table 1. The connection between DR and PG was lightly touched by Tucker et al. (2018), although the authors' main goal was to challenge the success of state-action-dependent baseline methods in benchmarks, and did not give a more detailed analysis on this connection.

More recently, Cheng et al. (2019) noticed that the previous variance reduction methods in PG overlooked the correlation across the times steps and ignored the randomness in the future steps (e.g., Grathwohl et al., 2018; Gu et al., 2017; Liu et al., 2018; Wu et al., 2018). They used the law of the total variance to derive a trajectory-wise control variate estimator, which is subsumed by our general form of PG derived from DR in Section 5 as a special case.

[1] Department of Computer Science, University of Illinois Urbana-Champaign. Correspondence to: Nan Jiang <nanjiang@illinois.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

Table 1. OPE estimators and their corresponding PG estimators. Time index in the subscript often specifies the omitted function arguments, e.g., $\widetilde{V}_t^{\pi'} := \widetilde{V}^{\pi'}(s_t)$; see Section 3.4 for details. Also note that $b \equiv V^\pi$ considered in the variance column is for simplicity and is not the optimal baseline (Tang & Abbeel, 2010).

	OPE	Finite Diff ⇒	PG	(Co)Variance of PG in Deterministic MDPs, with $b \equiv V^\pi$ and $\widetilde{Q}^{(\cdot)} \equiv Q^{(\cdot)}$
Traj-IS	$\rho_{[0:T]} \sum_{t=0}^{T} \gamma^t r_t$	Proposition 8 (Tang & Abbeel)	$\sum_{t=0}^{T} \nabla \log \pi_{\boldsymbol{\theta}}^t \sum_{t'=0}^{T} \gamma^{t'} r_{t'}$	Omitted (worse than step-IS)
Step-IS	$\sum_{t=0}^{T} \gamma^t \rho_{[0:t]} r_t$	Proposition 1	$\sum_{t=0}^{T} \nabla \log \pi_{\boldsymbol{\theta}}^t \sum_{t'=t}^{T} \gamma^{t'} r_{t'}$	$\mathbb{E}[\sum_{t=0}^{T} \gamma^{2t} \mathrm{Cov}_t[\nabla Q_t^{\pi_{\boldsymbol{\theta}}} + Q_t^{\pi_{\boldsymbol{\theta}}} \sum_{t'=0}^{t} \nabla \log \pi_{\boldsymbol{\theta}}^{t'} \| s_t]]$
Baseline	$b_0 + \sum_{t=0}^{T} \gamma^t \rho_{[0:t]} \Big(r_t + \gamma b_{t+1} - b_t \Big)$	Proposition 2	$\sum_{t=0}^{T} \nabla \log \pi_{\boldsymbol{\theta}}^t \Big(\sum_{t'=t}^{T} \gamma^{t'} r_{t'} - \gamma^t b_t \Big)$	$\mathbb{E}[\sum_{t=0}^{T} \gamma^{2t} \mathrm{Cov}_t[\nabla Q_t^{\pi_{\boldsymbol{\theta}}} + A_t^{\pi_{\boldsymbol{\theta}}} \sum_{t'=0}^{t} \nabla \log \pi_{\boldsymbol{\theta}}^{t'} \| s_t]]$
Doubly Robust	Recursive Version (Jiang & Li, 2016) $\widehat{\mathrm{DR}}_t^{\pi'} := \widetilde{V}_t^{\pi'}$ $+\rho_t \Big(r_t + \gamma \widehat{\mathrm{DR}}_{t+1}^{\pi'} - \widetilde{Q}_t^{\pi'} \Big)$ Expanded Version $\widetilde{V}_0^{\pi'} + \sum_{t=0}^{T} \gamma^t \rho_{[0:t]} \Big(r_t + \gamma \widetilde{V}_{t+1}^{\pi'} - \widetilde{Q}_t^{\pi'} \Big)$	Theorem 3	$\widetilde{Q}$ does not change with $\boldsymbol{\theta}$ (Cheng et al., 2019) $\sum_{t=0}^{T} \Big\{ \nabla \log \pi_{\boldsymbol{\theta}}^t \Big[\sum_{t'=t}^{T} \gamma^{t'} r_{t'}$ $+ \sum_{t'=t+1}^{T} \gamma^{t'} \Big(\widetilde{V}_{t'}^{\pi_{\boldsymbol{\theta}}} - \widetilde{Q}_{t'}^{\pi_{\boldsymbol{\theta}}} \Big) \Big]$ $+\gamma^t \Big(\nabla \widetilde{V}_t^{\pi_{\boldsymbol{\theta}}} - \widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}} \nabla \log \pi_{\boldsymbol{\theta}}^t \Big) \Big\}$	$\mathbb{E}[\sum_{t=0}^{T} \gamma^{2t} \mathrm{Cov}_t[\nabla Q_t^{\pi_{\boldsymbol{\theta}}} \| s_t]]$
			$\widetilde{Q}$ is a function of $\boldsymbol{\theta}$ (new) $\sum_{t=0}^{T} \Big\{ \nabla \log \pi_{\boldsymbol{\theta}}^t \Big[\sum_{t'=t}^{T} \gamma^{t'} r_{t'}$ $+ \sum_{t'=t+1}^{T} \gamma^{t'} \Big(\widetilde{V}_{t'}^{\pi_{\boldsymbol{\theta}}} - \widetilde{Q}_{t'}^{\pi_{\boldsymbol{\theta}}} \Big) \Big]$ $+\gamma^t \Big(\nabla \widetilde{V}_t^{\pi_{\boldsymbol{\theta}}} - \nabla \widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}$ $-\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}} \nabla \log \pi_{\boldsymbol{\theta}}^t \Big) \Big\}$	$\mathbf{0}$ (zero matrix)
Actor Critic	$\sum_{t=0}^{T} \gamma^t \rho_{[0:t]} \Big(f_t - \gamma f_{t+1} \Big)$	Proposition 9	$\sum_{t=0}^{T} \gamma^t \nabla \log \pi_{\boldsymbol{\theta}}^t \cdot f_t$	Omitted (biased estimator)

3. Preliminaries

3.1. Markov Decision Processes (MDPs)

We consider episodic RL problems with a fixed horizon, formulated as an MDP $M = (\mathcal{S}, \mathcal{A}, P, R, T, \gamma, s_0)$, where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the action space. For the ease of exposition we assume both $\mathcal{S}$ and $\mathcal{A}$ are finite and discrete.[1] $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, $R : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ is the reward function, and T is the horizon (or episode length). It is optional but we also include a discount factor $\gamma \in [0, 1)$ for more flexibility, which will later allow us to express the estimators in the IS and the PG literature in consistent notations. s_0 is the deterministic start state, which is without loss of generality. We will also assume that state contains the time step information (so that value functions are stationary); in other words, each state can only appear at a particular time step. Overall, these assumptions are only made for notational simplicity, and do not limit the generality of our derivations.

A (stochastic) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ induces a random trajectory $s_0, a_0, r_0, s_1, a_2, r_2, s_3, \ldots, s_T, a_T, r_T$, where $a_t \sim \pi(s_t)$, $r_t \sim R(s_t, a_t)$, and $s_{t+1} \sim P(s_t, a_t)$ for all $0 \le t \le T$. The ultimate measure of the performance of π is the expected return, defined as

$$J(\pi) := \mathbb{E}\left[\textstyle\sum_{t=0}^{T} \gamma^t r_t \,|\, a_{0:T} \sim \pi\right],$$

where $a_{0:T}$ is the shorthand for $a_t \sim \pi(s_t)$ for $0 \le t \le T$. It will be useful to define the state-value and Q-value functions: for s that may appear in time step t (recall that we assume t is encoded in s),

$$V^{\pi}(s) := \mathbb{E}\left[\sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'} \,|\, s_t = s, a_{t:T} \sim \pi\right],$$
$$Q^{\pi}(s,a) := \mathbb{E}\left[\sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'} \,|\, s_t = s, a_t = a, a_{t+1:T} \sim \pi\right].$$

For simplicity we treat s_{T+1} as a special terminal (absorbing) state, such that any (approximate or estimated) value function always evaluates to 0 on s_{T+1}.

3.2. Off-Policy Evaluation and Importance Sampling

Off-policy evaluation (OPE) is the problem of estimating the expected return of a policy π' from data collected using a different policy π. Importance sampling (IS) is a standard technique for OPE. Given a trajectory $s_0, a_0, r_0, s_1, a_2, r_2, s_3, \ldots, s_T, a_T, r_T$ where all actions are taken according to π, (step-wise) IS forms the following unbiased estimate of $J(\pi')$ (Precup, 2000):

$$\widehat{J}(\pi') = \sum_{t=0}^{T} \gamma^t \prod_{t'=0}^{t} \frac{\pi'(a_{t'}|s_{t'})}{\pi(a_{t'}|s_{t'})} r_t. \tag{2}$$

The estimator for a dataset of multiple trajectories will be simply the average of the above estimator applied to each trajectory. Since such a pattern is found in all estimators we consider (including the PG estimators), we will always consider only a single trajectory in the analyses.

The term $\frac{\pi'(a_t|s_t)}{\pi(a_t|s_t)}$ is often called the importance weight/ratio. We will use ρ_t as its shorthand, and $\rho_{[t_1:t_2]}$ is the shorthand for its cumulative product, $\prod_{t'=t_1}^{t_2} \rho_{t'}$, with $\rho_{[t_1:t_2]} := 1$ when $t_1 > t_2$. With the above shorthand, the step-wise IS estimator can be succinctly expressed as

$$\widehat{J}(\pi') = \sum_{t=0}^{T} \gamma^t \rho_{[0:t]} r_t. \tag{3}$$

We will be referring to multiple OPE estimators throughout the paper. Instead of giving each estimator a separate variable name, we will just use a generic notation $\widehat{J}(\cdot)$, and the specific estimator it refers to should be clear from the surrounding text and theorem statements.

Doubly Robust (DR) Estimator (Jiang & Li, 2016; Thomas & Brunskill, 2016)
The DR estimator uses an approximate value function $\widetilde{Q}^{\pi'}$ to reduce the variance of IS via control variates. In its expanded form, the estimator is

$$\widehat{J}(\pi') = \widetilde{V}^{\pi'}(s_0) + \sum_{t=0}^{T} \gamma^t \rho_{[0:t]} \Big(r_t + \gamma \widetilde{V}^{\pi'}(s_{t+1}) - \widetilde{Q}^{\pi'}(s_t, a_t) \Big), \tag{4}$$

where $\widetilde{V}^{\pi'}(s) := \mathbb{E}_{a \sim \pi'(s)}[\widetilde{Q}^{\pi'}(s,a)]$. Jiang & Li (2016) showed that DR has maximally reduced variance, in the sense that when $\widetilde{Q}^{\pi'}$ is accurate, there exists RL problems (typically tree-MDPs) where the variance of the estimator is equal to the Cramer-Rao lower bound of the estimation problem. As we will see later in Section 5, the PG estimator induced by DR also achieves the state-of-the-art variance reduction, and the variance when both $\widetilde{Q}$ and $\nabla_{\boldsymbol{\theta}} \widetilde{Q}$ are accurate also coincides with the C-R bound for PG.

3.3. Policy Gradient

Consider the problem of finding a good policy over a parameterized class, $\{\pi_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. Each policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \Delta(\mathcal{A})$ is stochastic and we assume that $\pi_{\boldsymbol{\theta}}(a|s)$ is differentiable w.r.t. $\boldsymbol{\theta}$. Policy gradient algorithms (Williams, 1992) perform (stochastic) gradient descent on the objective $J(\pi_{\boldsymbol{\theta}})$,

[1] Note that both PG and IS occur no explicit dependence on $|\mathcal{S}|$ or $|\mathcal{A}|$, and estimators derived for the discrete case can be extended to continuous state and action spaces.

and the following expression is an unbiased gradient based on a single trajectory (Sutton et al., 2000):

$$\sum_{t=0}^{T}\left(\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a_t|s_t)\sum_{t'=t}^{T}\gamma^{t'}r_{t'}\right). \tag{5}$$

Note that although most PG results are derived for the infinite-horizon discounted case, they can be immediately applied to our setup, since our formulation in Section 3.1 can be turned into an infinite-horizon discounted MDP by treating s_{T+1} as an absorbing state.

3.4. Further Notations

Since we always consider the estimators based on a single on-policy trajectory, all expectations $\mathbb{E}[\cdot]$ are w.r.t. that on-policy distribution induced by π (for OPE) or $\pi_{\boldsymbol{\theta}}$ (for PG). Following the notations in Jiang & Li (2016), we use $\mathbb{E}_t[\cdot]$ as a shorthand for the conditional expectation $\mathbb{E}[\cdot|s_0,a_0,\ldots,s_{t-1},a_{t-1}]$, and similarly $\mathbb{V}_t[\cdot]$ and $\mathrm{Cov}_t[\cdot]$ for the conditional (co)variance. We will often see the usage $\mathbb{E}_t[\cdot|s_t]$, which simply means $\mathbb{E}[\cdot|s_0,a_0,\ldots,s_{t-1},a_{t-1},s_t]$.

Omitted function arguments Since all value-functions of the form V^π (or Q^π) are always applied on s_t (or s_t,a_t) in the trajectory, we will sometimes omit such arguments and use V_t^π as a shorthand for $V^\pi(s_t)$ (and Q_t^π for $Q^\pi(s_t,a_t)$). Similarly, we write π_t as a shorthand for $\pi(a_t|s_t)$, and $\pi_{\boldsymbol{\theta}}^t$ as a shorthand for $\pi_{\boldsymbol{\theta}}(a_t|s_t)$.

4. Warm-up: Deriving PG from IS

In this section we show how the most common forms of PG can be derived from the corresponding IS estimators. Although these results will be later subsumed by our main theorem in Section 5, it is still instructive to derive the connection between IS and PG from the simpler cases.

Vanilla PG

Proposition 1. *The standard PG (Eq.(5)) can be derived from taking finite difference over step-wise IS (Eq.(3)).*

Proof. Denote $\boldsymbol{e_i}$ as the i-th standard basis vectors in $\boldsymbol{\theta}=(\theta_1,\theta_2,...,\theta_d)\in\mathbb{R}^d$ space, and denote ε_i as a small scalar for $i=1,2,...,d$. Then, we apply step-wise IS on the policy $\pi':=\pi_{\boldsymbol{\theta}+\varepsilon_i\boldsymbol{e}_i}$ for arbitrary $i=1,2,...,d$:

$$\begin{aligned}\widehat{J}(\boldsymbol{\theta}+\varepsilon_i\boldsymbol{e}_i) &= \sum_{t=0}^{T}\rho_{[0:t]}\gamma^t r_t = \sum_{t=0}^{T}\gamma^t r_t\prod_{t'=0}^{t}\frac{\pi^{t'}_{\boldsymbol{\theta}+\varepsilon_i\boldsymbol{e}_i}}{\pi^{t'}_{\boldsymbol{\theta}}}\\ &= \sum_{t=0}^{T}\gamma^t r_t\prod_{t'=0}^{t}\left(1+\frac{\pi^{t'}_{\boldsymbol{\theta}+\varepsilon_i\boldsymbol{e}_i}-\pi^{t'}_{\boldsymbol{\theta}}}{\pi^{t'}_{\boldsymbol{\theta}}}\right)\\ &= \sum_{t=0}^{T}\gamma^t r_t\left(1+\sum_{t'=0}^{t}\frac{\Delta\pi^{t'}_{\theta_i}}{\pi^{t'}_{\boldsymbol{\theta}}}\right)+o(\varepsilon_i)\\ &= \sum_{t=0}^{T}\gamma^t r_t+\sum_{t=0}^{T}\frac{\Delta\pi^{t}_{\theta_i}}{\pi^{t}_{\boldsymbol{\theta}}}\sum_{t'=t}^{T}\gamma^{t'}r_{t'}+o(\varepsilon_i).\end{aligned}$$

where $\Delta\pi^t_{\theta_i}$ is a shorthand of $\langle\nabla_{\boldsymbol{\theta}}\pi^t_{\boldsymbol{\theta}},\varepsilon_i\boldsymbol{e}_i\rangle$. Then,

$$\begin{aligned}\frac{\partial\widehat{J}(\boldsymbol{\theta})}{\partial\theta_i} &= \lim_{\varepsilon_i\to 0}\frac{\widehat{J}(\boldsymbol{\theta}+\varepsilon_i\boldsymbol{e}_i)-\widehat{J}(\boldsymbol{\theta})}{\varepsilon_i}\\ &= \lim_{\varepsilon_i\to 0}\sum_{t=0}^{T}\frac{\Delta\pi^t_{\theta_i}/\pi^t_{\boldsymbol{\theta}}}{\varepsilon_i}\sum_{t'=t}^{T}\gamma^{t'}r_{t'}\\ &= \sum_{t=0}^{T}\frac{\partial\log\pi^t_{\boldsymbol{\theta}}}{\partial\theta_i}\sum_{t'=t}^{T}\gamma^{t'}r_{t'}.\end{aligned}$$

As a result, the estimator derived from Eq.(3) should be:

$$\begin{aligned}&\left(\frac{\partial\widehat{J}(\boldsymbol{\theta})}{\partial\theta_1},\frac{\partial\widehat{J}(\boldsymbol{\theta})}{\partial\theta_2},...,\frac{\partial\widehat{J}(\boldsymbol{\theta})}{\partial\theta_d}\right)^\top\\ &=\sum_{t=0}^{T}\left(\nabla_{\boldsymbol{\theta}}\log\pi^t_{\boldsymbol{\theta}}\sum_{t'=t}^{T}\gamma^{t'}r_{t'}\right).\end{aligned}$$

□

PG with a Baseline Using a state baseline is a simple and popular form of variance reduction for PG. Below we show that there exists an unbiased OPE estimator (Eq.(7)) that yields such a PG estimator via the procedure in Eq.(1).

Proposition 2. *PG with a baseline (Greensmith et al., 2004)*

$$\sum_{t=0}^{T}\left(\nabla_{\boldsymbol{\theta}}\log\pi^t_{\boldsymbol{\theta}}\left(\sum_{t'=t}^{T}\gamma^{t'}r_{t'}-\gamma^t b(s_t)\right)\right) \tag{6}$$

can be derived by taking finite difference over the following OPE estimator,

$$\widehat{J}(\pi')=\sum_{t=0}^{T}\gamma^t\rho_{[0:t]}\Big(r_t-b(s_t)+\gamma b(s_{t+1})\Big)+b(s_0). \tag{7}$$

(Recall that $b(s_{T+1})=0$.) Furthermore, Eq.(7) is an unbiased estimator of $J(\pi')$.

We defer the proof to Appendix A, which also includes the connections between a few other pairs of OPE and PG estimators presented in Table 1.

Remark Whenever we encounter unfamiliar OPE or PG estimators in the derivation, we always verify their unbiasedness from scratch. (We omit such verification for those well-known estimators.) However, a PG estimator that is derived from a known and unbiased OPE estimator should be automatically unbiased, thanks to the linearity (and hence exchangeability) of differentiation and expectation; see further details in Appendix B.

5. A General Form of PG Derived From DR

In the previous section we have derived some popular forms of PG from their IS counterparts. However, as Cheng et al. (2019) noticed, the variance reduction in popular PG algorithms are relatively naïve. From our perspective, this is evidenced by the fact that the IS counterparts of these popular PG estimators—which often uses a "baseline" that carries the semantics of state-value functions—are naïve OPE estimators and do not fully exploit variance reduction opportunities in IS.

In this section, we derive a very general form of PG from the unbiased estimator in the IS family that arguably performs the maximal amount of variance reduction, known as the doubly robust estimator (Jiang & Li, 2016; Thomas & Brunskill, 2016), which requires an approximate Q-value function of $\pi_{\boldsymbol{\theta}}$, $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$. We show that a special case of the resulting PG estimator is exactly equivalent to that derived by Cheng et al. (2019) recently from a control variate perspective. This special case treats $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ as not varying with $\boldsymbol{\theta}$, whereas our more general estimator can further leverage the gradient information $\nabla\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ to reduce even more variance. Furthermore, the two popular forms of PG examined in Section 4 are also subsumed as the special cases of our estimator.

5.1. Derivation of DR-PG

Theorem 3. *Let* $\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}}(s) := \widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}(s,\pi_{\boldsymbol{\theta}}(s)) = \mathbb{E}_{a\sim\pi_{\boldsymbol{\theta}}(s)}[\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}(s,a)]$. *The following estimator is an unbiased policy gradient that can be derived by taking finite difference over the doubly robust estimator for OPE:*

$$\sum_{t=0}^{T}\Big\{\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big[\sum_{t_1=t}^{T}\gamma^{t_1}r_{t_1}+\sum_{t_2=t+1}^{T}\gamma^{t_2}\Big(\widetilde{V}_{t_2}^{\pi_{\boldsymbol{\theta}}}-\widetilde{Q}_{t_2}^{\pi_{\boldsymbol{\theta}}}\Big)\Big] \\ +\gamma^t\Big(\nabla_{\boldsymbol{\theta}}\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}}-\nabla_{\boldsymbol{\theta}}\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}-\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big)\Big\}. \tag{8}$$

Remark 4. *Since* $\nabla_{\boldsymbol{\theta}}\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}}(s) = \nabla_{\boldsymbol{\theta}}\ \mathbb{E}_{a\sim\pi_{\boldsymbol{\theta}}(s)}[\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}(s,a)]$, *the dependencies on* $\pi_{\boldsymbol{\theta}}$ *in the subscript and the superscript will both contribute to the gradient calculation.*[2]

Remark 5. *While* $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ *and* $\nabla_{\boldsymbol{\theta}}\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ *look related in notation, they have independent degrees of freedoms and can be estimated using separate procedures; see Appendix C for further details.*

Proof. We first show how Eq.(8) can be derived from DR. We start from the recursive form of DR (Jiang & Li, 2016):

$$\widehat{\mathrm{DR}}_t^{\pi'} = \widetilde{V}_t^{\pi'} + \frac{\pi'_t}{\pi_t}\Big(r_t + \gamma\widehat{\mathrm{DR}}_{t+1}^{\pi'} - \widetilde{Q}_t^{\pi'}\Big). \tag{9}$$

where $\widetilde{Q}$ and $\widetilde{V}$ are the approximate value functions, and $\widetilde{V}_t^{\pi} = \sum_a \pi(a|s_t)\widetilde{Q}_t^{\pi}(s_t,a)$. Note that $\widehat{\mathrm{DR}}_0^{\pi'}$ is equivalent to the expanded form given in Eq.(4) (Thomas & Brunskill, 2016). Denote e_i as the i-th standard basis vectors in $\boldsymbol{\theta}\in\mathbb{R}^d$ space, and denote ε_i as a scalar. Let $\Delta\pi_{\theta_i}^t$ be the shorthand of $\langle\nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}^t, \varepsilon_i\boldsymbol{e}_i\rangle$. Then, we apply DR on the policy $\pi' := \pi_{\boldsymbol{\theta}+\varepsilon_i\boldsymbol{e}_i}$ for $i = 1, 2, ..., d$:

$$\begin{aligned}&\widehat{\mathrm{DR}}_t^{\pi_{\boldsymbol{\theta}+\varepsilon_i\boldsymbol{e}_i}} - \widehat{\mathrm{DR}}_t^{\pi_{\boldsymbol{\theta}}} \\ =&\widetilde{V}_t^{\pi_{\boldsymbol{\theta}+\varepsilon_i\boldsymbol{e}_i}} - \widetilde{V}_t^{\pi_{\boldsymbol{\theta}}} + \frac{\Delta\pi_{\theta_i}^t}{\pi_{\boldsymbol{\theta}}^t}\Big(r_t+\gamma\widehat{\mathrm{DR}}_{t+1}^{\pi_{\boldsymbol{\theta}}}-\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}\Big) \\ &+\Big(\gamma\widehat{\mathrm{DR}}_{t+1}^{\pi_{\boldsymbol{\theta}+\varepsilon_i\boldsymbol{e}_i}} - \gamma\widehat{\mathrm{DR}}_{t+1}^{\pi_{\boldsymbol{\theta}}} - \widetilde{Q}_t^{\pi_{\boldsymbol{\theta}+\varepsilon_i\boldsymbol{e}_i}} + \widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}\Big) + o(\varepsilon_i).\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\partial\widehat{\mathrm{DR}}_t^{\pi_{\boldsymbol{\theta}}}}{\partial\theta_i} &= \lim_{\varepsilon_i\to 0}\frac{\widehat{\mathrm{DR}}_t^{\pi_{\boldsymbol{\theta}+\varepsilon_i\boldsymbol{e}_i}} - \widehat{\mathrm{DR}}_t^{\pi_{\boldsymbol{\theta}}}}{\varepsilon_i} \\ &=\frac{\partial\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}}}{\partial\theta_i} + \frac{\partial\log\pi_{\boldsymbol{\theta}}^t}{\partial\theta_i}(r_t+\gamma\widehat{\mathrm{DR}}_{t+1}^{\pi_{\boldsymbol{\theta}}}-\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}) \\ &\quad+\gamma\frac{\partial\widehat{\mathrm{DR}}_{t+1}^{\pi_{\boldsymbol{\theta}}}}{\partial\theta_i} - \frac{\partial\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}}{\partial\theta_i}.\end{aligned}$$

As a result,

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\widehat{\mathrm{DR}}_t^{\pi_{\boldsymbol{\theta}}} =&\Big(\frac{\partial\widehat{\mathrm{DR}}_t^{\pi_{\boldsymbol{\theta}}}}{\partial\theta_1}, \frac{\partial\widehat{\mathrm{DR}}_t^{\pi_{\boldsymbol{\theta}}}}{\partial\theta_2}, ..., \frac{\partial\widehat{\mathrm{DR}}_t^{\pi_{\boldsymbol{\theta}}}}{\partial\theta_d}\Big)^{\top} \\ =&\nabla_{\boldsymbol{\theta}}\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}} + \nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big(r_t+\gamma\widehat{\mathrm{DR}}_{t+1}^{\pi_{\boldsymbol{\theta}}}-\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}\Big) \\ &+\gamma\nabla_{\boldsymbol{\theta}}\widehat{\mathrm{DR}}_{t+1}^{\pi_{\boldsymbol{\theta}}} - \nabla_{\boldsymbol{\theta}}\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}.\end{aligned} \tag{10}$$

We can continue to expand (10) and finally get the following estimator:

$$\sum_{t=0}^{T}\Big\{\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big[\sum_{t_1=t}^{T}\gamma^{t_1}r_{t_1}+\sum_{t_2=t+1}^{T}\gamma^{t_2}\Big(\widetilde{V}_{t_2}^{\pi_{\boldsymbol{\theta}}}-\widetilde{Q}_{t_2}^{\pi_{\boldsymbol{\theta}}}\Big)\Big] \\ +\gamma^t\Big(\nabla_{\boldsymbol{\theta}}\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}}-\nabla_{\boldsymbol{\theta}}\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}-\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big)\Big\}.$$

Next, we show that the estimator is unbiased.

$$\begin{aligned}&\mathbb{E}\Big[\sum_{t=0}^{T}\Big\{\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big[\sum_{t_1=t}^{T}\gamma^{t_1}r_{t_1}+\sum_{t_2=t+1}^{T}\gamma^{t_2}\Big(\widetilde{V}_{t_2}^{\pi_{\boldsymbol{\theta}}}-\widetilde{Q}_{t_2}^{\pi_{\boldsymbol{\theta}}}\Big)\Big] \\ &\quad+\gamma^t\Big(\nabla_{\boldsymbol{\theta}}\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}}-\nabla_{\boldsymbol{\theta}}\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}-\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big)\Big\}\Big] \\ =&\underbrace{\mathbb{E}\Big[\sum_{t=0}^{T}\Big(\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big(\sum_{t_1=t}^{T}\gamma^{t_1}r_{t_1}\Big)\Big)\Big]}_{p_1} \\ &+\mathbb{E}\Big[\sum_{t=0}^{T}\Big(\underbrace{\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big[\sum_{t_2=t+1}^{T}\gamma^{t_2}\Big(\widetilde{V}_{t_2}^{\pi_{\boldsymbol{\theta}}}-\widetilde{Q}_{t_2}^{\pi_{\boldsymbol{\theta}}}\Big)\Big]}_{p_2^t}\Big)\Big]\end{aligned}$$

[2] In contrast, the $\nabla_{\boldsymbol{\theta}}\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}}$ term in the estimator of Cheng et al. (2019) only differentiate w.r.t. the subscript ($\mathbb{E}_{a\sim\pi_{\boldsymbol{\theta}}(s)}$), as they treat $\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}$ as not varying with $\pi_{\boldsymbol{\theta}}$.

$$+\mathbb{E}\Big[\sum_{t=0}^{T}\gamma^t\underbrace{\Big(\nabla_{\boldsymbol{\theta}}\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}} - \frac{\nabla_{\boldsymbol{\theta}}[\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}\pi_{\boldsymbol{\theta}}^t]}{\pi_{\boldsymbol{\theta}}^t}\Big)}_{p_3^t}\Big].$$

Since p_1 is the usual PG estimator, it suffices to show that p_2^t and p_3^t are equal to 0 in expectation. For p_2^t,

$$\mathbb{E}_t[p_2^t] = \mathbb{E}_t\Big[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big(\sum_{t_2=t+1}^{T}\gamma^{t_2}\mathbb{E}_{t_2}\Big[\widetilde{V}_{t_2}^{\pi_{\boldsymbol{\theta}}} - \widetilde{Q}_{t_2}^{\pi_{\boldsymbol{\theta}}}\Big|s_{t_2}\Big]\Big)\Big] = 0,$$

where $\mathbb{E}_{t_2}\Big[\widetilde{V}_{t_2}^{\pi_{\boldsymbol{\theta}}} - \widetilde{Q}_{t_2}^{\pi_{\boldsymbol{\theta}}}\Big|s_{t_2}\Big] = 0$ because PG is on-policy ($a_{t_2} \sim \pi_{\boldsymbol{\theta}}(s_{t_2})$). Similarly,

$$\begin{aligned}\mathbb{E}_t[p_3^t] &= \sum_a \pi_{\boldsymbol{\theta}}(a|s_t)\nabla_{\boldsymbol{\theta}}\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}} - \sum_a \nabla_{\boldsymbol{\theta}}[\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}\pi_{\boldsymbol{\theta}}^t] \\ &= \nabla_{\boldsymbol{\theta}}\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}} - \nabla_{\boldsymbol{\theta}}[\sum_a \widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}\pi_{\boldsymbol{\theta}}^t] = 0. \qquad \square\end{aligned}$$

It turns out that our estimator subsumes many previous ones as its special cases.

Special case when $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ is not a function of $\boldsymbol{\theta}$ When we treat $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ not as a function of $\boldsymbol{\theta}$, i.e., $\nabla\widetilde{Q}^{\pi_{\boldsymbol{\theta}}} \equiv 0$, $\nabla\widetilde{V}^{\pi_{\boldsymbol{\theta}}} \equiv \sum_a \widetilde{Q}^{\pi_{\boldsymbol{\theta}}}\nabla\pi_{\boldsymbol{\theta}}$, the estimator becomes[3]

$$\begin{aligned}&\sum_{t=0}^{T}\Big\{\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big[\sum_{t_1=t}^{T}\gamma^{t_1}r_{t_1} + \sum_{t_2=t+1}^{T}\gamma^{t_2}\Big(\widetilde{V}_{t_2}^{\pi_{\boldsymbol{\theta}}} - \widetilde{Q}_{t_2}^{\pi_{\boldsymbol{\theta}}}\Big)\Big] \\ &+ \gamma^t\Big(\nabla_{\boldsymbol{\theta}}\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}} - \widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big)\Big\}, \qquad (11)\end{aligned}$$

which is exactly the same as the one given by Cheng et al. (2019). We will compare the variance of the two estimators below and discuss when our general form can reduce more variance.

Special case when $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}(s,a)$ depends on neither $\boldsymbol{\theta}$ nor a As a more restrictive special case, when $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ is only a function of its state argument, we essentially recover the baseline method. This is obvious by comparing the correponding OPE estimator of PG with baseline to DR, and noticing that they are equivalent when we let $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}(s,a) := b(s)$.

Special case when $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}} \equiv 0$ As a further special case, when the approximate Q-value function is always 0, we recover the standard PG estimator, which corresponds to step-wise IS.

5.2. Variance Analysis

In this section, we analyze the variance of the DR-PG estimator given in Eq.(8).

Theorem 6. *The covariance matrix of the estimator Eq.(8) is*

$$\begin{aligned}\mathbb{E}\Big[\sum_{n=0}^{T}\gamma^{2n}\Big(&\mathbb{V}_{n+1}[r_n]\Big(\sum_{t=0}^{n}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big)\Big(\sum_{t=0}^{n}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big)^\top \\ &+ \mathrm{Cov}_n\Big[\nabla_{\boldsymbol{\theta}}Q_n^{\pi_{\boldsymbol{\theta}}} - \nabla_{\boldsymbol{\theta}}\widetilde{Q}_n^{\pi_{\boldsymbol{\theta}}} \\ &\quad + \Big(\sum_{t=0}^{n}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big)\Big(Q_n^{\pi_{\boldsymbol{\theta}}} - \widetilde{Q}_n^{\pi_{\boldsymbol{\theta}}}\Big)\Big|s_n\Big] \\ &+ \mathrm{Cov}_n\Big[\nabla_{\boldsymbol{\theta}}V_n^{\pi_{\boldsymbol{\theta}}} + \Big(\sum_{t=0}^{n-1}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}^t\Big)V_n^{\pi_{\boldsymbol{\theta}}}\Big]\Big)\Big]. \qquad (12)\end{aligned}$$

where $\mathrm{Cov}_n[\cdot]$ *denotes the covariance matrix of a column vector (defined as* $\mathrm{Cov}_n[\mathbf{v}] := \mathbb{E}_n[\mathbf{v}\mathbf{v}^\top] - \mathbb{E}_n[\mathbf{v}]\mathbb{E}_n[\mathbf{v}]^\top$*), and we omit 0 in* $\mathbb{E}_0$ *and* Cov_0.

We defer the proof to Appendix D. Besides, since many common estimators are special cases of DR-PG, we can obtain their variances as direct corollaries of Theorem 3.

Discussions As we can see, the approximate value-function can help reduce the second term in Eq.(12) when both $Q^{\pi_{\boldsymbol{\theta}}}$ and $\nabla_{\boldsymbol{\theta}}Q^{\pi_{\boldsymbol{\theta}}}$ are well approximated by $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ and $\nabla_{\boldsymbol{\theta}}\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ respectively. Comparing with Cheng et al. (2019), which is our special case with $\nabla_{\boldsymbol{\theta}}\widetilde{Q}^{\pi_{\boldsymbol{\theta}}} \equiv \mathbf{0}$, we can see that as long as $\nabla_{\boldsymbol{\theta}}\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ is a better approximation of $\nabla_{\boldsymbol{\theta}}Q^{\pi_{\boldsymbol{\theta}}}$ than $\mathbf{0}$, the new estimator will generally have lower variance than the previous one. We note that such a situation is very common in variance reduction by control variates; we refer the readers to Jiang & Li (2016) for very similar discussions when they compare DR to step-wise IS.

5.3. Cramer-Rao Lower Bound for Policy Gradient

We now state the Cramer-Rao lower bound for policy gradient, which is a lower bound for *any* unbiased estimator for the PG problem. As we will see, the DR-PG estimator achieves the C-R bound of PG when the MDP has a tree structure and both $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ and $\nabla_{\boldsymbol{\theta}}\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ are accurate, a property inherited directly from the DR estimator in OPE (Jiang & Li, 2016, Theorem 2). As a special case, when we further assume that the environment is fully deterministic (but the policy can still be stochastic), **DR-PG is the only estimator that achieves 0 variance with accurate side information**, and other estimators have non-zero variance in general (see Table 1).

Theorem 7 (Informal)**.** *For tree-structured MDPs (i.e., each state only appears at a unique time step and can be reached*

[3] Note that here $\nabla_{\boldsymbol{\theta}}\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}}$ is in general non-zero even when $\nabla_{\boldsymbol{\theta}}\widetilde{Q}_t^{\pi_{\boldsymbol{\theta}}} \equiv \mathbf{0}$, as $\nabla_{\boldsymbol{\theta}}\widetilde{V}_t^{\pi_{\boldsymbol{\theta}}}$ additionally depends on $\boldsymbol{\theta}$ through the expectation over actions drawn from $\pi_{\boldsymbol{\theta}}$ when we convert Q-value to V-value.

by a unique trajectory), the Cramer-Rao lower bound of PG is

$$\mathbb{E}\Big[\sum_{t=0}^{T}\gamma^{2t}\Big\{\mathbb{V}_{t+1}[r_t]\Big[\Big(\sum_{t_1=0}^{t}\frac{\partial\log\pi_{\boldsymbol{\theta}}^{t_1}}{\partial\theta_i}\Big)\Big]^2+ \mathbb{V}_t\Big[\Big(V_t^{\pi_{\boldsymbol{\theta}}}\sum_{t_1=0}^{t-1}\frac{\partial\log\pi_{\boldsymbol{\theta}}^{t_1}}{\partial\theta_i}+\frac{\partial V_t^{\pi_{\boldsymbol{\theta}}}}{\partial\theta_i}\Big)\Big]\Big\}\Big],$$

which coincides with the variance of DR-PG when $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}\equiv Q^{\pi_{\boldsymbol{\theta}}}$ *and* $\nabla_{\boldsymbol{\theta}}\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}\equiv\nabla_{\boldsymbol{\theta}}Q^{\pi_{\boldsymbol{\theta}}}$*.*

Please refer to Appendix E for the formal definitions and theorem statements, where we prove the more general results for DAG MDPs (Theorem 17) and induce the lower bound for tree-MDPs as a direct corollary (Remark 18).

5.4. Practical Considerations

It is worth pointing out that the new estimator requires more information than $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$: it also requires $\nabla_{\boldsymbol{\theta}}\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$, which is sometimes not available, e.g., when $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ is obtained by applying a model-free algorithm on a separate dataset. However, when an approximate dynamics model of the MDP is available (as considered by Cheng et al. (2019); Jiang & Li (2016)), both $\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ and $\nabla_{\boldsymbol{\theta}}\widetilde{Q}^{\pi_{\boldsymbol{\theta}}}$ can be computed by running simulations in the approximate model *for each data point*, where the former can be estimated by Monte-Carlo and the latter can be estimated by the PG estimators. Since we need to draw multiple trajectories starting from each (s_t, a_t) in the dataset, the approach will be computationally intensive and not suitable for situations where the original problem is also a simulation. The new estimator is most likely useful when the bottleneck is the sample efficiency in the real environment and computation in the approximate model is relatively cheap.

Despite the computational intensity, in the next section we provide proof-of-concept experimental results showing the variance reduction benefits of the new estimator compared to prior baselines.

6. Experiments

In this section we empirically validate the effectiveness of DR-PG. Most of our experiment settings follow exactly from Cheng et al. (2019) (we reuse their code).

6.1. Setup

Compared Methods We empirically demonstrate the variance reduction effect and the optimization results of the new DR-PG estimator, and compare it to the following methods: (a) Standard PG, (b) Standard PG with state-dependent baseline, (c) Standard PG with state-action-dependent baseline, (d) Standard PG with trajectory-wise baseline. For simplicity we will drop the prefix "Standard PG" when referring to the methods (b)–(d). See Appendix F for the detailed implementations of these methods.

Environments and Approximate Models We use the CartPole Environment in OpenAI Gym (Brockman et al., 2016) with DART physics engine (Lee et al., 2018), and set the horizon length to be 1000. We follow Cheng et al. (2019) for the choice of neural network architecture and training methods in building the policy π, value function estimator $\widetilde{V}$, and dynamics model $\widetilde{d}$.

Implementation of the Estimators For state-baseline, we use $\widetilde{V}$ as the baseline function. For the state-action-baseline, Traj-CV, and DR-PG, we additionally need $\widetilde{Q}(s_t, a_t)$, which is computed by the combination of $\widetilde{V}$ and $\widetilde{d}$: $\widetilde{Q}(s_t, a_t) := r(s_t, a_t) + \delta\widetilde{V}(\widetilde{d}(s_t, a_t))$, where δ is a hyperparameter that plays the role of discount factor introduced by Cheng et al. (2019).

For each state s_t we use Monte Carlo (1000 samples) to compute the expectation (over a_t) of $Q(s_t, a_t)\nabla\log\pi(s_{t'}, a_{t'})$, where $t' = t$ for state-action baseline and $t' = t, t+1, ..., T$ for Traj-CV and our method. All these design choices are taken from Cheng et al. (2019) as-is.

Our DR-PG requires additional estimation of $\nabla Q^{\pi_{\boldsymbol{\theta}}}$ and its expectation $\mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\nabla Q^{\pi_{\boldsymbol{\theta}}}]$, and we obtain them by Monte Carlo. To estimate $\nabla Q^{\pi_{\boldsymbol{\theta}}}(s, a)$, we sample n_q trajectories with π and $\widetilde{d}$, starting from (s, a) with the maximum length no larger than L. Since we are solving another policy gradient problem now (one in the biased dynamics model), we choose state-baseline with $\widetilde{V}$ as a computationally-cheap variance reduction method to speed up the computation (c.f. Section 5.4), and use a different discount factor γ' (again for variance reduction (Baxter & Bartlett, 2001; Jiang et al., 2015)). As for $\mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\nabla Q^{\pi_{\boldsymbol{\theta}}}(s, a)]$, we first sample n_v actions at state s, and then use the above procedure to estimate $\nabla Q^{\pi_{\boldsymbol{\theta}}}(s, a_i)$ for $i \in \{1, 2, 3, \dots, n_v\}$. Finally, we compute the mean of them as the expectation. We choose $n_q = 20, n_v = 20, L = 30$ and $\gamma' = 0.9$ in the actual experiments. We observe that even with relatively small n_q, n_v, there is already significant variance reduction effect for our method. See Appendix F for further details.

6.2. Variance Reduction Comparison

We first compare the variance reduction benefits of DR-PG against several baseline methods. The variance reduction ratio is defined as $\frac{\widehat{\mathbb{V}}_G-\widehat{\mathbb{V}}_{DR}}{\widehat{\mathbb{V}}_G}$, where $\widehat{\mathbb{V}}_G$ denotes the sum of the policy gradients estimator G's variance over all 194 parameters (i.e., the trace of the covariance matrix), and G can be Standard PG, state-dependent baseline, state-action-dependent baseline, or trajectory-wise control variate.

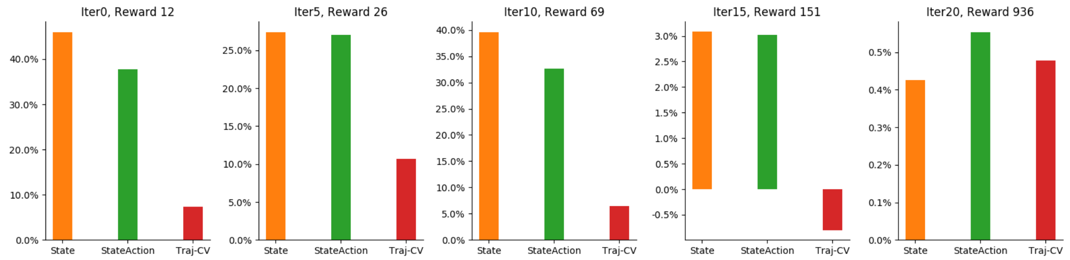

Figure 1. The variance reduction ratio of DR-PG comparing with (1) State-dependent baseline (orange), (2) State-action-dependent baseline (green), (3) Trajectory-wise control variate (red). We omit the results of Standard PG, because DR-PG enjoys a great variance reduction ratio more than 60% in each iteration. Y-axes show the variance reduction ratio. In each sub-title, we indicate the iteration number and the evaluation results of the policy at that iteration.

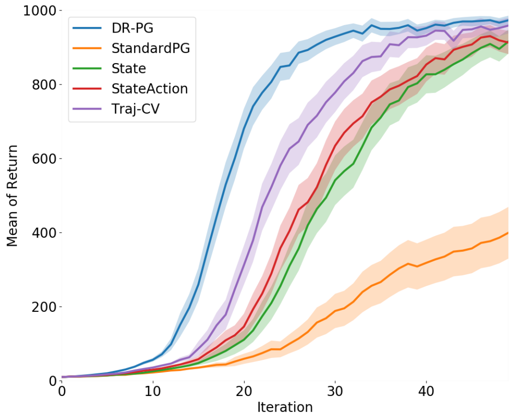

Figure 2. Comparison of different PG estimators in policy optimization. Y-axes show the mean of the expected returns of the policies over 150 trials learned by different PG methods. Error bars show double the standard errors, which correspond to 95% confidence intervals.

The results are shown in Figure 1. To display the variance reduction results in different training stages, we first train a randomly initialized agent with DR-PG for multiple iterations and stop the training when the policy is near optimal (nearly 20 iterations in total in this case). Every 5 iteration, we save the policy π_i, as well as the value function estimator $\widetilde{V}_i$ and dynamic estimator $\widetilde{d}_i$.

For each $(\pi_i, \widetilde{V}_i, \widetilde{d}_i)$ tuple, we first use 10000 sampled trajectories to build state-dependent baseline estimators and compute the mean of as the (estimated) groundtruth. (We use state-dependent baseline because this method suffers less variance than standard PG and is computationally cheap compared to other control variate methods). Next, we sample another 500 trajectories and calculate the mean squared error w.r.t. the approximate true gradient mentioned above for each gradient estimator, which gives an estimation of the estimators' variance (since all estimators considered here are unbiased).

As we can see in Figure 1, DR-PG has better variance reduction effect than the other methods. At the initial training stage (Iterations 0-10), DR-PG can be uniformly better than the others, and the variance reduction ratio is quite large. When the policy is close to optimal (Iterations 15-20), the complexity of the true value function will increase, as the trajectories will last longer and the true values of different states become much more distinct than before. As a result, it is more difficult for the value function estimator to make accurate predictions, hence the estimation of $\widetilde{Q}^\pi$ and $\nabla\widetilde{Q}^\pi$ are less accurate and the variance reduction ratio decreases. However, our method still has advantage over the others in these cases. Moreover, we will show in the next section that such decrease in the final training stage will not stop DR-PG from achieving a near-optimal policy with less amount of data from real environment, i.e., attaining better sample efficiency.

6.3. Policy Optimization

In this experiment, we directly compare the policy optimization performance of different PG methods, i.e., they generate different sequences of policies now (as opposed to computing the gradient for the same policies as in the previous experiment). In each iteration, we record the mean of the accumulated rewards over 5 trajectories sampled from true environment, and use them to represent the performance of the policy. We repeat multiple trials of the entire experiment with different random seeds and plot the mean expected return in Figure 2. As can be clearly seen, DR-PG achieves a near-optimal performance with a significantly less amount of data drawn from the real environment compared to the baseline algorithms.

6.4. Computational Cost

To understand the computational overhead of our method due to having to compute $\widetilde{Q}$ and $\nabla\widetilde{Q}$, we compare the computational cost of applying the PG estimators to train the policy till near-optimal, when the policy value averaged over 5 trajectories exceeds 900 for the first time (the optimal return is around 1000). The total CPU/GPU usage is reported in Figure 3, where we omit the standard PG because it costs much more than the others due to extended number of training iterations. As we can see, DR-PG requires a reasonable amount of additional computational resources compared to the other estimators.

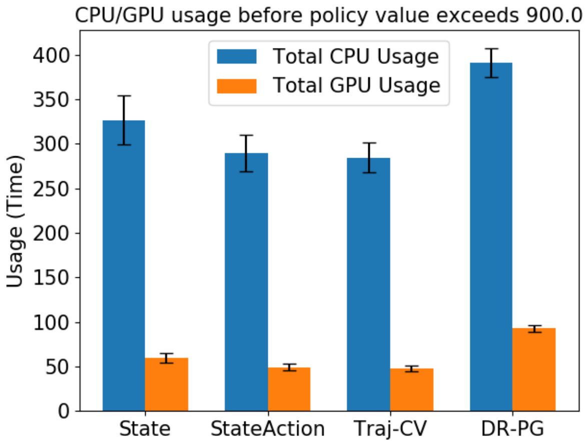

Figure 3. Comparison of computational costs. Y-axes show the total CPU/GPU usage. Error bars show double the standard errors.

7. Conclusion

This paper investigates a direct connection between variance reduction techniques for on-policy policy gradient and for off-policy evaluation with importance sampling. From the DR estimator for OPE, we derive a very general form of PG that subsumes many previous estimators as special cases, and achieve more variance reduction in the ideal situation with accurate side information.

Acknowledgement

The research project is motivated by a question asked by Lihong Li in 2015. The authors gratefully thank Ching-An Cheng and his coauthors for the code of the trajectory-wise control variates method and the valuable comments.

References

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Cheng, C.-A., Yan, X., and Boots., B. Trajectory-wise control variates for variance reduction in policy gradient methods. In *Proceedings of the 2019 Conference on Robot Learning (CoRL)*, 2019.

Grathwohl, W., Choi, D., Wu, Y., Roeder, G., and Duvenaud, D. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. In *International Conference on Learning Representations*, 2018.

Greensmith, E., Bartlett, P. L., and Baxter, J. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(Nov): 1471–1530, 2004.

Gu, S., Lillicrap, T. P., Ghahramani, Z., Turner, R. E., and Levine, S. Q-prop: Sample-efficient policy gradient with an off-policy critic. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661, 2016.

Jiang, N., Kulesza, A., Singh, S., and Lewis, R. The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1181–1189, 2015.

Lee, J., Grey, M. X., Ha, S., Kunz, T., Jain, S., Ye, Y., Srinivasa, S. S., Stilman, M., and Liu, C. K. DART: dynamic animation and robotics toolkit. *J. Open Source Software*, 3(22):500, 2018.

Liu, H., Feng, Y., Mao, Y., Zhou, D., Peng, J., and Liu, Q. Action-dependent control variates for policy optimization via stein identity. In *International Conference on Learning Representations*, 2018.

Moore, T. J. *A Theory of Cramer-Rao Bounds for Constrained Parametric Models*. PhD thesis, 2010.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Tang, J. and Abbeel, P. On a connection between importance sampling and the likelihood ratio policy gradient. In *Advances in Neural Information Processing Systems*, pp. 1000–1008, 2010.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148, 2016.

Tucker, G., Bhupatiraju, S., Gu, S., Turner, R., Ghahramani, Z., and Levine, S. The mirage of action-dependent baselines in reinforcement learning. In *International Conference on Machine Learning*, pp. 5022–5031, 2018.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Wu, C., Rajeswaran, A., Duan, Y., Kumar, V., Bayen, A. M., Kakade, S., Mordatch, I., and Abbeel, P. Variance reduction for policy gradient with action-dependent factorized baselines. In *International Conference on Learning Representations*, 2018.