Statistically Efficient Off-Policy Policy Gradients

Nathan Kallus [1] **Masatoshi Uehara** [2]

Abstract

Policy gradient methods in reinforcement learning update policy parameters by taking steps in the direction of an estimated gradient of policy value. In this paper, we consider the statistically efficient estimation of policy gradients from off-policy data, where the estimation is particularly non-trivial. We derive the asymptotic lower bound on the feasible mean-squared error in both Markov and non-Markov decision processes and show that existing estimators fail to achieve it in general settings. We propose a meta-algorithm that achieves the lower bound without any parametric assumptions and exhibits a unique 3-way double robustness property. We discuss how to estimate nuisances that the algorithm relies on. Finally, we establish guarantees at the rate at which we approach a stationary point when we take steps in the direction of our new estimated policy gradient.

1. Introduction

Learning sequential decision policies from observational off-policy data is an important problem in settings where exploration is limited and simulation is unreliable. A key application is reinforcement learning (RL) for healthcare (Gottesman et al., 2019). In such settings, data is limited and it is crucial to use the available data *efficiently*. Recent advances in off-policy evaluation (Kallus & Uehara, 2019a;b) have shown how efficiently leveraging problem structure, such as Markovianness, can significantly improve off-policy evaluation and tackle such sticky issues as the curse of horizon (Liu et al., 2018). An important next step is to translate these successes in off-policy *evaluation* to off-policy *learning*. In this paper we tackle this question by studying the efficient estimation of the *policy gradient* from off-policy data and the implications of this for learning via estimated-policy-gradient ascent.

*Equal contribution [1]Cornell University, Ithaca, NY, USA [2]Harvard University, Massachusetts, Boston, USA. Correspondence to: Masatoshi Uehara <ueharamasatoshi136@gmail.com>.

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

Table 1. Comparison of off-policy policy gradient estimators. Here, $f = \Theta(g)$ means $0 < \liminf f/g \leq \limsup f/g < \infty$ (not to be confused with the policy parameter space Θ). In the second row, nuisances must be estimated at $n^{-1/2}$-rates, and in the rows below it, nuisances may be estimated at slow non-parametric rates.

Estimator	MSE	Efficient	Nuisances
Reinforce, Eq. (4)	$2^{\Theta(H)}\Theta(1/n)$		none
PG, Eq. (5)	$2^{\Theta(H)}\Theta(1/n)$		q (parametric)
EOPPG (NMDP)	$2^{\Theta(H)}\Theta(1/n)$	✓	$q, \nabla q$
EOPPG (MDP)	$\Theta(H^4/n)$	✓	$q, \mu, \nabla q, \nabla \mu$

Policy gradient algorithms (Sutton & Barto, 2018, Chapter 13) enable one to effectively learn complex, flexible policies in potentially non-tabular, non-parametric settings and are therefore very popular in both on-policy and off-policy RL. We begin by describing the problem and our contributions, before reviewing the literature in Section 1.2.

Consider a $(H+1)$-long Markov decision process (MDP), with states $s_t \in \mathcal{S}_t$, actions $a_t \in \mathcal{A}_t$, rewards $r_t \in \mathbb{R}$, initial state distribution $p_0(s_0)$, transition distributions $p_t(s_{t+1} \mid s_t, a_t)$, and reward distribution $p_t(r_t \mid s_t, a_t)$, for $t = 0, \dots, H$. A policy $(\pi_t(a_t \mid s_t))_{t \leq H}$ induces a distribution over trajectories $\mathcal{T} = (s_0, a_0, r_0, \dots, s_T, a_H, r_H, s_{H+1})$:

$$p_\pi(\mathcal{T}) = \tag{1}$$
$$p_0(s_0) \textstyle\prod_{t=0}^{H} \pi_t(a_t \mid s_t) p_t(r_t \mid s_t, a_t) p_t(s_{t+1} \mid s_t, a_t).$$

Given a class of policies $\pi_t^\theta(a_t \mid s_t)$ parametrized by $\theta \in \Theta \in \mathbb{R}^D$, we seek the parameters with greatest average reward, defined as

$$J(\theta) = \mathbb{E}_{p_{\pi^\theta}}\left[\textstyle\sum_{t=0}^{H} r_t\right]. \qquad \text{(Policy Value)}$$

A generic approach is to repeatedly move θ in the direction of the policy gradient (PG), defined as

$$Z(\theta) = \nabla_\theta J(\theta) \qquad \text{(Policy Gradient)}$$
$$= \mathbb{E}_{p_{\pi^\theta}}\left[\textstyle\sum_{t=0}^{H} r_t \sum_{k=0}^{t} \nabla_\theta \log \pi_k^\theta(a_k \mid s_k)\right]$$

For example, in the *on-policy* setting, we can generate trajectories from π^θ, $\mathcal{T}^{(1)}, \dots, \mathcal{T}^{(n)} \sim p_{\pi^\theta}$, and the (GPOMDP variant of the) REINFORCE algorithm (Baxter & Bartlett, 2001) advances in the direction of the stochastic gradient

$$\hat{Z}^{\text{on-policy}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{H} r_t^{(i)} \sum_{k=0}^{t} \nabla_\theta \log \pi_k^\theta(a_k^{(i)} \mid s_k^{(i)}).$$

In the *off-policy* setting, however, we *cannot* generate trajectories from any given policy and, instead our data consists only of trajectory observations from one fixed policy,

$$\mathcal{T}^{(1)}, \dots, \mathcal{T}^{(n)} \sim p_{\pi^b}, \qquad \text{(Off-policy data)}$$

where π^b is known as the *behavior policy*. With this data, $\hat{Z}^{\text{on-policy}}(\theta)$ is no longer a stochastic gradient (*i.e.*, it is biased *and* inconsistent) and we must seek other ways to estimate $Z(\theta)$ in order to make policy gradient updates.

This paper addresses the *efficient* estimation of $Z(\theta)$ from off-policy data and its use in off-policy policy learning. Specifically, our contributions are:

(Section 2) We calculate the asymptotic lower bound on the minimal-feasible mean-squared error (MSE) in estimating the policy gradient, which is of order $\mathcal{O}(H^4/n)$. In addition, we demonstrate that existing off-policy policy gradient approaches fail to achieve this bound and may even have exponential dependence on the horizon.

(Section 3.1) We propose a meta-algorithm called Efficient Off-Policy Policy Gradient (EOPPG) that achieves this bound without any parametric assumptions. In addition, we prove it enjoys a unique 3-way double robustness property.

(Section 3.3) We show how to estimate the nuisance functions needed for our meta-algorithm by introducing the concepts of Bellman equations for the gradient of q-function and stationary distributions.

(Section 4) We establish guarantees for the rate at which we approach a stationary point when we take steps in the direction of our new estimated policy gradient. Based on efficiency results for our gradient estimator, we can prove the regret's horizon dependence is H^2.

1.1. Notation and definitions

We define the following variables (note the implicit dependence on θ):

$$\begin{aligned}
g_t &= \nabla_\theta \log \pi_{\theta,t}(a_t \mid s_t), && \text{(Policy score)}\\
q_t &= \mathbb{E}_{p_{\pi^\theta}}\left[\textstyle\sum_{k=t}^H r_k \mid s_t, a_t\right], && (q\text{-function})\\
v_t &= \mathbb{E}_{p_{\pi^\theta}}\left[\textstyle\sum_{k=t}^H r_k \mid s_t\right], && (v\text{-function})\\
\tilde{\nu}_t &= \tfrac{\pi_t^\theta(a_t|s_t)}{\pi_t^b(a_t|s_t)}, && \text{(Density Ratio)}\\
\nu_{t':t} &= \textstyle\prod_{k=t'}^t \tilde{\nu}_k, && \text{(Cumulative Density Ratio)}\\
\tilde{\mu}_t &= \tfrac{p_{\pi^\theta}(s_t)}{p_{\pi^b}(s_t)} && \text{(Marginal State Density Ratio)}\\
\mu_t &= \tilde{\mu}_t\tilde{\nu}_t, && \text{(Marginal State-Action Density Ratio)}\\
d_t^q &= \nabla_\theta q_t, d_t^v = \nabla_\theta v_t, d_t^\mu = \nabla_\theta \mu_t, d_t^\nu = \nabla_\theta \nu_{0:t}.
\end{aligned}$$

Note that all of the above are simply functions of the trajectory, $\mathcal{T}$, and θ. To make this explicit, we sometimes write, for example, $q_t = q_t(s_t, a_t)$ and refer to q_t as a function. Similarly, when we estimate this function by $\hat{q}_t$ we also refer to $\hat{q}_t$ as the random variable gotten by evaluating it on the trajectory, $\hat{q}_t(s_t, a_t)$.

We write $a \lessapprox b$ to mean that there exists a *universal* constant C satisfying $a \le Cb$, such as a number like 5, which doesn't depend on any instance-specific parameters. We let $\|\cdot\|_2$ denote the Euclidean vector norm and $\|\cdot\|_{\text{op}}$ denote the matrix operator norm.

All expectations, variances, and probabilities *without* subscripts are understood to be with respect to p_{π^b}. Given a vector-valued function of trajectory, f, we define its L_2 norm as

$$\|f\|_{L_b^2}^2 = \mathbb{E}\|f(\mathcal{T})\|_2^2.$$

Further, given trajectory data, $\mathcal{T}^{(1)}, \dots, \mathcal{T}^{(n)}$, we define the *empirical expectation* as

$$\mathbb{E}_n f = \mathbb{E}_n[f(\mathcal{T})] = \tfrac{1}{n}\textstyle\sum_{i=1}^n f(\mathcal{T}^{(i)}).$$

MDP and NMDP. Throughout this paper, we focus on the MDP setting where the trajectory distribution p_π is given by Eq. (1). For completeness, we also consider relaxing the Markov assumption, yielding a *non-Markov* decision process (NMDP), where the trajectory distribution $p_\pi(\mathcal{T})$ is

$$p_0(s_0)\textstyle\prod_{t=0}^H \pi_t(a_t \mid \mathcal{H}_{s_t}) p_t(r_t \mid \mathcal{H}_{a_t}) p_t(s_{t+1} \mid \mathcal{H}_{a_t}),$$

where $\mathcal{H}_{a_t}$ is $(s_0, a_0, \cdots, a_t)$ and $\mathcal{H}_{s_t}$ is $(s_0, a_0, \cdots, s_t)$.

Assumptions. Throughout we assume that $\forall t \le H$: $0 \le r_t \le R_{\max}$, $\|g_t\|_{\text{op}} \le G_{\max}$, $\tilde{\nu}_t \le C_1$, $\tilde{\mu}_t \le C_2'$. And, we define $C_2 = C_1 C_2'$ so that $\mu_t \le C_2$.

The bounds on $\tilde{\nu}_t, \mu_t$ are often called (sequential) overlap or concentrability conditions. The bounds on g_t is used to bound the variance of the policy gradient. A similar assumption is made in Agarwal et al. (2019). The above uniform bounds may also be replaced with bounds on second moments at the cost of stronger conditions on nuisance estimate convergence in, *e.g.*, Theorem 7; we omit these details to focus on the more common L_2 mode of estimate convergence and uniform bounds on rewards, density ratios, and policy scores.

1.2. Related literature

1.2.1. Off-policy policy gradients

A standard approach to dealing with off-policy data is to correct the policy gradient equation using *importance sampling* (IS) using the cumulative density ratios, $\nu_{0:t}$ (see, *e.g.*, Papini et al., 2018, Appendix A; Hanna & Stone, 2018).

This allows us to rewrite the policy gradient $Z(\theta)$ as an expectation over p_{π^b} and then estimate it using an equivalent empirical expectation.

The off-policy version of the classic REINFORCE algorithm (Williams, 1992) recognizes

$$Z(\theta) = \mathbb{E}\left[\nu_{0:H}\left(\sum_{t=0}^{H} r_t\right)\left(\sum_{t=0}^{H} g_t\right)\right] \tag{2}$$

(recall that $\mathbb{E}$ is understood as $\mathbb{E}_{p_{\pi^b}}$) and uses the estimated policy gradient given by replacing $\mathbb{E}$ with $\mathbb{E}_n$. The GPOMDP variant (Baxter & Bartlett, 2001) refines this by

$$Z(\theta) = \mathbb{E}\left[\nu_{0:H}\sum_{t=0}^{H} r_t \sum_{s=0}^{t} g_s\right], \tag{3}$$

whose empirical version ($\mathbb{E}_n$) has *less* variance and is therefore preferred. A further refinement is given by a step-wise IS (Precup et al., 2000) as in Deisenroth et al. (2013):

$$Z(\theta) = \mathbb{E}\left[\sum_{t=0}^{H} \nu_{0:t} r_t \sum_{s=0}^{t} g_s\right]. \tag{4}$$

Following Degris et al. (2012), Chen et al. (2019) replace $\nu_{0:t}$ with $\tilde{\nu}_t$ in Eq. (4) to reduce variance, but this is an *approximation* that incurs non-vanishing bias.

By exchanging the order of summation in Eq. (4) and recognizing $q_t = \mathbb{E}\left[\sum_{j=t}^{H} \nu_{t+1:j} r_j \mid s_t, a_t\right]$, we obtain a policy gradient in terms of the q-function (Sutton et al., 1998),

$$Z(\theta) = \mathbb{E}\left[\sum_{t=0}^{H} \nu_{0:t} g_t q_t\right]. \tag{5}$$

The off-policy policy gradient (Off-PAC) of Degris et al. (2012) is obtained by replacing $\nu_{0:t}$ with $\tilde{\nu}_t$ in Eq. (5), estimating q_t by $\hat{q}_t$ and plugging it in, and taking the empirical expectation. Replacing $\nu_{0:t}$ with $\tilde{\nu}_t$ is intended to reduce variance but it is an *approximation* that ignores the state distribution mismatch (essentially, μ_t) and incurs non-vanishing bias. Since it amounts to a reweighting and the unconstrained optimal policy remains optimal on any input distribution, in the tabular and fully unconstrained case considered in Degris et al. (2012), we may still converge. But this fails in the general non-parametric, non-tabular setting. We therefore focus only on *consistent* estimates of $Z(\theta)$ in this paper (which requires zero or vanishing bias).

Many of the existing off-policy RL algorithms including DDPG (Silver et al., 2014) and Off-PAC with emphatic weightings (Imani et al., 2018) also use the above trick, *i.e.*, ignoring the state distribution mismatch. Various recent work deals with this problem (Dai et al., 2019; Liu et al., 2019; Tosatto et al., 2020). These, however, both assume the existence of a stationary distribution and are not efficient. We do not assume the existence of a stationary distribution since many RL problems have a finite horizon and/or do not have a stationary distribution. Moreover, our gradient estimates are efficient in that they achieve the MSE lower bound among all regular estimators.

1.2.2. Other Literature

Online off-policy PG. Online policy gradients have shown marked success in the last few years (Schulman et al., 2015). Various work has investigated incorporating offline information into online policy gradients (Gu et al., 2017; Metelli et al., 2018). Compared with this setting, our setting is completely off-policy with no opportunity of collecting new data from arbitrary policies, as considered in, *e.g.*, Athey & Wager (2017); Kallus (2018); Kallus & Zhou (2018); Swaminathan & Joachims (2015) for $H = 0$ and Chen et al. (2019); Fujimoto et al. (2019) for $H \geq 1$.

Variance reduction in PG. Variance reduction has been a central topic for PG (Greensmith et al., 2004; Schulman et al., 2016; Tang & Abbeel, 2010; Wu et al., 2018). These papers generally consider a given class of estimators given by an explicit formula (such as given by all possible baselines) and show that some estimator is optimal among the class. In our work, the class of estimators among which we are optimal is *all* regular estimators, which both extremely general and also provides minimax bounds in any vanishing neighborhood of p_{π^b} (van der Vaart, 1998, Thm. 25.21).

Off-policy evaluation (OPE). OPE is the problem of estimating $J(\theta)$ for a given θ from off-policy data. Step-wise IS (Precup et al., 2000) and direct estimation of q-functions (Munos & Szepesvári, 2008) are two common approaches for OPE. However, the former is known to suffer from the high variance and the latter from model misspecification. To alleviate this, the doubly robust estimate combines the two; however, the asymptotic MSE still explodes exponentially in the horizon like $\Omega(C_1^H H^2/n)$ (Jiang & Li, 2016; Thomas & Brunskill, 2016). Kallus & Uehara (2019b) show that the efficiency bound in the MDP case is actually polynomial in H and give an estimator achieving it, which combines marginalized IS (Xie et al., 2019) and q-modeling using cross-fold estimation. This achieves MSE $O(C_2 H^2/n)$. Kallus & Uehara (2019a) further study efficient OPE in the infinite horizon MDP setting with non-iid batch data.

Offline policy learning There are many types of methods for offline policy learning (batch RL) such as fitted Q-iteration (Munos & Szepesvári, 2008), bellman residual minimization (Antos et al., 2008), and minimax learning (Chen & Jiang, 2019). We focus on the policy gradient approach since it can be easily applied very generally, in particular when actions are continuous. As far as we know, there are few studies of offline policy gradient with regret guarantee as in our Section 4.

2. Efficiency Bound for Estimating $Z(\theta)$

Our target estimand is $Z(\theta)$ so a natural question is what is the least-possible error we can achieve in estimating it. In parametric models, the Cramér-Rao bound lower bounds the variance of all unbiased estimators and, due to Hájek (1970), also the asymptotic MSE of *all* (regular) estimators. Our model, however, is nonparametric as it consists of *all* MDP distributions, *i.e.*, *any* choice for $p_0(s_0)$, $p_t(r_t \mid s_t, a_t)$, $p_t(r_t \mid s_t, a_t)$, and $\pi_t(a_t \mid s_t)$ in Eq. (1). Semiparametric theory gives an answer to this question. We first informally state the key property of the *efficient influence function* (EIF) from semiparametric theory. The EIF is a function defined in terms of only the estimand (map from data generating distribution to quantity of interest), model (set of possible data generating distributions), and instance in the model (the specific unknown data generating distribution). It provides a lower bound on the feasible asymptotic MSE in estimating the estimand, which is sometimes achievable. And, we will then show how to achieve the corresponding bound in the next section. We present the key property in terms of our own model, which is all MDP distributions, and our estimand, which is $Z(\theta)$. For additional detail, see Appendix B and Tsiatis (2006); van der Vaart (1998).

Theorem 1 (Informal description of van der Vaart (1998), Theorem 25.20)**.** *The EIF $\xi_{\text{MDP}}(\mathcal{T}; p_{\pi^b})$ satisfies that for any MDP distribution p_{π^b} and any regular estimator $\hat{Z}(\theta)$,*

$$\inf_{\|v\|_2 \leq 1} v^T(\text{AMSE}[\hat{Z}(\theta)] - \text{var}[\xi_{\text{MDP}}])v = 0,$$

where $\text{AMSE}[\hat{Z}(\theta)] = \int zz^T dF(z)$ is the second moment of F the limiting distribution of $\sqrt{n}(\hat{Z}(\theta) - Z(\theta))$.

This also implies $\|\text{AMSE}[\hat{Z}(\theta)]\|_{\text{op}} \geq \|\text{var}[\xi_{\text{MDP}}]\|_{\text{op}}$. Here, $\text{var}[\xi_{\text{MDP}}]$ is called the *efficiency bound* (note it is a covariance *matrix*). Estimators satisfying $\text{AMSE}[\hat{Z}(\theta)] = \text{var}[\xi_{\text{MDP}}]$ are called *efficient*. A regular estimator is any whose limiting distribution is insensitive to small changes of order $O(1/\sqrt{n})$ to p_{π^b} that keep it an MDP distribution (see van der Vaart, 1998, Chapter 25). So the above provides a lower bound on the variance of all regular estimators, which is a very general class. It is so general that the bound also applies to *all* estimators at all in a local asymptotic minimax sense (see van der Vaart, 1998, Theorem 25.21).

Technically, we first need to prove that the EIF ξ_{MDP} exists in order to obtain the bound in Theorem 1. The following result does so and derives it explicitly (in terms of unknown nuisance functions). The result after does the same in the NMDP model. (Note that, while by the usual convention the EIF refers to a function with 0 mean, instead we let the EIF have mean $Z(\theta)$ everywhere as it simplifies the presentation. Since adding a constant does not change the variance, Theorem 1 is unchanged.)

Theorem 2. *The EIF of $Z(\theta)$ under MDP, ξ_{MDP}, exists and is equal to*

$$\textstyle\sum_{j=0}^{H}(d_j^{\mu}(r_j - q_j) - \mu_j d_j^q + \mu_{j-1} d_j^v + d_{j-1}^{\mu} v_j),$$

where $\mu_{-1} = 1$, $d_{-1}^{\mu} = 0$.

And, in particular,

$$\|\text{var}[\xi_{\text{MDP}}]\|_{\text{op}} \leq C_2 R_{\max}^2 G_{\max}(H+1)^2(H+2)^2/4.$$

Theorem 3. *The EIF of $Z(\theta)$ under NMDP, ξ_{NMDP}, exists and is equal to*

$$\textstyle\sum_{j=0}^{H}(d_j^{\nu}(r_j - q_j) - \nu_{0:j} d_j^q + \nu_{0:j-1} d_j^v + d_{j-1}^{\nu} v_j),$$

where $\nu_{0:-1} = 1$, $d_{-1}^{\nu} = 0$. (Note that here $\nu_{0:j}, d_j^{\nu}, d_j^q, d_j^q$ are actually functions of all of $\mathcal{H}_{a_j}$ and not just of (s_j, a_j) as in MDP case in Theorem 2.)

And, in particular,

$$\|\text{var}[\xi_{\text{NMDP}}]\|_{\text{op}} \leq C_1^H R_{\max}^2 G_{\max}(H+1)^2(H+2)^2/4.$$

Roughly speaking, the EIFs of $Z(\theta)$ in Theorems 2 and 3 are derived by differentiating the EIFs of $J(\theta)$ obtained in Kallus & Uehara (2019b) with respect to θ, since we have the relation $Z(\theta) = \nabla J(\theta)$. That is why d_j^{μ}, d_j^q appear in addition to μ_j, q_j in the EIFs above. In fact,

$$\xi_{\text{MDP}} = \nabla\left\{v_0 + \textstyle\sum_{j=0}^{H} \mu_j(r_j - q_j + v_{j+1})\right\},$$
$$\xi_{\text{NMDP}} = \nabla\left\{v_0 + \textstyle\sum_{j=0}^{H} \nu_{0:j}(r_j - q_j + v_{j+1})\right\}.$$

Formulae for $\text{var}[\xi_{\text{MDP}}]$ and $\text{var}[\xi_{\text{NMDP}}]$ are given in Appendix C. Theorem 3 showed $\text{var}[\xi_{\text{NMDP}}]$ is at most exponential; we next show it is also at least exponential.

Theorem 4. *Suppose that $\tilde{\nu}_t \geq C_3$ and that $\text{var}[(\sum_h g_h)(r_H - q_H) \mid \mathcal{H}_{a_H}] \succeq cI$. Then, $\|\text{var}[\xi_{\text{NMDP}}]\|_{\text{op}} \geq C_3^{2H} c$.*

Theorems 3 and 4 show that the curse of horizon is generally *unavoidable* in NMDP since the lower bound in is at least *exponential* in horizon. On the other hand, Theorem 2 shows there is a possibility we can avoid the curse of horizon in MDP in the sense that the lower bound is at most polynomial in horizon as long as C_2 is bounded as H grows. This is true, for example, if $p_{\pi^b}(s_t)$ converges in distribution, which will necessarily occur if the MDP is ergodic.

We show that REINFORCE necessarily suffers from the curse of horizon.

Theorem 5. *The MSE of step-wise REINFORCE Eq. (4) is*

$$\textstyle\sum_{k=0}^{H+1} \mathbb{E}[\nu_{k-1}^2 \times \text{var}[\mathbb{E}[\sum_{t=k-1}^{H} \nu_{k:t} r_t \sum_{s=k-1}^{t} g_s \mid \mathcal{H}_{a_k}] \mid \mathcal{H}_{a_{k-1}}]],$$

which is no smaller than the MSE of REINFORCE Eq. (2) and GOMDP-REINFROCE Eq. (3). (Equation references refer to the estimate given by replacing $\mathbb{E}$ with $\mathbb{E}_n$.)

Algorithm 1 Efficient Off-Policy Policy Gradient

Take a K-fold random partition $(I_k)_{k=1}^K$ of the observation indices $\{1,\dots,n\}$ such that the size of each fold, $|I_k|$, is within 1 of n/K.
Let $\mathcal{L}_k = \{\mathcal{T}^{(i)} : i \in I_k\}$, $\mathcal{U}_k = \{\mathcal{T}^{(i)} : i \notin I_k\}$
for $k \in \{1,\cdots,K\}$ **do**
Using only $\mathcal{L}_k$ as data, construct nuisance estimators $\hat{q}_j^{(k)}$, $\hat{\mu}_j^{(k)}$, $\hat{d}_j^{q(k)}$, $\hat{d}_j^{\mu(k)}$ for $\forall j \le H$ (see Section 3.3)
By integrating/summing w.r.t $a_j \sim \pi_j^\theta(a_j \mid s_j)$, set

$$\hat{v}_j(s_j) = \mathbb{E}_{\pi_j^\theta}[\hat{q}_j \mid s_j],\ \hat{d}_j^v(s_j) = \mathbb{E}_{\pi_j^\theta}[\hat{d}_j^q + \hat{q}_j g_j \mid s_j] \quad (6)$$

Construct an intermediate estimate:

$$\hat{Z}_k(\theta) = \mathbb{E}_{\mathcal{U}_k}\big[\textstyle\sum_{j=0}^H \big(\hat{d}_j^{\mu(k)}(r_j - \hat{q}_j^{(k)}) - \hat{\mu}_j^{(k)}\hat{d}_j^{q(k)} + \hat{\mu}_{j-1}^{(k)}\hat{d}_j^{v(k)} + \hat{d}_{j-1}^{\mu(k)}\hat{v}_j^{(k)}\big)\big],$$

where $\mathbb{E}_{\mathcal{U}_k}$ is the empirical expectation over $\mathcal{U}_k$
end for
Return $\hat{Z}^{\mathrm{EOPPG}}(\theta) = \frac{1}{K}\sum_{k=1}^K \hat{Z}_k$.

Theorem 6. *Suppose that $\tilde{v}_t \ge C_3$ and that $\mathrm{var}[r_H g_H \mid \mathcal{H}_{a_H}] \succeq cI$. Then, the operator norm of the variance of step-wise REINFORCE is lower bounded by cC_3^{2H}/n.*

3. Efficient Policy Gradient Estimation

In this section we develop an estimator, EOPPG, for $Z(\theta)$ achieving the lower bound in Theorem 2 under weak nonparametric rate assumptions.

3.1. The Meta-Algorithm

Having derived the EIF of $Z(\theta)$ under MDP in Theorem 2, we use a meta-algorithm based on estimating the unknown functions (aka nuisances) $\mu_j, d_j^q, q_j, d_j^\mu$ and plugging them into ξ_{MDP}, as described in Algorithm 1, which we call the Efficient Off-Policy Policy Gradient (EOPPG). In particular, we use a cross-fitting technique (Chernozhukov et al., 2018; van der Vaart, 1998) to avoid technical smoothness conditions on our nuisance estimates. We refer to this as a meta-algorithm as it relies on given nuisances estimators: we show to construct these in Section 3.3.

Note Eq. (6) in Algorithm 1 is computed simply by taking an integral over a_j (or, sum, for finite actions) with respect to the *known* measure (or, mass function) $\pi_j^\theta(a_j \mid s_j)$.

We next prove that EOPPG achieves the efficiency bound under MDP and enjoys a 3-way double robustness (see Fig. 1). We require the following about our nuisance estimators, which arises from the boundedness assumed in Section 1.1.

Assumption 1. *$\forall k \le K$, $\forall j \le H$, we have $0 \le \hat{q}_j^{(k)} \le R_{\max}(H+1-j)$, $\hat{\mu}_j^{(k)} \le C_2$, $\|\hat{d}_j^{q(k)}\|_{\mathrm{op}}, \|\hat{d}_j^{\mu(k)}\|_{\mathrm{op}} \le C_4$.*

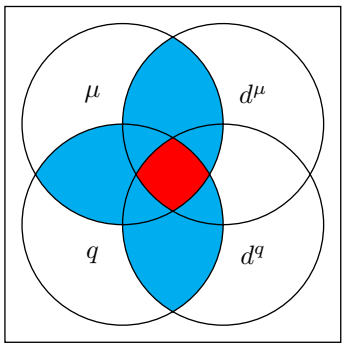

Figure 1. Doubly robust and efficient structure of EOPPG. Every circle represents the event that the corresponding nuisance is well-specified. The cyan-shaded region represents the event that $\hat{Z}^{\mathrm{EOPPG}}(\theta)$ is consistent. The red-shaded region represents the event that $\hat{Z}^{\mathrm{EOPPG}}(\theta)$ is efficient (when nuisances are consistently estimated non-parametrically at slow rates).

Theorem 7 (Efficiency)**.** *Suppose $\forall k \le K$, $\forall j \le H$,*

$$\|\hat{\mu}_j^{(k)} - \mu_j\|_{L_b^2} = \mathrm{o}_p(n^{-\alpha_1}),\ \|\hat{d}_j^{\mu(k)} - d_j^\mu\|_{L_b^2} = \mathrm{o}_p(n^{-\alpha_2}),$$
$$\|\hat{q}_j^{(k)} - q_j\|_{L_b^2} = \mathrm{o}_p(n^{-\alpha_3}),\ \|\hat{d}_j^{q(k)} - d_j^q\|_{L_b^2} = \mathrm{o}_p(n^{-\alpha_4}),$$

$\min(\alpha_1,\alpha_2)+\min(\alpha_3,\alpha_4) \ge 1/2$ and $\alpha_1,\alpha_2,\alpha_3,\alpha_4 > 0$. Then, $\sqrt{n}(\hat{Z}^{\mathrm{EOPPG}}(\theta) - Z(\theta)) \xrightarrow{d} \mathcal{N}(0, \mathrm{var}[\xi_{\mathrm{MDP}}])$.

An important feature of Theorem 7 is that the required nuisance convergence rates are nonparametric (slower than $n^{-1/2}$) and no metric entropy condition (*e.g.*, Donsker) is needed. In particular, the result does not depend on the particular nuisance estimates used, and we experience no variance inflation due to plugging-in estimates instead of true nuisances. While usually we can expect inflation due to nuisance variance (*e.g.*, PG Eq. (5) generally has MSE worse than $\Theta(n^{-1/2})$ if we use an estimate $\hat{q}$ with a nonparametric rate), we avoid this due to the special doubly robust structure of ξ_{MDP} that renders our estimate insensitive to the way nuisances are estimated.

To establish this doubly robust structure – the key step of the proof – we show that $\hat{Z}^{\mathrm{EOPPG}}(\theta)$ is equal to

$$\mathbb{E}_n[\xi_{\mathrm{MDP}}] + K^{-1}\textstyle\sum_{k=1}^K\sum_{j=0}^H \mathrm{Bias}_{k,j} + \mathrm{o}_p(n^{-1/2}), \quad (7)$$

where $\|\mathrm{Bias}_{k,j}\|_2$ is upper bounded up to constant by the following

$$\begin{aligned}\|\mathrm{Bias}_{k,j}\|_2 \lesssim\!\!\!\approx\ & \|\hat{\mu}_j^{(k)} - \mu_j\|_{L_b^2}\|\hat{d}_j^{q(k)} - d_j^q\|_{L_b^2} \\ &+ \|\hat{d}_j^{\mu(k)} - d_j^\mu\|_{L_b^2}\|\hat{q}_j^{(k)} - q_j\|_{L_b^2} \\ &+ \|\hat{\mu}_{j-1}^{(k)} - \mu_{j-1}\|_{L_b^2}\|\hat{d}_j^{v(k)} - d_j^v\|_{L_b^2} \\ &+ \|\hat{d}_{j-1}^{\mu(k)} - d_{j-1}^\mu\|_{L_b^2}\|\hat{v}_j^{(k)} - v_j\|_{L_b^2}.\end{aligned} \quad (8)$$

After establishing this key result, Eqs. (7) and (8), Theorem 7 follows by showing that the bias term is $\mathrm{o}_p(n^{-1/2})$

and applying CLT. This asymptotic results can also be extended to a finite-sample results by assuming finite-sample bounds on nuisance estimates, following Kallus & Uehara (2019b).

We also obtain the following from Eq. (7) via LLN.

Theorem 8 (3-way double robustness). *Suppose* $\forall k \leq K, \forall j \leq H$, $\|\hat{\mu}_j^{(k)} - \mu_j^\dagger\|_{L_b^2}, \|\hat{d}_j^{q(k)} - d_j^{q\dagger}\|_{L_b^2}, \|\hat{q}_j^{(k)} - q_j^\dagger\|_{L_b^2}, \|\hat{d}_j^{\mu(k)} - d_j^{\mu\dagger}\|_{L_b^2}$ *all converge to 0 in probability. Then* $\hat{Z}^{\mathrm{EOPPG}}(\theta) \to_p Z(\theta)$ *if one the following hold:* $\mu_j^\dagger = \mu_j, d_j^{\mu\dagger} = d_j^\mu$; $q_j^\dagger = q_j, d_j^{q\dagger} = d_j^q$; *or* $\mu_j^\dagger = \mu_j, q_j^\dagger = q_j$.

That is, as long as (a) $\hat{\mu}, \hat{d}^\mu$ are correct, (b) $\hat{q}, \hat{d}^q$ are correct, or (c) $\hat{\mu}, \hat{q}$ are correct, EOPPG is still consistent. The reason the estimator is not consistent when only $\hat{d}^q, \hat{d}^\mu$ are correct is because $\hat{d}^v$ is constructed using both $\hat{q}, \hat{d}^q$ (see Eq. (6)). Commonly double robustness refers to a situation with two nuisances where an estimator is consistent as long as either nuisance estimate is consistent (Rotnitzky & Vansteelandt, 2014). In doubly robust OPE in MDPs, these nuisances are μ and q (Kallus & Uehara, 2019b). Here, for policy gradient estimation, we have *four* nuisances and we have a new 3-way double robustness wherein there are three pairs of nuisances where only one pair need be consistently estimated to make the final estimator consistent.

3.2. Special Cases

Example 1 (On-policy case). *If* $\pi^b = \pi^\theta$, *then*

$$\xi_{\mathrm{NMDP}} = \textstyle\sum_{j=0}^H ((\sum_{i=j}^H r_i + v_{i+1} - q_i) g_j + d_j^v - d_j^q),$$
$$\xi_{\mathrm{MDP}} = \textstyle\sum_{j=0}^H (d_j^\mu (r_j - q_j) - d_j^q + d_j^v + d_{j-1}^\mu v_j),$$

where $d_j^\mu = \mathbb{E}[\sum_{i=0}^j g_i(a_i \mid s_i) \mid a_j, s_j]$. *(Recall that* q_j, d_j^q *are functions of* $\mathcal{H}_{a_j}$ *in NMDP but only of* (s_j, a_j) *in MDP; similarly for* v_j, d_j^v *and* $\mathcal{H}_{s_j}$ *compared to just* s_j.*)*

In the on-policy case, Cheng et al. (2019); Huang & Jiang (2019) propose estimators equivalent to estimating q, d^q *and plugging into the above equation for* ξ_{NMDP}. *Using our results (establishing the efficiency bound and that* ξ_{NMDP} *is the EIF under NMDP) these estimators can then be shown to be efficient for NMDP (either under a Donsker condition or using cross-fitting instead of their in-sample estimation). These are not efficient under MDP, however, and* ξ_{MDP} *will still have lower variance. However, in the on-policy case,* $C_1 = 1$, *so the curse of horizon does not affect* ξ_{NMDP} *and since it requires fewer nuisances it might be preferable.*

Example 2 (Logged bandit case). *If* $H = 0$ *(one decision point), then* $\xi_{\mathrm{MDP}} = \xi_{\mathrm{NMDP}}$ *are both equal to*

$$\tilde{\nu}_0 (r_0 - q_0) g_0 + \mathbb{E}_{\pi_0^\theta(a_0|s_0)}[q_0 g_0 \mid s_0].$$

We can construct an estimator by cross-fold estimation of q_0 *(note the last expectation is just an integral/sum with respect to the measure* $\pi^\theta(a_0 \mid s_0)$ *for a given* s_0*). While policy gradients are used in the logged bandit case in the counterfactual learning community (e.g. Swaminathan & Joachims, 2015, which use the gradient* $\tilde{\nu}_0 r_0 g_0$*), as far as we know, no one uses this efficient estimator for the gradient even in the logged bandit case, where NMDP and MDP are the same.*

Example 3. *By Theorem 8, each of the following is a new policy gradient estimator that is consistent given consistent estimates of its respective nuisances:*
a) $\hat{\mu} = 0, \hat{d}^\mu = 0$*:* $\mathbb{E}_n[\hat{d}_0^v]$,
b) $\hat{q} = 0, \hat{d}^q = 0$*:* $\mathbb{E}_n[\sum_{j=0}^H \hat{d}_j^\mu r_j]$,
c) $\hat{d}^q = 0$, $\hat{d}^\mu = 0$*:* $\mathbb{E}_n[\sum_{j=0}^H \mathbb{E}_{\pi^\theta}[\hat{\mu}_{j-1}\hat{q}_j g_j \mid s_j]]$,
where the inner expectation is only over $a_j \sim \pi^\theta(a_j \mid s_j)$.

Example 4 (Stationary infinite-horizon case). *Suppose the MDP transition and reward probabilities and the behavior and target policy (*π^θ*) are all stationary (i.e., time invariant so that* $\pi = \pi_t$, $g = g_t$, $p_t = p$, etc.*). Suppose moreover that, as* $H \to \infty$ *the Markov chain on the variables* $\{(s_t, a_t, r_t) : t = 0, 1, \dots\}$ *is ergodic under either the behavior or target policy. Consider estimating the derivative of the long-run average reward* $Z^\infty(\theta) = \lim_{H\to\infty} Z(\theta)/H$. *By taking the limit of* ξ_{MDP}/H *as* $H \to \infty$, *we obtain*

$$\xi_{\mathrm{MDP}}^\infty \overset{dist}{=} d^\mu(s', a')(r' - q(s', a')) - \mu(s', a') d^q(s', a') + \mu(s, a) d^v(s') + d^\mu(s, a) v(s'),$$

where (s, a, r, s', a') *are distributed as the stationary distribution of* $(s_t, a_t, r_t, s_{t+1}, a_{t+1})$ *under the behavior policy,* $\mu(s, a)$ *is the ratio of stationary distributions of* (s_t, a_t) *under the target and behavior policies,* $q(s, a)$ *and* $v(s)$ *are the long-run average* q*- and* v*-functions under the target policy, and* d^μ, d^q, d^v *are the derivatives with respect to* θ.

It can be shown that under appropriate conditions, $\xi_{\mathrm{MDP}}^\infty$ *is in fact the EIF for* $Z^\infty(\theta)$ *if our data were iid observations of* (s, a, r, s', a') *from the stationary distribution under the behavior policy. If our data consists, as it does, of* n *observations of* $(H+1)$*-long trajectories, then we can instead construct the estimator*

$$\tfrac{1}{n(H+1)} \textstyle\sum_{i=1}^n \sum_{j=0}^H \big(d^\mu(s_j^{(i)}, a_j^{(i)})(r_j^{(i)} - q(s_j^{(i)}, a_j^{(i)})) - \mu(s_j^{(i)}, a_j^{(i)}) d^q(s_j^{(i)}, a_j^{(i)}) + \mu(s_{j-1}^{(i)}, a_{j-1}^{(i)}) d^v(s_j^{(i)}) + d^\mu(s_{j-1}^{(i)}, a_{j-1}^{(i)}) v(s_j^{(i)})\big),$$

where the nuisances μ, d^μ, q, d^q *are appropriately estimated in a cross-fold manner as in Algorithm 1. Following similar arguments as in Kallus & Uehara (2019a), which study infinite-horizon OPE, one can show that this extension of EOPPG maintains its efficiency and 3-way robustness guarantees as long as our data satisfies appropriate mixing conditions (which ensures it appropriately approximates observing draws from the stationary distribution). Fleshing out these details is beyond the scope of this paper.*

3.3. Estimation of Nuisance Functions

We next explain how to estimate the nuisances d_j^μ and d_j^q. The estimation of q_j is covered by Chen & Jiang (2019); Munos & Szepesvári (2008) and the estimation of μ_j by Kallus & Uehara (2019b); Xie et al. (2019). In this section, we focus on the case where the behavior policy (and thus $\tilde{\nu}_t$) is known. When the behavior policy is unknown, each method can be adapted by estimating the behavior policy first and then plugging it in. The error from estimating the behavior policy will be additive and hence might not deteriorate the rates of the below methods unless it is strictly slower.

Monte-Carlo approach. First we explain a Monte-Carlo way to estimate d_j^q, d_j^μ. We use the following theorem.

Theorem 9 (Monte Carlo representations of d_j^μ, d_j^q)**.**

$$d_j^q = \mathbb{E}\left[\sum_{t=j+1}^H r_t \nu_{j+1:t} \sum_{\ell=j+1}^t g_\ell \mid a_j, s_j\right],$$
$$d_j^\mu = \mathbb{E}\left[\nu_{0:j} \sum_{\ell=0}^j g_\ell \mid a_j, s_j\right].$$

Based on this result, we can simply learn d_j^q, d_j^μ using any regression algorithm. Specifically, we construct the response variables $y_{d_j^q}^{(i)} = \sum_{t=j+1}^H r_t^{(i)} \nu_{j+1:t}^{(i)} \sum_{\ell=j+1}^t g_\ell^{(i)}$, $y_{d_j^\mu}^{(i)} = \nu_{0:j}^{(i)} \sum_{\ell=0}^j g_\ell^{(i)}$, and we regress these on $(a_j^{(i)}, s_j^{(i)})$. For example, we can use empirical risk minimization:

$$\hat{d}_j^q = \arg\min_{f\in\mathcal{F}} \tfrac{1}{n}\sum_{i=1}^n \left(y_{d_j^q}^{(i)} - f(a_j, s_j)\right)^2,$$
$$\hat{d}_j^\mu = \arg\min_{f\in\mathcal{F}} \tfrac{1}{n}\sum_{i=1}^n \left(y_{d_j^\mu}^{(i)} - f(a_j, s_j)\right)^2.$$

Examples for $\mathcal{F}$ include RKHS norm balls, an expanding subspace of L_2 (*i.e.*, a sieve), and neural networks. Rates for such estimators can, for example, be derived from Bartlett et al. (2005); Wainwright (2019).

A careful reader might wonder whether estimating nuisances in this way causes the final EOPPG estimator to suffer from the curse of horizon, since $\nu_{0:j}$ can be exponentially growing in j. However, as long as we have suitable nonparametric rates (in n) for the nuisances as in Theorem 7, the asymptotic MSE of $\hat{Z}^{\mathrm{EOPPG}}(\theta)$ does *not* depend on the estimation error of the nuisances. These errors only appear in higher-order (in n) terms and therefore vanish. This is still an important concern in finite samples, which is why we next present an alternative nuisance estimation approach.

Recursive approach. Next, we explain a recursive way to estimate d_j^q, d_j^μ. This relies on the following result.

Theorem 10 (Bellman equations of d_j^q, d_j^μ)**.**

$$d_j^q(s_j, a_j) = \mathbb{E}[d_{j+1}^v \mid s_j, a_j],\ d_j^v(s_j) = \mathbb{E}_{\pi^\theta}[d_j^q + g_j q_j \mid s_j]$$
$$d_j^\mu(s_j, a_j) = \mathbb{E}[d_{j-1}^\mu \tilde{\nu}_j \mid s_j, a_j] + \mu_j g_j.$$

Algorithm 2 Estimation of d_j^q (Recursive way)

Input: q-estimates $\hat{q}_j$, hypothesis classes $\mathcal{F}^{d_j^q}$
Set $\hat{d}_H^v = \hat{d}_H^q = 0$
for $j = H-1, H-2, \cdots$ **do**
  Set $\hat{d}_j^q \in \arg\min_{f\in\mathcal{F}^{d_j^q}} \sum_{i=1}^n \left(\hat{d}_{j+1}^v(s_{j+1}^{(i)}) - f(s_j^{(i)}, a_j^{(i)})\right)^2$
  Set $\hat{d}_j^v(s_j) = \mathbb{E}_{\pi_j^\theta}[\hat{d}_j^q + \hat{q}_j g_j \mid s_j]$
  (by integrating/summing w.r.t $a_j \sim \pi_j^\theta(a_j \mid s_j)$)
end for

Algorithm 3 Estimation of d_j^μ (Recursive way)

Input: μ-estimates $\hat{\mu}_j$, hypothesis classes $\mathcal{F}^{d_j^\mu}$
Set $\hat{d}_0^\mu = \nu_0 g_0$
for $j = 1, 2, \cdots$ **do**
  Set $\hat{d}_j^\mu = \arg\min_{f\in\mathcal{F}^{d_j^\mu}} \sum_{i=1}^n \big(f(s_j^{(i)}, a_j^{(i)}) - \tilde{\nu}_j^{(i)} \hat{d}_{j-1}^\mu(s_{j-1}^{(i)}, a_{j-1}^{(i)}) - \hat{\mu}_j^{(i)} g_j^{(i)}\big)^2$
end for

Algorithm 4 Off-policy projected gradient ascent

Input: An initial point $\theta_1 \in \Theta$ and step size schedule α_t
for $t = 1, 2, \cdots$ **do**
  $\tilde{\theta}_{t+1} = \theta_t + \alpha_t \hat{Z}^{\mathrm{EOPPG}}(\theta_t)$
  $\theta_{t+1} = \mathrm{Proj}_\Theta(\tilde{\theta}_{t+1})$
end for

This is derived by differentiating the Bellman equations:

$$q_j(s_j, a_j) = \mathbb{E}[r + v_{j+1}(s_{j+1}) \mid s_j, a_j],$$
$$\mu_j(s_j, a_j) = \mathbb{E}[\mu_{j-1}(s_{j-1}, a_{j-1})\tilde{\nu}_j | s_j, a_j].$$

Following Theorem 10, we propose the recursive Algorithms 2 and 3 that estimate d_j^q using backwards recursion and d_j^μ using forward recursion.

Remark 1. *Morimura et al. (2010) discussed a way to estimate the gradient of the stationary distribution in an on-policy setting. In comparison, our setting is off-policy.*

Remark 2. *We have directly estimated d_j^μ. Another way is using $d_j^\mu = \tilde{\nu}_j \nabla_\theta \tilde{\mu}_j + \tilde{\mu}_j g_j$ and estimating $\nabla_\theta \tilde{\mu}_j$ recursively based on a Bellman equation for $\nabla \tilde{\mu}_j$, derived in a similar way to that for d_j^μ in Theorem 10.*

4. Off-policy Optimization with EOPPG

Next, we discuss how to use the EOPPG estimator presented in Section 3 for off-policy optimization using projected gradient ascent and the resulting guarantees. The algorithm is given in Algorithm 4.

Then, by defining an error $B_t = \hat{Z}^{\mathrm{EOPPG}}(\theta_t) - Z(\theta_t)$, we have the following theorem.

Theorem 11. *Assume the function $J(\theta)$ is differentiable*

and M-smooth in θ, $M < 1/(4\alpha_t)$, and $\tilde{\theta}_t = \theta_t$.[1] *Set $J^* = \max_{\theta\in\Theta} J(\theta)$. Then, $\{\theta_t\}_{t=1}^T$ in Algorithm 4 satisfies*

$$\frac{1}{T}\sum_{t=1}^T \alpha_t \|Z(\theta_t)\|_2^2 \le \frac{4(J^*-J(\theta_1))}{T} + \frac{3}{T}\sum_{t=1}^T \alpha_t \|B_t\|_2^2.$$

Theorem 11 gives a bound on the average derivative. That is, if we let $\hat{\theta}$ be chosen at random from $\{\theta_t\}_{t=1}^T$ with weights α_t, then via Markov's inequality,

$$Z(\hat{\theta}) = \mathcal{O}_p(\tfrac{4}{T}(J^* - J(\theta_1)) + \tfrac{3}{T}\textstyle\sum_{t=1}^T \alpha_t \|B_t\|_2^2).$$

So as long as we can bound the error term $\sum_t \alpha_t \|B_t\|_2^2/T$, we have that $\hat{\theta}$ becomes a near-stationary point.

This error term comes from the noise of the EOPPG estimator. A heuristic calculation based on Theorem 7 that ignores the fact that θ_t is actually random would suggest

$$\begin{aligned}\|B_t\|_2^2 &\lessapprox \mathrm{trace}(\mathrm{var}[\xi_{\mathrm{MDP}}]) + \mathrm{o}_p(1/n)\\ &\lessapprox \frac{DC_2 R_{\max}^2 G_{\max}(H+1)^2(H+2)^2}{n} + \mathrm{o}_p(1/n).\end{aligned}$$

To establish this formally, we recognize that θ_t is a random variable and bound the *uniform* deviation of EOPPG over all $\theta \in \Theta$. We then obtain the following high probability bound on the cumulative errors.

Theorem 12 (Bound for cumulative errors)**.** *Suppose the assumptions of Theorem 7 hold, that $\theta \to \xi_{\mathrm{MDP},j}$ is almost surely differentiable with derivatives bounded by L for $j \in \{1,\cdots,D\}$, where $\xi_{\mathrm{MDP},j}$ is a j-th component of ξ_{MDP}, and that Θ is compact with diameter Υ.*

Then, for any δ, there exists N_δ such that $\forall n \ge N_\delta$, we have that, with probability at least $1-\delta$,

$$\frac{1}{T}\textstyle\sum_t \|B_t\|_2^2 \lessapprox U_{n,T,\delta},$$
$$U_{n,T,\delta} = \frac{D(L^2 D\Upsilon^2 + C_2 G_{\max} R_{\max}^2 (H+1)^2(H+2)^2 \log(TD/\delta))}{n}.$$

This shows that, by letting $T = n^\beta$ ($\beta > 1$) be sufficiently large, we can obtain $Z(\hat{\theta}) = \mathcal{O}_p(H^4 C_2 \log(n)/n)$ for $\hat{\theta}$ chosen at random from $\{\theta_t\}_{t=1}^T$ as above. Note that if we had used other policy gradient estimators such as (step-wise) REINFORCE, PG as in Eq. (5), or off-policy variants of the estimators of Cheng et al. (2019); Huang & Jiang (2019), then the term C_1^H would have appeared in the bound and the curse of horizon would have meant that our learned policies would not be near-stationary for long-horizon problems.

Remark 3. *Many much more sophisticated gradient-based optimization methods equipped with our EOPPG gradient estimator can be used in place of the vanilla projected gradient ascent in Algorithm 4. We refer the reader to Jain & Kar (2017) for a review of non-convex optimization methods.*

[1] This means all iterates remain in Θ so the projection is identity. This is a standard condition in the analysis of non-convex optimization method that can be guaranteed under certain assumptions; see Khamaru & Wainwright (2018); Nesterov & Polyak (2006).

The concave case. The previous results study the guarantees of Algorithm 4 in terms of convergence to a stationary point, which is the standard form of analysis for non-convex optimization. If we additionally assume that $J(\theta)$ is a *concave* function then we can see how the efficiency of EOPPG translates to convergence to an optimal solution in terms of the *regret* compared to the optimal policy. In this case we set $\hat{\theta} = \frac{1}{T}\sum_{t=1}^T \theta_t$, for which we can prove the following:

Theorem 13 (Regret bound)**.** *Suppose the assumptions of Theorem 12 hold, that $J(\theta)$ is a concave function, and that Θ is convex. For a suitable choice of α_t we have that, for any δ, there exists N_δ such that $\forall n \ge N_\delta$, we have that, with probability at least $1-\delta$,*

$$J^* - J(\hat{\theta}) \lessapprox \Upsilon \frac{\sup_{\theta\in\Theta}\|Z(\theta)\|_2 + \sqrt{U_{n,T,\delta}}}{\sqrt{T}}.$$

Here, the first term is the optimization error if we knew the true gradient $Z(\theta)$. The second term is the approximation error due to the error in our estimated gradient $\hat{Z}^{\mathrm{EOPPG}}(\theta)$. When we set $T = n^\beta$ ($\beta > 1$), the final regret bound is

$$\mathcal{O}_p\left(\Upsilon R_{\max} H^2 \sqrt{D\beta G_{\max} C_2 \log(nD/\delta))}/\sqrt{n}\right).$$

The regret's horizon dependence is H^2. This is a crucial result since the regret with polyomial horizon dependence is a desired result in RL (Jiang & Agarwal, 2018). Again, if we had used other policy gradient methods, then an exponential dependence via C_1^H would appear. Moreover, the regret depends on C_2, which corresponds to a concentrability coefficient (Antos et al., 2008).

Remark 4. *Recent work studies the global convergence of online-PG algorithms without concavity (Agarwal et al., 2019; Bhandari & Russo, 2019). This may be applicable here, but our setting is completely off-policy and therefore different and requiring future work. Notably, the above focus on optimization rather than PG variance reduction. In a truly off-policy setting, the available data is limited and statistical efficiency is crucial and is our focus here.*

Remark 5 (Comparison with other results for off-policy policy learning)**.** *In the logged bandit case ($H = 0$), the regret bound of off-policy learning via exhaustive search (non-convex) optimization is $\mathcal{O}_p(\sqrt{\tau(\Pi)\log(1/\delta)/n})$, where $\tau(\Pi)$ represents the complexity of the hypothesis class (Foster & Syrgkanis, 2019; Zhou et al., 2018). In this bandit case, the nuisance functions of the EIF do not depend on the policy itself, making this analysis tractable. However, for our RL problem ($H > 0$), nuisance functions depend on the policy; thus, these techniques do not extend directly. Nie et al. (2019) do extend these types of regret results to an RL problem but where the problem has a special when-to-treat structure, not the general MDP case.*

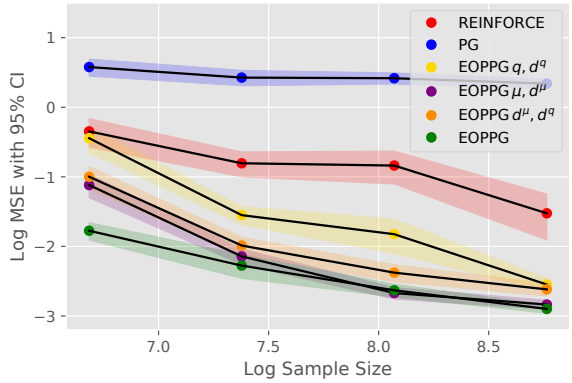

Figure 2. Comparison of MSE of gradient estimation

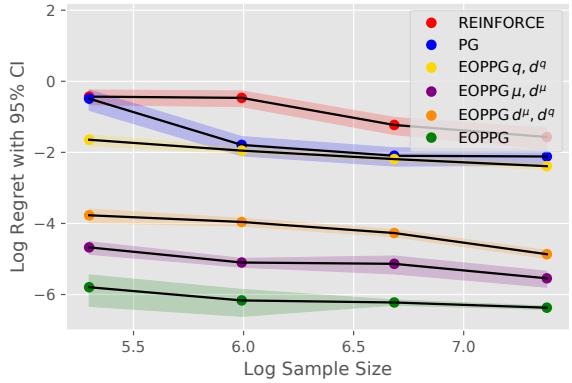

Figure 3. Comparison of regret after gradient ascent

5. Experiments

We conducted an experiment in a simple environment to confirm the theoretical guarantees of the proposed estimator. The setting is as follows. Set $\mathcal{S}_t = \mathbb{R}$, $\mathcal{A}_t = \mathbb{R}$, $s_0 = 0$. Then, set the transition dynamics as $s_t = a_{t-1} - s_{t-1}$, the reward as $r_t = -s_t^2$, the behavior policy as $\pi_t^b(a \mid s) = \mathcal{N}(0.8s, 0.2^2)$, the policy class as $\mathcal{N}(\theta s, 0.2^2)$, and horizon as $H = 49$. Then, $\theta^* = 1$ with optimal value $J^* = -1.96$, obtained by analytical calculation. We compare REINFORCE (Eq. (4)), PG (Eq. (5)), and EOPPG with $K = 2$. Nuisances functions q, μ, d^q, d^μ are estimated by polynomial sieve regressions (Chen, 2007). Additionally, to investigate 3-way double robustness, we consider corrupting the nuisance models by adding noise $\mathcal{N}(0, 1)$; we consider thus corrupting (q, d^q), (μ, d^μ), or (d^μ, d^q).

First, in Fig. 2, we compare the MSE of these gradient estimators at $\theta = 1.0$ using 100 replications of the experiment for each of $n = 800, 1600, 3200, 6400$. As can be seen, the performance of EOPPG is superior to existing estimators in terms of MSE, validating our efficiency results. We can also see that the EOPPG with misspecified models still converges, validating our 3-way double robustness results.

Second, in Fig. 3, we apply a gradient ascent as in Algorithm 4 with $\alpha_t = 0.15$ and $T = 40$. We compare the regret of the final policy, *i.e.*, $J(\theta^*) - J(\hat{\theta}_{40})$, using 60 replications of the experiment for each of $n = 200, 400, 800, 1600$. Notice that the lines decrease roughly as $1/\sqrt{n}$ but because of the large differences in values, the lines only appear somewhat flat. This shows that the efficiency and 3-way double robustness translate to good regret performance, as predicted by our policy learning analysis.

6. Conclusions

We established an MSE efficiency bound of order $\mathcal{O}(H^4/n)$ for estimating a policy gradient in an MDP in an off-policy manner. We proposed an estimator, EOPPG, that achieves the bound, enjoys 3-way double robustness, and leads to regret dependence of order $H^2/\sqrt{n}$ when used for policy learning. Notably, this is much smaller than other approaches, which incur exponential-in-H errors. This paper is only a first step toward efficient and effective off-policy policy gradients in MDPs. Remaining questions include how to estimate d^q, d^μ in a large-scale environments, the performance of more practical implementations that alternate in updating θ and nuisance estimates with only one gradient update, and extending our theory to the deterministic policy class as in Silver et al. (2014).

Acknowledgements

We thank the anonymous reviewers for their insightful comments and suggestions.

This material is based upon work supported by the National Science Foundation under Grant No. 1846210. Masatoshi Uehara was partially supported by the Masason Foundation.

References

Agarwal, A., Kakade, S., Lee, J., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. *arXiv preprint arXiv: 1908.00261*, 2019.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

Athey, S. and Wager, S. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *Annals of Statistics*, 33:1497–1537, 2005.

Baxter, J. and Bartlett, P. L. Infinite-Horizon Policy-

Gradient Estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore, 1993.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 1042–1051, 2019.

Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., and Chi, E. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on web search and data mining*, WSDM '19, pp. 456–464, 2019.

Chen, X. Chapter 76 large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6: 5549–5632, 2007.

Cheng, C.-A., Yan, X., and Boots, B. Trajectory-wise control variates for variance reduction in policy gradient methods. *arXiv preprint arxiv:1908.03263*, 2019.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68, 2018.

Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

Degris, T., White, M., and Sutton, R. Off-policy actor-critic. *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 179–186, 2012.

Deisenroth, M., Neumann, G., and Peters, J. A survey on policy search for robotics. *Foundations and Trends in Robotics - volume 1*, 2:1–142, 2013.

Foster, D. J. and Syrgkanis, V. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2052–2062, 2019.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. Guidelines for reinforcement learning in healthcare. *Nat Med*, 25: 16–18, 2019.

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

Gu, S. S., Lillicrap, T., Turner, R. E., Ghahramani, Z., Schölkopf, B., and Levine, S. Interpolated policy gradient: Merging on-policy and off-policy gradient estimation for deep reinforcement learning. In *Advances in Neural Information Processing Systems 30*, pp. 3846–3855. 2017.

Hájek, J. A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14:323–330, 1970.

Hanna, J. and Stone, P. Towards a data efficient off-policy policy gradient. In *AAAI Spring Symposium on Data Efficient Reinforcement Learning*, 2018.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2:157–325, 2015.

Huang, J. and Jiang, N. From importance sampling to doubly robust policy gradient. *arXiv preprint arXiv:1910.09066*, 2019.

Imani, E., Graves, E., and White, M. An off-policy policy gradient theorem using emphatic weightings. In *Advances in Neural Information Processing Systems 31*, pp. 96–106. 2018.

Jain, P. and Kar, P. Non-convex optimization for machine learning. *Found. Trends Mach. Learn.*, 10:142–336, December 2017.

Jiang, N. and Agarwal, A. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pp. 3395–3398, 2018.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. *In Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume*, pp. 652–661, 2016.

Kallus, N. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pp. 8895–8906, 2018.

Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*, 2019a.

Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *arXiv preprint arXiv:1908.08526*, 2019b.

Kallus, N. and Zhou, A. Confounding-robust policy improvement. In *Advances in neural information processing systems*, pp. 9269–9279, 2018.

Khamaru, K. and Wainwright, M. Convergence guarantees for a class of non-convex and non-smooth optimization problems. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2601–2610, 2018.

Klaassen, C. A. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, pp. 1548–1562, 1987.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5361–5371, 2018.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Off-policy policy gradient with state distribution correction. *In Proccedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2019)*, 2019.

Metelli, A. M., Papini, M., Faccio, F., and Restelli, M. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems 31*, pp. 5442–5454. 2018.

Morimura, T., Uchibe, E., Yoshimoto, J., Peters, J., and Doya, K. Derivatives of logarithmic stationary distributions for policy gradient reinforcement learning. *Neural Computation*, 22:342–376, 2010.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.

Nesterov, Y. and Polyak, B. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108:177–205, 2006.

Nie, X., Brunskill, E., and Wager, S. Learning when-to-treat policies. *arXiv preprint arXiv:1905.09751*, 2019.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. Stochastic variance-reduced policy gradient. In *International Conference on Machine Learning*, pp. 4026–4035, 2018.

Precup, D., Sutton, R. S., and Singh, S. P. Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 759–766, 2000.

Rotnitzky, A. and Vansteelandt, S. Double-robust methods. In *Handbook of missing data methodology. In Handbooks of Modern Statistical Methods*, pp. 185–212. Chapman and Hall/CRC, 2014.

Schulman, J., Heess, N., Weber, T., and Abbeel, P. Gradient estimation using stochastic computation graphs. In *Neural Information Processing Systems (NIPS), Montreal, Canada*, 2015.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Sen, B. A gentle introduction to empirical process theoryand applications. http://www.stat.columbia.edu/~bodhi/Talks/Emp-Proc-Lecture-Notes.pdf, 2018. Accessed: 2020–1-1.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 387–395, 2014.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Sutton, R. S., Precup, D., and Singh, S. P. Intra-option Learning about Temporally Abstract Actions. In *Proceedings of the 15th International Conference on Machine Learning*, pp. 556–564, 1998.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pp. 814–823, 2015.

Tang, J. and Abbeel, P. On a connection between importance sampling and the likelihood ratio policy gradient. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, pp. 1000–1008, 2010.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. *In Proceedings of the 33rd International Conference on Machine Learning*, pp. 2139–2148, 2016.

Tosatto, S., Carvalho, J., Abdulsamad, H., and Peters, J. A nonparametric offpolicy policy gradient. *arXiv preprint arXiv:2001.02435*, 2020.

Tsiatis, A. A. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, New York, NY, 2006.

van Der Laan, M. J. and Robins, J. M. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics,. Springer New York, New York, NY, 2003.

van der Laan, M. J. and Rose, S. *Targeted Learning :Causal Inference for Observational and Experimental Data*. Springer Series in Statistics. Springer New York : Imprint: Springer, New York, NY, 2018.

van der Vaart, A. W. *Asymptotic statistics*. Cambridge University Press, Cambridge, UK, 1998.

Wainwright, M. J. *High-Dimensional Statistics : A Non-Asymptotic Viewpoint*. Cambridge University Press, New York, 2019.

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., Kakade, S., Mordatch, I., and Abbeel, P. Variance reduction for policy gradient with action-dependent factorized baselines. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Xie, T., Ma, Y., and Wang, Y.-X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems 32*, pp. 9665–9675. 2019.

Zhou, Z., Athey, S., and Wager, S. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*, 2018.