On the Convergence Rate of Density-Ratio Based Off-Policy Policy Gradient Methods

Jiawei Huang
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
jiaweih@illinois.edu

Nan Jiang
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
nanjiang@illinois.edu

Abstract

We study the convergence properties of two optimization algorithms for off-policy policy gradient based on density-ratio learning. We establish general conditions that enable convergence and near-optimality guarantees, and show that these conditions can be satisfied in the linear case under standard assumptions. The keys to our analyses are the successful integration and application of stochastic first-order methods on solving saddle-point and non-convex optimization problems.

1 Introduction

Policy gradient (PG) is a very popular class of methods in empirical reinforcement-learning (RL) research, and has also attracted significant attention from the theoretical community recently [1]. Despite its appealing properties, classical PG typically requires on-policy roll-outs, making them not directly applicable to offline (or batch) RL. Recent development in marginalized importance sampling (MIS) methods [2, 3, 4, 5], however, has yielded promising off-policy policy-gradient estimators. For example, Nachum et al. [6] reformulated off-policy policy-optimization to a max-max-min problem, which faithfully optimizes the policy with sufficiently expressive function approximators [7]. A more general form of the problem considered by Yang et al. [5] is:

$$\begin{aligned}\max_{\pi\in\Pi}\max_{w\in\mathcal{W}}\min_{Q\in\mathcal{Q}}\mathcal{L}(\pi,w,Q) :=& \max_{\theta\in\Theta}\max_{\zeta\in Z}\min_{\xi\in\Xi}\mathcal{L}(\pi_\theta,w_\zeta,Q_\xi)\\ :=&(1-\gamma)\mathbb{E}_{s_0\sim\nu_0}[Q_\xi(s_0,\pi_\theta)]+\mathbb{E}_{d^\mu}[w_\zeta(s,a)\Big(r+\gamma Q_\xi(s',\pi_\theta)-Q_\xi(s,a)\Big)]\\ &+\lambda_Q\mathbb{E}_{d^\mu}[f(Q_\xi(s,a))]-\lambda_w\mathbb{E}_{d^\mu}[g(w_\zeta(s,a))] \qquad (1)\end{aligned}$$

where π, w, Q are respectively parameterized by $(\theta,\zeta,\xi)\in\Theta\times Z\times\Xi$ (Θ, Z and Ξ are all convex sets), and we use $\Pi,\mathcal{W},\mathcal{Q}$ to denote their function classes; ν_0 is the initial state distribution, d^μ denotes the normalized discounted state-action occupancy induced by behavior policy μ (see Sec. 2.1 for a formal definition); $Q_\xi(s,\pi_\theta)$ is short for $\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[Q_\xi(s,a)]$; f, g are regularizers.

Despite the promising formulation, the problem takes a complex max-max-min form, which makes the optimization challenging. In this paper, we study the convergence guarantees of two natural optimization strategies for (the empirical version of) Eq.(2), and establish the conditions under which we can prove convergence rate and characterize the quality of the solutions. The actual objective, based on a sample D from d^μ, is

$$\begin{aligned}\max_{\pi\in\Pi}\max_{w\in\mathcal{W}}\min_{Q\in\mathcal{Q}}\mathcal{L}^D(\pi,w,Q) :=& \max_{\theta\in\Theta}\max_{\zeta\in Z}\min_{\xi\in\Xi}\mathcal{L}^D(\pi_\theta,w_\zeta,Q_\xi)\\ :=&(1-\gamma)\mathbb{E}_{s_0\sim\nu_D}[Q_\xi(s_0,\pi_\theta)]+\mathbb{E}_{d^D}[w_\zeta(s,a)\Big(r+\gamma Q_\xi(s',\pi_\theta)-Q_\xi(s,a)\Big)]\\ &+\frac{\lambda_Q}{2}\mathbb{E}_{d^D}[Q_\xi^2(s,a)]-\frac{\lambda_w}{2}\mathbb{E}_{d^D}[w_\zeta^2(s,a)]. \qquad (2)\end{aligned}$$

Offline Reinforcement Learning Workshop at Neural Information Processing Systems, 2020.

Here we replace ν_0 with ν_D to denote the empirical initial distribution, and use d^D to denote the empirical state-action distribution in dataset. We also choose the regularizers to be quadratic functions.

In our analyses, we focus on the case when $\mathcal{L}^D$ is strongly-concave w.r.t. ζ and strongly-convex w.r.t. ξ, but do not require the concavity related to θ. The strong concavity/convexity, among other assumptions we will introduce in Section 2.2, can be shown to be satisfied in the linear case under very standard assumptions (Appendix F).

Due to regularization, generalization error, and mis-specification error, there is inevitable bias between the stationary points of $\mathcal{L}^D(\pi_\theta, w_\zeta, Q_\xi)$ and $J(\pi_\theta)$, respectively, where $J(\pi_\theta)$ is the expected return of π_θ. Therefore, we focus on the convergence to the biased stationary point defined below.

Definition 1.1 (Biased stationary point)**.**

$$\mathbb{E}[\|\nabla_\theta J(\pi_\theta)\|] \leq \varepsilon + \varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg} \tag{3}$$

where $\varepsilon_{reg}, \varepsilon_{func}, \varepsilon_{data}$ are biases caused by regularization, mis-specified function class, and finite-sample effects, respectively, as we will explain in Section 2. All norms in this paper is ℓ_2 norm unless specified otherwise. The expectation is over the randomness of the algorithm (e.g., the randomness in SGD) and not that of the data.

Paper Outline Our first algorithm, converts the original max-max-min problem to a max-min problem $\max_{(\theta,\zeta)\in\Theta\times Z} \min_{\xi\in\Xi} \mathcal{L}(\pi_\theta, w_\zeta, Q_\xi)$, by simultaneously optimizing θ and ζ. Under the assumptions identified in Section 2.2, we prove that the stationary point returned by any stochastic optimization algorithm for non-convex-strongly-concave problems is also a biased stationary point in Definition 1.1. As a result, the $O(\varepsilon^{-3})$ convergence rate can be established based on a recent result on non-convex-strongly-concave optimization [8].

We then study another algorithm, where we iteratively solve the inner strongly-concave-strongly-convex max-min problem $\max_{\zeta\in Z} \min_{\xi\in\Xi} \mathcal{L}(\pi_\theta, w_\zeta, Q_\xi)$ for fixed θ and the outer non-convex optimization problem $\max_{\theta\in\Theta} \mathcal{L}(\pi_\theta, w_\zeta, Q_\xi)$ for fixed ζ and ξ. For the inner loop, we assume an oracle that solves the saddle-point problem, and provide a concrete example in Appendix E. For the outer loop, the main technique difficulty is that, the loss function $\mathcal{L}(\pi_\theta, w_{\zeta_t}, Q_{\xi_t})$ varies across iterations because we update ζ_t, ξ_t in the inner loop, which prevents us from adapting existing non-convex optimization algorithms directly. We resolve this difficulty by coordinating the inner and the outer loops so that we can relate the variation $\|\zeta_{t+1} - \zeta_t\|$ and $\|\xi_{t+1} - \xi_t\|$ with $\|\theta_{t+1} - \theta_t\|$. The convergence rate to a biased stationary point of our second strategy is $O(\varepsilon^{-4})$.

1.1 Related works

Recently, there has been a lot of interest in turning MIS methods for off-policy evaluation [3, 9, 2] into off-policy policy-optimization algorithms. Liu et al. [10] presented OPPOSD with convergence guarantees, but the convergence relies on accurately estimating the density ratio and the value function via MIS, which were treated as a black box without further analysis. [6, 7] discussed policy optimization given arbitrary off-policy dataset, but no convergence analysis was performed. Another style of off-policy policy-improvement algorithms is off-policy actor-critic [11, 12, 13]. Although [13] presented a provably convergent algorithm, where only asymptotic convergence was proved and no finite convergence rate was given.

Meanwhile, along with the progress of the variance reduction techniques for non-convex optimization, there are several emerging works analyzing convergence rates in RL settings [14, 15, 16, 17, 18]. However, all of them require on-policy interaction with the environment, whereas our focus is the off-policy setting.

2 Preliminary

2.1 Markov Decision Process

We consider an infinite-horizon discounted MDP $(\mathcal{S}, \mathcal{A}, R, P, \gamma, \nu_0)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively, which we assume to be finite but can be arbitrarily large. $R : \mathcal{S} \times \mathcal{A} \to \Delta([0, 1])$ is the reward function. $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, γ is the discount factor and ν_0 denotes the initial state distribution.

For arbitrary policy π, we use $d^\pi(s,a) = (1-\gamma)\mathbb{E}_{\tau\sim\pi, s_0\sim\nu_0}[\sum_{t=0}^\infty \gamma^t p(s_t = s, a_t = a)]$ to denote the normalized discounted state-action occupancy, where $\tau \sim \pi, s_0 \sim \nu_0$ means a trajectory $\tau = \{s_0, a_0, s_1, a_1, ...\}$ is sampled according to the rule that $s_0 \sim \nu_0, a_0 \sim \pi(\cdot|s_0), s_1 \sim P(\cdot|s_0, a_0), a_1 \sim \pi(\cdot|s_1)$, ..., and $p(s_t = s, a_t = a)$ denotes the probability that the t-th state-action pair are exactly (s,a). We also use $Q^\pi(s,a) = \mathbb{E}_{\tau\sim\pi, s_0=s, a_0=a}[\sum_{t=0}^\infty \gamma^t r(s_t, a_t)]$ to denote the Q-function of π. It is well-known that Q^π satisfies the Bellman Equation:

$$Q^\pi(s,a) = \mathcal{T}^\pi Q^\pi(s,a) := \mathbb{E}_{r\sim R(s,a), s'\sim P(\cdot|s,a), a'\sim\pi(\cdot|s')}[r + \gamma Q^\pi(s', a')].$$

Define $J(\pi) = \mathbb{E}_{s\sim\nu_0, a\sim\pi(\cdot|s_0)}[Q^\pi(s,a)] = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^\pi}[r(s,a)]$ as the expected return of policy π. If π is parameterized by θ and differentiable, the policy-gradient theorem [19] states that

$$\nabla_\theta J(\pi_\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^\pi}[Q^\pi(s,a)\nabla_\theta \log \pi(a|s)].$$

In the off-policy setting, we can only get access to d^μ, the discounted state-action occupancy w.r.t. another policy μ. Then we can rewrite $\nabla_\theta J(\pi)$ by introducing the importance ratio $w^\pi(s,a) := \frac{d^\pi(s,a)}{d^\mu(s,a)}$.

$$\nabla_\theta J(\pi_\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^\mu}[w^\pi(s,a)Q^\pi(s,a)\nabla_\theta \log \pi(a|s)].$$

In the rest of the paper, we will refer μ as the behavior policy, and refer π as the target policy whose performance we are interested in.

In practice, usually, we are only provided with an off-line dataset instead of the exact distribution d^μ, which we denote as $D = \{(s_i, a_i, r_i, s_i')\}_{i=1}^{|D|}$. Each tuple is sampled by $s_i, a_i \sim d^\mu, r_i \sim R(s_i, a_i), s_i' \sim P(\cdot|s_i, a_i)$, and we use d^D to denote the empirical state-action distribution.

2.2 Assumptions and Definitions

We now introduce the assumptions and definitions that will later enable us to establish the convergence guarantees and characterize the solution quality. We will also introduce some algorithm-specific assumptions later. While some of the assumptions (e.g., Assumption C) are quite strong, in Appendix F we show they are automatically satisfied in the linear setting under more standard assumptions.

Assumption A (Smoothness)**.**

(a) For any $s, a \in \mathcal{S}\times\mathcal{A}$ and $\theta \in \Theta$, $\pi_\theta(s,a)$ is second-order differentiable w.r.t. θ, and there exist constants G and H, s.t.

$$\|\nabla_\theta \log \pi_\theta(a|s)\| \le G, \qquad \|\nabla_\theta^2 \log \pi_\theta(a|s)\|_{op} \le H \tag{4}$$

where $\|\cdot\|_{op}$ is the matrix operator norm.

(b) For any $\xi, \xi_1, \xi_2 \in \Xi, \zeta, \zeta_1, \zeta_2 \in Z, (s,a) \in \mathcal{S}\times\mathcal{A}$, there are constants $C_\mathcal{Q}, C_\mathcal{W}, L_Q, L_w$, s.t.

$$|Q_\xi(s,a)| \le C_\mathcal{Q}; \quad |Q_{\xi_1}(s,a) - Q_{\xi_2}(s,a)| \le L_Q\|\xi_1 - \xi_2\|;$$
$$|w_\zeta(s,a)| \le C_\mathcal{W}; \quad |w_{\zeta_1}(s,a) - w_{\zeta_2}(s,a)| \le L_w\|\zeta_1 - \zeta_2\|;$$

Usually, in practice, we normalize the expectation of w_ζ to 1, so $C_\mathcal{W} > 1$ in general.

(c) Let $v \in V = \Theta \times Z \times \Xi$ denote a vector formed by concatenating θ, ζ, ξ. For any $v, v_1, v_2 \in V$, $\mathcal{L}^D$ defined in Eq.(2) is differentiable w.r.t. v, and there exists constant L s.t.

$$\begin{aligned}
&\|\nabla_v \mathcal{L}^D(v_1) - \nabla_v \mathcal{L}^D(v_2)\| : \\
=&\|\nabla_\theta \mathcal{L}^D(v_1) - \nabla_\theta \mathcal{L}^D(v_2)\| + \|\nabla_\zeta \mathcal{L}^D(v_1) - \nabla_\zeta \mathcal{L}^D(v_2)\| + \|\nabla_\xi \mathcal{L}^D(v_1) - \nabla_\xi \mathcal{L}^D(v_2)\| \\
\le& L\|\theta_1 - \theta_2\| + L\|\zeta_1 - \zeta_2\| + L\|\xi_1 - \xi_2\|
\end{aligned}$$

Assumption B (Exploratory Data)**.** Recall the behavior policy is denoted as μ. We assume there exists a constant $C > 0$, for arbitrary $\pi \in \Pi$ and any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$w^\pi(s,a) := \frac{d^\pi(s,a)}{d^\mu(s,a)} \le C, \qquad w^\pi_{d^\mu}(s,a) := \frac{d^\pi_{d^\mu}(s,a)}{d^\mu(s,a)} \le C$$

where $d^\pi_{d^\mu}(s,a) := (1-\gamma)\mathbb{E}_{\tau\sim\pi, s_0,a_0\sim d^\pi(\cdot,\cdot)}[\sum_{t=0}^\infty \gamma^t p(s_t = s, a_t = a)]$ is the normalized discounted state-action occupancy by treating d^μ as initial distribution.

Assumption C (Strongly-Convex-Strongly-Concave)**.** We use u_Z and u_Ξ to denote the dimension of vector parameters ζ and ξ. Given arbitrary $\theta \in \Theta, \zeta \in \mathbb{R}^{u_Z}$, $\mathcal{L}^D(\theta, \zeta, \cdot)$ is μ_ξ-strongly convex w.r.t. $\xi \in \Xi$. Given arbitrary $\theta \in \Theta, \xi \in \mathbb{R}^{u_\Xi}$, $\mathcal{L}^D(\theta, \cdot, \xi)$ is μ_ζ-strongly concave w.r.t. $\zeta \in Z$.

Remark 2.1. *In fact, the regularization terms is necessary if we want Assumption C to hold when one of w^π and Q^π is realizable. We defer the discussion to Appendix B.*

Assumption D. Denote $(\zeta^*_\theta, \xi^*_\theta)$ as the saddle point of $\mathcal{L}^D(\theta, \zeta, \xi)$ without constraint on ζ and ξ. For arbitrary π_θ parameterized by $\theta \in \Theta$, $(\zeta^*_\theta, \xi^*_\theta) \in Z \times \Xi$.

Remark 2.2. *Based on Assumption A, C, since both Z and Ξ are convex sets, Assumption D implies that*

$$\|\nabla_\zeta \mathcal{L}^D(\theta, \zeta^*_\theta, \xi^*_\theta)\| = \|\nabla_\xi \mathcal{L}^D(\theta, \zeta^*_\theta, \xi^*_\theta)\| = 0$$

Definition 2.3 (Generalization Error)**.** Suppose there exists a constant ε'_{data}, for arbitrary $\pi_\theta, w_\zeta, Q_\xi \in \Pi \times \mathcal{W} \times \mathcal{Q}$, we have:

$$|\mathcal{L}(\pi_\theta, w_\zeta, Q_\xi) - \mathcal{L}^D(\pi_\theta, w_\zeta, Q_\xi)| \leq \varepsilon'_{data}$$

$$\|\nabla_\theta \mathcal{L}(\pi_\theta, w^*_\mu, Q^*_\mu) - \nabla_\theta \mathcal{L}^D(\pi_\theta, w^*_\mu, Q^*_\mu)\|^2 \leq \varepsilon'_{data}$$

where $(w^*_\mu, Q^*_\mu) := \arg\max_{w \in \mathcal{W}} \min_{Q \in \mathcal{Q}} \mathcal{L}(\pi, w, Q)$.

Proposition 2.4. *Denote $\varepsilon_{data} := (2\kappa_\zeta \kappa_\xi + 2\kappa_\zeta + 2\kappa_\xi + \sqrt{2}/2)\sqrt{2\varepsilon'_{data}}$, where ε'_{data} is defined in Definition 2.3, $\kappa_\zeta = L/\mu_\zeta$, $\kappa_\xi = L/\mu_\xi$. Under Assumption A and C, we have:*

$$\|\nabla_\theta \max_{w \in \mathcal{W}} \min_{Q \in \mathcal{Q}} \mathcal{L}(\pi_\theta, w, Q) - \nabla_\theta \max_{w \in \mathcal{W}} \min_{Q \in \mathcal{Q}} \mathcal{L}^D(\pi_\theta, w, Q)\| \leq \varepsilon_{data}$$

We defer the proof to Appendix A.

Definition 2.5 (Mis-specification Error)**.**

(1) For arbitrary $\pi \in \Pi$, denote $w_{\zeta^\pi} := \arg\min_{w \in \mathcal{W}} \|w - w^\pi_\mathcal{L}\|^2_\Lambda$ parameterized by $\zeta^\pi \in Z$, where $w^\pi_\mathcal{L} = \arg\max_{w \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \min_{Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \mathcal{L}(\pi, w, Q)$. We define

$$\varepsilon_1 := \max_{\pi \in \Pi} \|w_{\zeta^\pi} - w^\pi_\mathcal{L}\|^2_\Lambda$$

(2) For arbitrary policy $\pi \in \Pi$ and $w \in \mathcal{W}$, denote $Q_{\xi^\pi_w} := \arg\min_{Q \in \mathcal{Q}} \mathcal{L}(\pi, w, Q)$ parameterized by $\xi^\pi_w \in \Xi$. We define

$$\varepsilon_2 := \max_{w \in \mathcal{W}, \pi \in \Pi} \|Q_{\xi^\pi_w} - \arg\min_{Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \mathcal{L}(\pi, w, Q)\|^2_\Lambda$$

A consequence of Assumptions A and C is Proposition 2.6, that we can use ε_1 and ε_2 defined in Definition 2.5 to bound the weighted difference between the saddle points of $\mathcal{L}^D(\pi, w, Q)$ with and without constraining w and Q on $\mathcal{W} \times \mathcal{Q}$, respectively, which is crucial to analyzing the bias resulting from the mis-specified function classes. We defer its proof to Appendix A.

Proposition 2.6. *Under Assumption A and C, for arbitrary $\pi \in \Pi$, we have:*

$$\mathbb{E}_{d^\mu}[|w^*_\mu(s,a) - w^\pi_\mathcal{L}(s,a)|^2] \leq \varepsilon_\mathcal{W} := 4\frac{\lambda^2_{\max}}{\lambda_Q \lambda_w}\varepsilon_1 + 2\frac{L^2_w \lambda_{\max}}{\mu_\zeta}\varepsilon_2$$

$$\mathbb{E}_{d^\mu}[|Q^*_\mu(s,a) - Q^\pi_\mathcal{L}(s,a)|^2] \leq \varepsilon_\mathcal{Q} := 8\frac{\lambda^3_{\max}}{\lambda^2_Q \lambda_w}\varepsilon_1 + (2 + 4\frac{L^2_w \lambda^2_{\max}}{\lambda_Q \mu_\zeta})\varepsilon_2$$

where (w^*_μ, Q^*_μ) denotes the saddle point of $\mathcal{L}(\pi, w, Q)$ constrained by $w, Q \in \mathcal{W} \times \mathcal{Q}$, $(w^\pi_\mathcal{L}, Q^\pi_\mathcal{L})$ denotes the saddle point of $\mathcal{L}(\pi, w, Q)$ without any constraint on w and Q, $\lambda_{\max} = \max\{\lambda_Q, \lambda_w\}$, L_w is defined in Assumption A, μ_ζ is defined in Assumption C.

2.3 Main goal of the analyses

First, by applying the triangle inequality, we have:

$$\|\nabla_\theta J(\pi_\theta)\| \leq \|\nabla_\theta \max_{w\in\mathcal{W}} \min_{Q\in\mathcal{Q}} \mathcal{L}^D(\pi_\theta, w, Q)\| + \|\nabla_\theta J(\pi_\theta) - \nabla_\theta \max_{w\in\mathcal{W}} \min_{Q\in\mathcal{Q}} \mathcal{L}^D(\pi_\theta, w, Q)\|$$

where w^*, Q^* denotes the saddle point of $\mathcal{L}^D(\pi_\theta, w, Q)$ constrained by $w, Q \in \mathcal{W}\times\mathcal{Q}$. Optimizing the loss function $\mathcal{L}^D(\pi, w, Q)$ may offer us a better θ to decrease the first term, while based on above Assumptions, we can bound the second term in the following Theorem.

Theorem 2.7. *[Bias] Under Assumption A, B, C, given arbitrary $\theta \in \Theta$, we have*

$$\|\nabla_\theta \max_{w\in\mathcal{W}} \min_{Q\in\mathcal{Q}} \mathcal{L}^D(\pi_\theta, w, Q) - \nabla_\theta J(\pi_\theta)\| \leq \varepsilon_{reg} + \varepsilon_{func} + \varepsilon_{data}$$

where ε_{data} is defined in Proposition 2.4, and

$$\varepsilon_{func} = \frac{G}{1-\gamma}\Big(\sqrt{C\varepsilon_{\mathcal{Q}}} + C_{\mathcal{W}}\sqrt{\frac{\gamma\varepsilon_{\mathcal{Q}}C}{1-\gamma}} + \sqrt{\frac{\gamma\varepsilon_{\mathcal{Q}}\varepsilon_{\mathcal{W}}C}{1-\gamma}} + \gamma C_{\mathcal{Q}}\sqrt{\varepsilon_{\mathcal{W}}}\Big)$$

($\varepsilon_{\mathcal{W}}$ and $\varepsilon_{\mathcal{Q}}$ defined in Prop. 2.6)

$$\varepsilon_{reg} = \frac{G}{1-\gamma}\Big(\frac{C^2}{(1-\gamma)}(\frac{\lambda_w\lambda_Q}{1-\gamma}+\lambda_w) + \frac{\gamma C(\lambda_Q+\lambda_Q\lambda_w C)}{(1-\gamma)^3} + \frac{C^2(\lambda_Q+\lambda_Q\lambda_w C)}{(1-\gamma)^3}(\frac{\lambda_w\lambda_Q}{1-\gamma}+\lambda_w)\sqrt{\frac{\gamma C}{1-\gamma}}\Big)$$

We defer its proof to Appendix B.

As we can see, $\|\nabla_\theta \max_{w\in\mathcal{W}} \min_{Q\in\mathcal{Q}} \mathcal{L}^D(\pi_\theta, w, Q) - \nabla_\theta J(\pi_\theta)\|$ can be controlled by three terms. ε_{data} reflects the generalization error, and should be small if we have plenty of data. ε_{reg} depends on the magnitude of regularization, and will decrease as λ_w and λ_Q. As for ε_{func}, it depends on the approximation error $\varepsilon_{\mathcal{W}}$ and $\varepsilon_{\mathcal{Q}}$, which are propotional to ε_1 and ε_2. Besides, because μ_ζ should be proportional to λ_w and L_w does not depend on regularization, the coefficients before ε_1 and ε_2 should not vary a lot as we change λ_w and λ_Q while keeping $\lambda_w \approx \lambda_Q$ (but ε_1 and ε_2 may change with λ_w and λ_Q). In general, a larger dataset, better function classes and smaller λ_w and λ_Q may result in smaller bias, while smaller regularization can lead to weaker strong-concavity or strong-convexity of the loss function and make the convergence slower.

Based on the discussion above, our goal is to find stochastic optimization algorithms, which can return us π_θ after consuming $Poly(\varepsilon^{-1})$ samples from dataset (we omit the dependence on others such as μ_ζ, μ_ξ and etc.), satisfying the following biased stationary condition in Definition 1.1:

$$\mathbb{E}[\|\nabla_\theta J(\pi_\theta)\|] \leq \varepsilon + \varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg} \tag{5}$$

where ε_{data} is defined in 2.3 and ε_{func} and ε_{reg} are defined in Theorem 2.7.

Since D can be extremely large, we consider stochastic optimization, and introduce another crucial assumption about the stochastic gradient:

Assumption E (Variance of Estimated Gradient)**.** We use $\mathbb{E}_{s,a,r,s',a_0,a'}[\cdot]$ as a short note of

$$\mathbb{E}_{(s,a,r,s')\sim d^D, a_0\sim\pi(\cdot|s), a'\sim\pi(\cdot|s')}[\cdot]$$

and use $\mathcal{L}^{(s,a,r,s',a_0,a')}(\theta,\zeta,\xi)$ to denote the gradient estimation with only one sample defined by:

$$(1-\gamma)Q_\xi(s,a_0)\pi_\theta(a_0|s)\mathbb{I}[s\in S_0] + w_\zeta(s,a)\Big(r+\gamma Q_\xi(s',a')\pi_\theta(a'|s') - Q_\xi(s,a)\Big) + \frac{\lambda_Q}{2}Q_\xi^2(s,a) - \frac{\lambda_w}{2}w_\zeta^2(s,a)$$

where $\mathbb{I}[s\in\mathcal{S}_0]$ equals 1 only if s is generated at the first step in a trajectory and equals 0 otherwise (note that we allow the case when a state in the initial state sets can be visited at step $t \geq 1$.). We assume that, there exists a positive constant σ, for arbitrary $\theta,\zeta,\xi \in \Theta\times Z\times\Xi$, we have:

$$\mathbb{E}_{s,a,r,s',a_0,a'}[\|\nabla_\theta\mathcal{L}^{(s,a,r,s',a_0,a')}(\theta,\zeta,\xi) - \nabla_\theta\mathcal{L}^D(\theta,\zeta,\xi)\|^2] \leq \sigma^2$$
$$\mathbb{E}_{s,a,r,s',a_0,a'}[\|\nabla_\zeta\mathcal{L}^{(s,a,r,s',a_0,a')}(\theta,\zeta,\xi) - \nabla_\zeta\mathcal{L}^D(\theta,\zeta,\xi)\|^2] \leq \sigma^2$$
$$\mathbb{E}_{s,a,r,s',a_0,a'}[\|\nabla_\xi\mathcal{L}^{(s,a,r,s',a_0,a')}(\theta,\zeta,\xi) - \nabla_\xi\mathcal{L}^D(\theta,\zeta,\xi)\|^2] \leq \sigma^2$$

Remark 2.8. *The upper bound on the variance of the gradients w.r.t. θ,ζ and ξ are usually assumed to be different. Here we use σ to refer to the maximum of these upper bounds to simplify notations.*

3 Strategy 1: Converting Max-Max-Min to Max-min problem

A heuristic optimization strategy for (2) is to rewrite the original max-max-min problem $\max_\theta \max_\zeta \min_\xi \mathcal{L}^D(\theta,\zeta,\xi)$ to a max-min problem $\max_{\theta,\zeta} \min_\xi \mathcal{L}^D(\theta,\zeta,\xi)$. Given Assumption A and C, we know $\max_{\theta,\zeta} \min_\xi \mathcal{L}^D(\theta,\zeta,\xi)$ is a standard non-concave-strongly-convex problem, which can be solved efficiently based on the recent progress on non-convex-strongly-concave optimization [20, 8].

In this section, we prove the equivalence between the stationary point of the non-convex-strongly-concave saddle-point problem and the stationary point of our policy gradient objective:

Theorem 3.1. *[Equivalence Between Stationary Points] Under Assumption A, C and D, suppose there exists a $\theta \in \Theta$ s.t. $\|\nabla_\theta \max_{\zeta\in Z} \min_{\xi\in\Xi} \mathcal{L}^D(\theta,\zeta,\xi)\| = 0$ and there is an Algorithm provides us one stationary point (θ_T,ζ_T,ξ_T) of the non-concave-strongly-convex problem $\max_{\theta,\zeta} \min_\xi \mathcal{L}^D(\theta,\zeta,\xi)$ after running T iterations, which statisfying the following conditions in expectation over the randomness of algorithm.*

$$\begin{aligned}&\mathbb{E}[\|\nabla_{\theta,\zeta}\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\|]\\ :=&\mathbb{E}[\|\nabla_\theta\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\| + \|\nabla_\zeta\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\|] \le \frac{\varepsilon}{(\kappa_\xi+1)(\kappa_\zeta+1)}\end{aligned} \tag{6}$$

where $\phi_\theta(\zeta) = \arg\min_{\xi\in\Xi} \mathcal{L}^D(\theta,\zeta,\xi)$. Then, we have

$$\mathbb{E}[\|\nabla_\theta J(\pi_{\theta_T})\|] \le \varepsilon + \varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg}$$

In Appendix C, we will give the detailed proof. Besides, we also list algorithm examples which can return us stationary points satisfying Eq.(6).

4 Strategy 2: Stochastic Recursive Momentum with Saddle-Point Oracle

In this section, we propose a new algorithm, based on stochastic recursive momentum and a saddle-point oracle. We will provide a concrete example of the oracle algorithm in the Appendix E.

Definition 4.1 (Oracle Algorithm)**.** Suppose we have an oracle algorithm $Oracle$. For arbitrary strongly-concave-strongly-convex problem $f(\zeta,\xi)$ with saddle point $(\zeta^*,\xi^*) \in Z\times\Xi$, and arbitrary $0<\beta\le 1$ and $c>0$, starting from a random initializer $(\zeta_0,\xi_0)\in Z\times\Xi$ and executing finite steps, $Oracle$ returns a solution (ζ_K,ξ_K) satisfying

$$\mathbb{E}[\|\zeta_K-\zeta^*\|^2 + \|\xi_K-\xi^*\|^2] \le \frac{\beta}{2}\mathbb{E}[\|\zeta_0-\zeta^*\|^2 + \|\xi_0-\xi^*\|^2] + c \tag{7}$$

Next, we present our oracle based stochastic recursive momentum algorithm (O-SRM), inspired by the on-policy SRM [17]. In our algorithm, we choose $\Theta = \mathbb{R}^{u_\Theta}$ where u_Θ is the dimension of Θ. As a result, we will not do projection after update θ and there must exist stationary points of $J(\pi_\theta)$ and $\max_{\zeta\in Z}\min_{\xi\in\Xi}\mathcal{L}^D(\theta,\zeta,\xi)$. We will use $\nabla_\theta\mathcal{L}^B(\theta,\zeta,\xi)$ as a short note of the empirical version of the gradient estimator, i.e.

$$\nabla_\theta\mathcal{L}^B(\theta,\zeta,\xi) = \frac{1}{|B|}\sum_B (1-\gamma)Q(s^i,a_0^i)\nabla_\theta\log\pi(a_0^i|s^i)\mathbb{I}[s^i\in S_0] + \gamma w(s^i,a^i)Q(s'^i,a'^i)\nabla_\theta\log\pi(a'^i|s'^i)$$

where (s^i, a^i, r^i, s'^i) for $i = 1, 2, ..., |B|$ are elements in B sampled from d^D, and $a_0^i \sim \pi(\cdot|s^i), a'^i \sim \pi(\cdot|s'^i)$.

Algorithm 1: O-SRM

Input: Total number of iteration T; Learning rate $\eta_\theta, \eta_\zeta, \eta_\xi$; Dataset distribution d^D; Oracle parameter β.
Initialize $\theta_0, \zeta_{-1}, \xi_{-1}$
$\zeta_0, \xi_0 \leftarrow \text{Oracle}(T_1, \eta_\zeta, \eta_\xi, \theta_0, \zeta_{-1}, \xi_{-1}, d^D)$
Sample $B_0 \sim d^D$ with batch size $|B_0|$ and estimate $g_\theta^0 = \nabla_\theta \mathcal{L}^{B_0}(\theta_0, \zeta_0, \xi_0)$
for $t = 0, 1, 2, ...T-1$ **do**
  $\theta_{t+1} \leftarrow \theta_t + \eta_\theta g_\theta^t$
  $\zeta_{t+1}, \xi_{t+1} \leftarrow Oracle(\beta, \theta_{t+1}, \zeta_t, \xi_t, d^D, \beta)$
  Sample $B \sim d^D$;
  $g_\theta^{t+1} = (1-\alpha)\Big(g_\theta^t - \nabla_\theta \mathcal{L}^B(\theta_t, \zeta_t, \xi_t)\Big) + \nabla_\theta \mathcal{L}^B(\theta_{t+1}, \zeta_{t+1}, \xi_{t+1})$
end
Output: Sample $\theta_{out} \sim \text{Unif}\{\theta_0, \theta_1, ..., \theta_T\}$ and output π_θ.

4.1 Additional Assumptions for Algorithm 1

Assumption F (Diameter)**.** We use Z and Ξ to denote the sets of parameters ζ and ξ, respectively, we assume Z and Ξ are both convex and bounded set, and there exists a constant d, such that the diameters of Z and Ξ are bounded by d.

4.2 Algorithm Analysis

We first derive the smoothness of $J(\pi_\theta)$:

Proposition 4.2. *Under Assumption A, $J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta, s_0\sim\nu_0}[\sum_{t=0}^\infty \gamma^t r(s_t, a_t)]$ is L_J smooth with*

$$L_J := \frac{H}{(1-\gamma)^2} + \frac{(1+\gamma)G^2}{(1-\gamma)^3}$$

Theorem 4.3. *Under Assumption A-F and H, given arbitrary ε, by choosing Algorithm 3 as the $Oracle$, Algorithm 1 will return us a policy $\pi_{\theta_{out}}$, satisfying*

$$\mathbb{E}[\|\nabla_\theta J(\pi_{\theta_T})\|] \le \varepsilon + \sqrt{3}(\varepsilon_{reg} + \varepsilon_{data} + \varepsilon_{func})$$

if the hyper-parameters in Alg. 1 and 3 satisfy the following constraints:

$$T = [\max\{96, \frac{16L_J}{\varepsilon^2}, \frac{48}{(1-\gamma)\varepsilon^2}\sqrt{\Big(2C_{\zeta,\mu}C_{w,Q} + H^2 C_{\mathcal{Q}}^2 C_{\mathcal{W}}^2\Big)}, \frac{72C_{w,Q}d^2}{\varepsilon\sigma}\}] = O(\varepsilon^{-2});$$

$$|B| = \max\{13, \frac{12\sigma}{\varepsilon}\}; \quad |N| = [\frac{48(2L^2+3C_{w,Q})\sigma^2}{\min\{\frac{\mu_\zeta\eta_\zeta}{4}, \frac{\mu_\xi\eta_\xi}{4}\}\varepsilon^2}(\frac{\eta_\zeta}{\mu_\zeta} + \frac{\eta_\xi}{\mu_\xi})]; \quad K = c_{oracle}\log(\frac{1}{\beta});$$

$$\alpha = 12/|B| = \min\{\frac{12}{13}, \frac{\varepsilon}{\sigma}\}; \quad \beta = \min\{\frac{\varepsilon^2}{L^2}, \frac{(1-\gamma)^2\varepsilon^4}{C_{\zeta,\mu}L^2}, \frac{\alpha}{2}(1-\alpha)^2\}; \quad B_0 = [\frac{8\sigma^2}{\varepsilon^2}]$$

$$\eta_\theta = \min\{\frac{1}{2L_J}, \Big(108\Big[\frac{C_{\zeta,\mu}L^2\beta}{18(1-\beta)} + \frac{1}{12}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{W}}^2C_{\mathcal{Q}}^2\Big)\Big]\Big)^{-1/2}\}$$

where $[\cdot]$ is the upper rounding function, $C_{w,Q} = G^2L_w^2C_{\mathcal{Q}}^2 + G^2C_{\mathcal{W}}^2L_Q^2$, $C_{\zeta,\mu} = \kappa_\mu^2(\kappa_\xi+1)^2 + \kappa_\xi^2(\kappa_\mu+1)^2$, L_J is defined in Prop. 4.2, η_ζ and η_ξ satisfy the constraints in Theorem E.1 and c_{oracle} is an independent constant.

Besides, the total gradient computation to obtain θ_{out} should be $|B_0|+|B|\cdot T+|N|\cdot K\cdot T = O(\varepsilon^{-4})$.

We defer the proofs to Appendix D.

5 Conclusion

In this paper, we study two natural optimization strategies for density-ratio based off-policy policy gradients, establish their convergence rates, and characterize the quality of the results. In the future, it will be interesting to extend the results to other settings with milder assumptions, or improve the dependence on ε^{-1} on the convergence rate of our second strategy.

References

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

[2] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5361–5371, 2018.

[3] Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2019.

[4] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

[5] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*, 2020.

[6] Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

[7] Nan Jiang and Jiawei Huang. Minimax confidence interval for off-policy evaluation and policy optimization. *arXiv preprint arXiv:2002.02081*, 2020.

[8] Luo Luo, Ye Haishan, and Zhang Tong. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. 2020.

[9] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 2315–2325, 2019.

[10] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *CoRR*, abs/1904.08473, 2019. URL `http://arxiv.org/abs/1904.08473`.

[11] Thomas Degris, Martha White, and Richard S. Sutton. Off-policy actor-critic. *CoRR*, abs/1205.4839, 2012. URL `http://arxiv.org/abs/1205.4839`.

[12] Ehsan Imani, Eric Graves, and Martha White. An off-policy policy gradient theorem using emphatic weightings. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 96–106, 2018.

[13] Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation, 2019.

[14] Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. *arXiv preprint arXiv:1806.05618*, 2018.

[15] Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. *arXiv preprint arXiv:1905.12615*, 2019.

[16] Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019.

[17] Huizhuo Yuan, Xiangru Lian, Ji Liu, and Yuren Zhou. Stochastic recursive momentum for policy gradient methods. *arXiv preprint arXiv:2003.04302*, 2020.

[18] F. Huang, Shangqian Gao, Jian Pei, and H. Huang. Momentum-based policy gradient methods. *ArXiv*, abs/2007.06680, 2020.

[19] Richard S Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1057–1063. MIT Press, 2000.

[20] Tianyi Lin, Chi Jin, and Michael I Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331*, 2019.

[21] Tianyi Lin, Chi Jin, Michael Jordan, et al. Near-optimal algorithms for minimax optimization. *arXiv preprint arXiv:2002.02417*, 2020.

[22] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. In *Advances in Neural Information Processing Systems*, pages 393–403, 2019.

[23] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1042–1051, 2019.

[24] Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of LSTD. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 615–622. Omnipress, 2010. URL `https://icml.cc/Conferences/2010/papers/598.pdf`.

[25] Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *J. Mach. Learn. Res.*, 13:3041–3074, 2012. URL `http://dl.acm.org/citation.cfm?id=2503339`.