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

Anonymous Author(s)
Affiliation
Address
email

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:

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

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

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

Submitted to Offline Reinforcement Learning Workshop in 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

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 E).

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 discuss the practicality of such an oracle in Appendix D. 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 algorithm is also $O(\varepsilon^{-3})$.

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 E 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 $dim(Z)$ and $dim(\Xi)$ to denote the dimension of vector parameters ζ and ξ. Given arbitrary $\theta \in \Theta, \zeta \in Z$, $\mathcal{L}^D(\theta, \zeta, \cdot)$ is μ_ξ-strongly convex w.r.t. $\xi \in \mathbb{R}^{dim(\Xi)}$. Given arbitrary $\theta \in \Theta, \xi \in \Xi$, $\mathcal{L}^D(\theta, \cdot, \xi)$ is μ_ζ-strongly concave w.r.t. $\zeta \in \mathbb{R}^{dim(Z)}$.

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)**.** We will use ε_{data} to denote the generalization error defined in the following:

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

Definition 2.4 (Mis-specification Error)**.**

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

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

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

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

A consequence of Assumptions A and C is Proposition 2.5, that we can use ε_1 and ε_2 defined in Definition 2.4 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 2.5. *Under Assumption A and C, for arbitrary $\pi \in \Pi$, we have:*

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

where (w_μ^*, Q_μ^*) denotes the saddle point of $\mathcal{L}(\pi, w, Q)$ constrained by $w, Q \in \mathcal{W} \times \mathcal{Q}$, $(w_\mathcal{L}^\pi, Q_\mathcal{L}^\pi)$ 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 2.6. *[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 Definition 2.3, and

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

($\varepsilon_{\mathcal{W}}$ and $\varepsilon_{\mathcal{Q}}$ defined in Prop. 2.5)

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

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:

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

where ε_{data} is defined in 2.3 and ε_{func} and ε_{reg} are defined in Theorem 2.6.

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 we use S_0 to denote the set of initial states. We assume that, there exists a positive constant σ, for arbitrary $\theta,\zeta,\xi\in\Theta\times Z\times\Xi$, we have:

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

Remark 2.7. *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 an Algorithm provides us one stationary point* $(\theta_T, \zeta_T, \xi_T)$ *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.*

$$\begin{aligned}\mathbb{E}[\|\nabla_{\theta,\zeta}\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\|] :=& \mathbb{E}[\|\nabla_{\theta}\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\| + \|\nabla_{\zeta}\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\|] \\ \leq& \frac{\varepsilon}{(\kappa_\xi+1)^2}\end{aligned} \tag{6}$$

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

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

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 defer the discussion about the practicality of the oracle to Appendix D.

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 $0<\beta\leq 1$, starting from a random initializer (ζ_0,ξ_0) and executing $K = c_{oracle}\log(\frac{1}{\beta})$ steps, where c_{oracle} is a positive constant independent with β, $Oracle$ returns a solution (ζ_K,ξ_K) satisfying

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

Next, we present our oracle based stochastic recursive momentum algorithm (O-SRM), inspired by the on-policy SRM [17]. 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.

$$\begin{aligned}\nabla_\theta\mathcal{L}^B(\theta,\zeta,\xi) = \frac{1}{|B|}\sum_B &(1-\gamma)Q(s^i,a_0^i)\pi(a_0^i|s^i)\mathbb{I}[s^i\in S_0] \\ &+ w(s^i,a^i)\Big(r^i + \gamma Q(s'^i,a'^i)\pi(a'^i|s'^i) - Q(s^i,a^i)\Big) \\ &+ \frac{\lambda_Q}{2}Q^2(s^i,a^i) - \frac{\lambda_w}{2}w^2(s^i,a^i)\end{aligned}$$

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 $\theta\sim\text{Unif}\{\theta_0,\theta_1,...,\theta_T\}$ 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. *Given arbitrary ε, suppose $|B|$ and T satisfy the following constraints:*

$$\begin{aligned} T &\approx \max\{96, \frac{16L_J}{\varepsilon^2}\} = O(\varepsilon^{-2}) \\ |B|T &\approx \max\{\frac{576\sigma}{(1-\gamma)\varepsilon^3}\sqrt{2C_{\zeta,\mu}C_{w,Q} + H^2 C_{\mathcal{W}}^2 C_{\mathcal{Q}}^2}, \frac{864 C_{w,Q} d^2}{\varepsilon^2}\} = O(\varepsilon^{-3}) \end{aligned}$$

where $C_{w,Q} = G^2 L_w^2 C_{\mathcal{Q}}^2 + G^2 C_{\mathcal{W}}^2 L_Q^2$, $C_{\zeta,\mu} = \kappa_\mu^2(\kappa_\xi + 1)^2 + \kappa_\xi^2(\kappa_\mu + 1)^2$ and L_J is defined in Prop. 4.2, while other hyper-parameters satisfy:

$$\begin{aligned} \alpha =& \frac{|B|\varepsilon^2}{12\sigma}; \quad \beta \le \min\{\frac{\varepsilon^2}{L^2}, \frac{(1-\gamma)^2\varepsilon^4}{C_{\zeta,\mu}L^2}, \frac{\alpha}{2}(1-\alpha)^2\}; \quad B_0 = \frac{4\sigma^2}{\varepsilon^2} \\ \eta_\theta \le& \min\{\frac{1}{2L_J}, \Big(108\Big[\frac{C_{\zeta,\mu}L^2\beta}{18(1-\beta)} + \frac{1}{\alpha|B|}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2 C_{\mathcal{W}}^2 C_{\mathcal{Q}}^2\Big)\Big]\Big)^{-1/2}\} \end{aligned}$$

The Algorithm 2 will return us a policy π_{θ_T} after T steps with batch size $|B|$, satisfying

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

The total gradient computation of Algorithm 1 (ignoring Oracle) is $|B_0| + |B|T = O(\varepsilon^{-3})$.

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, and give concrete examples for the oracle in Section 4.

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`.

A Useful Lemma

Lemma A.1 (Lemma B.2 in [21]). *Define*

$$\Phi_\theta(\zeta) = \min_{\xi\in\Xi} \mathcal{L}^D(\theta,\zeta,\xi) \qquad \phi_\theta(\zeta) = \arg\min_{\xi\in\Xi} \mathcal{L}^D(\theta,\zeta,\xi), \quad for\ \zeta\in\mathbb{R}^{dim(Z)}$$
$$\Psi_\theta(\xi) = \max_{\zeta\in Z} \mathcal{L}^D(\theta,\zeta,\xi) \qquad \psi_\theta(\xi) = \arg\max_{\zeta\in Z} \mathcal{L}^D(\theta,\zeta,\xi), \quad for\ \xi\in\mathbb{R}^{dim(\Xi)}$$

Under Assumption A and C, for fixed θ, we have:

(1) The function $\phi_\theta(\cdot)$ is $\kappa_\xi = \frac{L}{\mu_\xi}$-Lipschitz.

(2) The function $\Phi_\theta(\cdot)$ is $2\kappa_\xi L = 2\frac{L^2}{\mu_\xi}$-smooth and μ_ζ-strongly concave with $\nabla\Phi_\theta(\cdot) := \nabla_\zeta\mathcal{L}^D(\theta,\zeta,\phi_\theta(\zeta))$.

(3) The function $\psi_\theta(\cdot)$ is $\kappa_\zeta = \frac{L}{\mu_\zeta}$-Lipschitz.

(4) The function $\Psi_\theta(\cdot)$ is $2\kappa_\zeta L = 2\frac{L^2}{\mu_\zeta}$-smooth and μ_ξ-strongly convex with $\nabla\Psi_\theta(\cdot) := \nabla_\xi\mathcal{L}^D(\theta,\psi_\theta(\xi),\xi)$.

Remark A.2 (For clarification). *In $\nabla\Phi_\theta(\cdot) := \nabla_\zeta\mathcal{L}^D(\theta,\zeta,\phi_\theta(\zeta))$, when we compute $\nabla_\zeta\mathcal{L}^D(\theta,\zeta,\phi_\theta(\zeta))$, we treat $\phi_\theta(\zeta)$ as a constant, instead of a function w.r.t. ζ. Therefore, for arbitrary ζ',ξ', based on Assumption A, we always have:*

$$\|\nabla\Phi_\theta(\cdot) - \nabla_\zeta\mathcal{L}^D(\theta,\zeta',\xi')\| \le L\|\zeta-\zeta'\| + L\|\phi_\theta(\zeta)-\xi'\|$$

We have a similar clarification w.r.t. $\nabla_\xi\Psi(\xi)$.

Lemma A.3. *For α-strongly-convex function $f(x)$ and β-strongly-concave function $g(x)$ w.r.t. $x\in X$, where $X\subseteq\mathbb{R}^n$ is a convex set. Then, we have*

$$\|x-x_f^*\| \le \frac{1}{\alpha}\|\nabla_x f(x)\| \tag{8}$$
$$\frac{\alpha}{2}\|x-x_f^*\|^2 \le f(x) - f(x_f^*) \tag{9}$$
$$\|x-x_g^*\| \le \frac{1}{\beta}\|\nabla_x g(x)\| \tag{10}$$
$$\frac{\beta}{2}\|x-x_f^*\|^2 \le g(x_g^*) - g(x) \tag{11}$$

where x_f^* and x_g^* the minimum and maximum of $f(x)$ and $g(x)$, respectively.

Proof. Since $f(x)$ is α-strongly-convex, we have

$$(\nabla_x f(x) - \nabla_x f(x_f^*))^\top(x-x_f^*) \ge \alpha\|x-x_f^*\|^2$$
$$f(x) \ge f(x_f^*) + \nabla_x f(x_f^*)^\top(x-x_f^*) + \frac{\alpha}{2}\|x-x_f^*\|^2$$

Since x_f^* is the minimizer of $f(x)$, we know that

$$\nabla_x f(x_f^*)^\top(x-x_f^*) \ge 0$$

Combining all the above inequalities together and we obtain

$$\|x-x_f^*\|^2 \le \frac{1}{\alpha}\nabla_x f(x)^\top(x-x_f^*) \le \frac{1}{\alpha}\|\nabla_x f(x)\|\|x-x_f^*\|$$
$$f(x) \ge f(x_f^*) + \frac{\alpha}{2}\|x-x_f^*\|^2$$

which implies

$$\|x-x_f^*\| \le \frac{1}{\alpha}\|\nabla_x f(x)\|$$
$$\frac{\alpha}{2}\|x-x_f^*\|^2 \le f(x) - f(x_f^*)$$

By applying the above results for $-g(x)$ which is a β-strongly-convex function and we can complete the proof. $\square$

Lemma A.4. *For positive definite matrix A, and arbitrary $\alpha > 0$, we have:*

$$(A^\top A)^{-1} \succ \Big((\alpha I + A)^\top (\alpha I + A)\Big)^{-1}$$

Proof. Suppose for symmetric matrix A and B, we have the relationship $A \succ B \succ 0$. According to the inverse matrix lemma, we have

$$B^{-1} - A^{-1} = B^{-1} - (B + (A - B))^{-1} = (B + B(A - B)^{-1} B)^{-1}$$

Because $A \succ B \succ 0$, we have $(B + B(A - B)^{-1} B)^{-1} \succ 0$, therefore $B^{-1} \succ A^{-1}$.

Then, we only need to prove

$$(\alpha I + A)^\top (\alpha I + A) \succ A^\top A$$

We have

$$(\alpha I + A)^\top (\alpha I + A) = \alpha^2 I + \alpha(A + A^\top) + A^\top A$$

Combining $A = A^\top \succ 0$ and $\alpha > 0$, we can finish the proof. $\square$

Lemma A.5 (Non-negative Elements)**.** *We use $P^\pi_* = (P^\pi)^\top \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ to denote the transpose of the transition kernel. All the elements in $(I - \gamma P_*)^{-1}$ are non-negative. Moreover, the element indexed by (s_i, a_j) in row and (s_p, a_q) in column equals to the discounted state-action occupancy of (s_i, a_j) starting from (s_p, a_q).*

Proof. For arbitrary initial state-action distribution vector $\mu_0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times 1}$, $(I - \gamma P_*)^{-1} \mu_0$ is a vector whose elements are unnormalized state-action occupancy with μ_0 as initial distribution, which is larger or equal to 0. As a result, by choosing standard basis vector as μ_0, we can finish the proof. $\square$

Proposition 2.5. *Under Assumption A and C, for arbitrary $\pi \in \Pi$, we have:*

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

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.

Proof. In the following, we will frequently consider two loss functions. The first one is $\mathcal{L}(\pi, w, Q)$ defined in Eq.(1), where w and Q are parameterized by ζ and ξ, respectively, and we will write $(w, Q) \in \mathcal{W} \times \mathcal{Q}$. The second one is $\mathcal{F}(\pi, x, y)$ defined by:

$$\mathcal{F}(\pi, x, y) = (1-\gamma)(\nu_0^\pi)^\top \Lambda^{-1/2} y + x^\top \Big(\Lambda^{1/2} R - (I - \gamma \Lambda^{1/2} P^\pi \Lambda^{-1/2}) y\Big) + \frac{\lambda_Q}{2} y^\top y - \frac{\lambda_w}{2} x^\top x$$

where $(x, y) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \times \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. For simplification, in the following, we will use $\max_x \min_y$ as a short note of $\max_{x \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \min_{y \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}}$.

As we can see, the difference between $\mathcal{L}(\pi, w, Q)$ and $\mathcal{F}(\pi, x, y)$ is not only that we don't have any constraint on x and y, but also that we absorb one $\Lambda^{1/2}$ into vector x and y. In another word, for arbitrary π, w, Q, we have

$$\mathcal{L}(\pi, w, Q) = \mathcal{F}(\pi, \Lambda^{1/2} w, \Lambda^{1/2} Q).$$

Obviously, $\mathcal{F}(\pi, x, y)$ is λ_w-strongly-concave-λ_Q-strongly-convex and $\lambda_{\max}$-smooth w.r.t. $x, y \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

Next, for arbitrary ζ, we have:

$$\mathcal{F}(\pi, \Lambda^{1/2} w_\zeta, \Lambda^{1/2} Q_{\xi_w^\pi}) = \min_{Q \in \mathcal{Q}} \mathcal{F}(\pi, \Lambda^{1/2} w_\zeta, \Lambda^{1/2} Q) \geq \min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\zeta, y)$$

where $Q_{\xi_w^\pi}$ is defined in Definition 2.4. Combining $\nabla_y \min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\zeta, y) = 0$, we have:

$$\begin{aligned}
&\frac{\lambda_{\max}}{2} \|\Lambda^{1/2} Q_{\xi_w^\pi} - \arg\min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\zeta, y)\|^2 \\
\geq& \mathcal{F}(\pi, \Lambda^{1/2} w_\zeta, \Lambda^{1/2} Q_{\xi_w^\pi}) - \min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\zeta, y) && \text{(Smoothness of } \mathcal{F}) \\
\geq& \min_{Q \in \mathcal{Q}} \mathcal{F}(\pi, \Lambda^{1/2} w_\zeta, \Lambda^{1/2} Q) - \min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\zeta, y) \\
\geq& \frac{\lambda_Q}{2} \|\Lambda^{1/2} \arg\min_{Q \in \mathcal{Q}} \mathcal{F}(\pi, \Lambda^{1/2} w_\zeta, \Lambda^{1/2} Q) - \arg\min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\zeta, y)\|^2 \\
& && \text{(Strongly Convexity of } \mathcal{F})
\end{aligned}$$

Recall that $w_\mu^* = \arg\max_{w \in \mathcal{W}} \min_{Q \in \mathcal{Q}} \mathcal{F}(\pi, \Lambda^{1/2} w, \Lambda^{1/2} Q)$, and we use ζ^* to denote it's parameter. Note that, $Q_{\xi^\pi_{w_\mu^*}} = Q_\mu^*$. By choosing $w_\zeta = w_\mu^*$ (i.e. $\zeta = \zeta^*$) in the above inequality, we have

$$\begin{aligned}
&\|\Lambda^{1/2} Q_\mu^* - \arg\min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, y)\|^2 \\
=& \|\Lambda^{1/2} \arg\min_{Q \in \mathcal{Q}} \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, \Lambda^{1/2} Q) - \arg\min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, y)\|^2 \\
\leq& \frac{\lambda_{\max}}{\lambda_Q} \|\Lambda^{1/2} Q_{\xi^\pi_{w_\mu^*}} - \arg\min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, y)\|^2 \leq \frac{\lambda_{\max}}{\lambda_Q} \varepsilon_2
\end{aligned} \tag{12}$$

where ε_2 is defined in Def. 2.4.

In the following, we use $w_\mathbb{R}^*$ parameterized by $\zeta_\mathbb{R}^*$ to denote $\arg\max_{w \in \mathcal{W}} \min_y \mathcal{F}(\pi, \Lambda^{1/2} w, y)$. According to Lemma A.1, $\min_y \mathcal{F}(\pi, x, y)$ is a $2\frac{\lambda_{\max}^2}{\lambda_Q}$-smooth and λ_w-strongly-concave function with gradient $\nabla_x \min_y \mathcal{F}(\pi, x, y)$. Since $\nabla_x \mathcal{F}(\pi, \Lambda^{1/2} w_\mathcal{L}^\pi, \Lambda^{1/2} Q_\mathcal{L}^\pi) = 0$, we have,

$$\begin{aligned}
&\frac{\lambda_w}{2} \|\Lambda^{1/2} w_\mathbb{R}^* - \Lambda^{1/2} w_\mathcal{L}^\pi\|^2 \\
\leq& \mathcal{F}(\pi, \Lambda^{1/2} w_\mathcal{L}^\pi, \Lambda^{1/2} Q_\mathcal{L}^\pi) - \min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\mathbb{R}^*, y) && \text{(Strong concavity of } \min_y \mathcal{F}(\pi, x, y)) \\
=& \mathcal{F}(\pi, \Lambda^{1/2} w_\mathcal{L}^\pi, \Lambda^{1/2} Q_\mathcal{L}^\pi) - \max_{w \in \mathcal{W}} \min_y \mathcal{F}(\pi, \Lambda^{1/2} w, y) \\
\leq& \mathcal{F}(\pi, \Lambda^{1/2} w_\mathcal{L}^\pi, \Lambda^{1/2} Q_\mathcal{L}^\pi) - \min_y \mathcal{F}(\pi, \Lambda^{1/2} w_{\zeta^\pi}, y) && (w_{\zeta^\pi} \text{ is defined in Def. 2.4}) \\
\leq& \frac{\lambda_{\max}^2}{\lambda_Q} \|\Lambda^{1/2} w_{\zeta^\pi} - \Lambda^{1/2} w_\mathcal{L}^\pi\|^2 && \text{(Smoothness of } \min_y \mathcal{F}(\pi, x, y)) \\
=& \frac{\lambda_{\max}^2}{\lambda_Q} \|w_{\zeta^\pi} - w_\mathcal{L}^\pi\|_\Lambda^2 = \frac{\lambda_{\max}^2}{\lambda_Q} \varepsilon_1 && \text{(see definition of } \varepsilon_1 \text{ in Def.2.4)}
\end{aligned}$$

which implies

$$\|\Lambda^{1/2} w_\mathbb{R}^* - \Lambda^{1/2} w_\mathcal{L}^\pi\|^2 \leq 2\frac{\lambda_{\max}^2}{\lambda_Q \lambda_w} \varepsilon_1 \tag{13}$$

Applying Lemma A.1 for $(w, Q) \in \mathcal{W} \times \mathcal{Q}$, we know $\min_{\xi \in \Xi} \mathcal{L}(\pi, w_\zeta, Q_\xi)$ is μ_ζ-strongly-concave w.r.t. ζ. Since ζ^* is the minimizer of $\min_{\xi \in \Xi} \mathcal{L}(\pi, w_\zeta, Q_\xi)$ and Z is a convex set, we have

$$\begin{aligned}
\frac{\mu_\zeta}{2} \|\zeta^* - \zeta_\mathcal{R}^*\|^2 \leq& \mathcal{L}(\pi, w_\mu^*, Q_\mu^*) - \min_{Q \in \mathcal{Q}} \mathcal{L}(\pi, w_\mathcal{R}^*, Q) \\
& \text{(Stong concavity of } \min_{Q \in \mathcal{Q}} \mathcal{L}(\pi, w, Q)\text{; Lemma A.3)} \\
=& \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, \Lambda^{1/2} Q_\mu^*) - \min_{Q \in \mathcal{Q}} \mathcal{F}(\pi, \Lambda^{1/2} w_\mathcal{R}^*, \Lambda^{1/2} Q) \\
\leq& \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, \Lambda^{1/2} Q_\mu^*) - \min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\mathcal{R}^*, y)
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, \Lambda^{1/2} Q_\mu^*) - \min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, y) \\
&\qquad\qquad\qquad\qquad (\text{Because } w_R^* = \arg\max_{w \in \mathcal{W}} \min_y \mathcal{F}(\pi, \Lambda^{1/2} w, y)) \\
&\leq \frac{\lambda_{\max}}{2} \|\Lambda^{1/2} Q_\mu^* - \arg\min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, y)\|^2 \\
&\qquad\qquad\qquad\qquad (\text{Smoothness of } \mathcal{F}(\pi, x, y) \text{ for fixed } x \text{ and } \nabla_y \min_y \mathcal{F} = 0) \\
&\leq \frac{\lambda_{\max}^2}{2\lambda_Q} \varepsilon_2
\end{aligned}
$$

318 In the last but two inequality, we use the fact that $\mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, \cdot)$ is $\lambda_{\max}$-smooth and
319 $\nabla_y \min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, Q) = 0$; in the last equality, we use Eq.(12). Combing (2) in Assump-
320 tion A, for arbitrary $s, a \in \mathcal{S} \times \mathcal{A}$, we have:

$$
|w_\mu^*(s,a) - w_{\mathbb{R}}^*(s,a)|^2 \leq L_w^2 \|\zeta^* - \zeta_{\mathcal{R}}^*\|^2 \leq \frac{L_w^2 \lambda_{\max}^2}{\lambda_Q \mu_\zeta} \varepsilon_2 \tag{14}
$$

321 Therefore, as a result of Eq.(13) and Eq.(14):

$$
\begin{aligned}
\mathbb{E}_{d^\mu}[|w_\mu^* - w_{\mathcal{L}}^\pi|^2] &\leq 2\mathbb{E}_{d^\mu}[|w_{\mathbb{R}}^* - w_{\mathcal{L}}^\pi|^2] + 2\mathbb{E}_{d^\mu}[|w_{\mathbb{R}}^* - w_\mu^*|^2] \\
&= 2\|\Lambda^{1/2} w_{\mathbb{R}}^* - \Lambda^{1/2} w_{\mathcal{L}}^\pi\|^2 + 2\mathbb{E}_{d^\mu}[|w_{\mathbb{R}}^* - w_\mu^*|^2] \\
&\leq 4\frac{\lambda_{\max}^2}{\lambda_Q \lambda_w} \varepsilon_1 + 2\frac{L_w^2 \lambda_{\max}^2}{\lambda_Q \mu_\zeta} \varepsilon_2 := \varepsilon_{\mathcal{W}}
\end{aligned}
$$

322 According to Lemma A.1 again, $\arg\min_y \mathcal{F}(\pi, x, y)$ is $\frac{\lambda_{\max}}{\lambda_w}$-Lipschitz w.r.t. x, we have

$$
\begin{aligned}
&\mathbb{E}_{d^\mu}[|Q_\mu^* - Q_{\mathcal{L}}^\pi|^2] = \|\Lambda^{1/2} Q_\mu^* - \Lambda^{1/2} Q_{\mathcal{L}}^\pi\|^2 \\
&\leq 2\underbrace{\|\Lambda^{1/2} Q_\mu^* - \arg\min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, Q)\|^2}_{bounded\ in\ Eq.(12)} + 2\|\arg\min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, y) - \Lambda^{1/2} Q_{\mathcal{L}}^\pi\|^2 \\
&\leq 2\frac{\lambda_{\max}}{\lambda_Q} \varepsilon_2 + 2\frac{\lambda_{\max}}{\lambda_w} \|\Lambda^{1/2} w_\mu^* - \Lambda^{1/2} w_{\mathcal{L}}^\pi\|^2 \\
&\leq 8\frac{\lambda_{\max}^3}{\lambda_Q \lambda_w^2} \varepsilon_1 + (2\frac{\lambda_{\max}}{\lambda_Q} + 4\frac{L_w^2 \lambda_{\max}^3}{\lambda_Q \lambda_w \mu_\zeta}) \varepsilon_2 := \varepsilon_{\mathcal{Q}}
\end{aligned}
$$

323 As a result,

$$
\varepsilon_{\mathcal{W}} = 4\frac{\lambda_{\max}^2}{\lambda_Q \lambda_w} \varepsilon_1 + 2\frac{L_w^2 \lambda_{\max}^2}{\lambda_Q \mu_\zeta} \varepsilon_2; \quad \varepsilon_{\mathcal{Q}} = 8\frac{\lambda_{\max}^3}{\lambda_Q \lambda_w^2} \varepsilon_1 + (2\frac{\lambda_{\max}}{\lambda_Q} + 4\frac{L_w^2 \lambda_{\max}^3}{\lambda_Q \lambda_w \mu_\zeta}) \varepsilon_2
$$

324 □

325 B The analysis of Bias

326 **Theorem B.1** (Bias resulting from regularization)**.** *Let's rewrite Eq.*(1) *in a vector-matrix form:*

$$
\max_{w \in \mathcal{W}} \min_{Q \in \mathcal{Q}} \mathcal{L}(\pi, w, Q) := (1-\gamma)(\nu_0^\pi)^\top Q + w^\top \Lambda \Big(R - (I - \gamma P^\pi) Q\Big) + \frac{\lambda_Q}{2} Q^\top \Lambda Q - \frac{\lambda_w}{2} w^\top \Lambda w
$$

327 *where ν_0^π and P^π denotes the initial state-action distribution and the transition matrix w.r.t. policy π,*
328 *respectively; $\Lambda \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ denotes the diagonal matrix whose diagonal elements are $d^\mu(\cdot, \cdot)$.*
329 *Denote $(w_{\mathcal{L}}^\pi, Q_{\mathcal{L}}^\pi)$ as the saddle point of $\mathcal{L}(\pi, w, Q)$ without any constraint on w and Q, then we*
330 *have:*

$$
\begin{aligned}
w_{\mathcal{L}}^\pi &= w^\pi + \Big(\lambda_w \lambda_Q I + (I - \gamma P^\pi) \Lambda^{-1} (I - \gamma P_*^\pi) \Lambda\Big)^{-1} \Big(\lambda_Q R - \lambda_Q \lambda_w w^\pi\Big) \\
Q_{\mathcal{L}}^\pi &= Q^\pi - \Big(\lambda_w \lambda_Q I + \Lambda^{-1} (I - \gamma P_*^\pi) \Lambda (I - \gamma P^\pi)\Big)^{-1} \Big(\lambda_w \lambda_Q Q^\pi + \lambda_w (1-\gamma) \Lambda^{-1} \nu_0^\pi)\Big)
\end{aligned}
$$

331 *where $w^\pi = \frac{d^\pi}{d^\mu}$ is the density ratio and Q^π is the Q function of π. we use $P_*^\pi = (P^\pi)^\top$ to denote*
332 *the transpose of the transition matrix.*

Proof. Recall the loss function

$$\mathcal{L}(\pi, w, Q) = (1-\gamma)(\nu_0^\pi)^\top Q + w^\top \Lambda R - w^\top \Lambda (I - \gamma P^\pi) Q + \frac{\lambda_Q}{2} Q^\top \Lambda Q - \frac{\lambda_w}{2} w^\top \Lambda w$$

By taking the derivatives w.r.t. Q, if K_Q is invertible, the optimal choice of Q should be:

$$Q = \frac{1}{\lambda_Q} \Lambda^{-1} ((I - \gamma P_*^\pi) \Lambda w - (1-\gamma)\nu_0^\pi)$$

Plug this result in, and we can obtain

$$\mathcal{L}(\pi, w, Q) = -\frac{1}{2\lambda_Q} \Big((1-\gamma)\nu_0^\pi - (I - \gamma P_*^\pi)\Lambda w \Big)^\top \Lambda^{-1} \Big((1-\gamma)(\nu_0^\pi) - (I - \gamma P_*^\pi)\Lambda w \Big) + w^\top \Lambda R - \frac{\lambda_w}{2} w^\top \Lambda w$$

Taking the derivative w.r.t. w, and set it to 0:

$$0 = \frac{1}{\lambda_Q} \Lambda (I - \gamma P^\pi) \Lambda^{-1} \Big((1-\gamma)(\nu_0^\pi) - (I - \gamma P_*^\pi) \Lambda w \Big) + \Lambda R - \lambda_w \Lambda w$$

As a result,

$$\begin{aligned}
w_{\mathcal{L}}^\pi =& \Big(\lambda_w I + \frac{1}{\lambda_Q} (I - \gamma P^\pi) \Lambda^{-1} (I - \gamma P_*^\pi) \Lambda \Big)^{-1} \Big(\frac{1}{\lambda_Q} (I - \gamma P^\pi) \Lambda^{-1} (1-\gamma) \nu_0^\pi + R \Big) \\
=& \Big(\lambda_w \lambda_Q I + (I - \gamma P^\pi) \Lambda^{-1} (I - \gamma P_*^\pi) \Lambda \Big)^{-1} \Big((I - \gamma P^\pi) \Lambda^{-1} (I - \gamma P_*^\pi) \Lambda \Lambda^{-1} (I - \gamma P_*^\pi)^{-1} (1-\gamma) \nu_0^\pi + \lambda_Q R \Big) \\
=& w^\pi + \Big(\lambda_w \lambda_Q I + (I - \gamma P^\pi) \Lambda^{-1} (I - \gamma P_*^\pi) \Lambda \Big)^{-1} \Big(\lambda_Q R - \lambda_Q \lambda_w w^\pi \Big)
\end{aligned}$$

and

$$\begin{aligned}
Q_{\mathcal{L}}^\pi =& \frac{1}{\lambda_Q} \Lambda^{-1} \Big((I - \gamma P_*^\pi) \Lambda w_{\mathcal{L}}^\pi - (1-\gamma) \nu_0^\pi \Big) \\
=& \frac{1}{\lambda_Q} \Lambda^{-1} \Big((I - \gamma P_*^\pi) \Lambda w_{\mathcal{L}}^\pi - (I - \gamma P_*^\pi) \Lambda w^\pi \Big) \\
=& \frac{1}{\lambda_Q} \Lambda^{-1} (I - \gamma P_*^\pi) \Lambda \Big(\lambda_Q \lambda_w \Lambda + \Lambda (I - \gamma P^\pi) \Lambda^{-1} (I - \gamma P_*^\pi) \Lambda \Big)^{-1} \Big(\lambda_Q \Lambda R - \lambda_Q \lambda_w \Lambda w^\pi \Big) \\
=& \Big(\lambda_w \lambda_Q (I - \gamma P_*^\pi)^{-1} \Lambda + \Lambda (I - \gamma P^\pi) \Big)^{-1} \Big(\Lambda R - \lambda_w \Lambda w^\pi \Big) \\
=& \Big(\lambda_w \lambda_Q (I - \gamma P_*^\pi)^{-1} \Lambda + \Lambda (I - \gamma P^\pi) \Big)^{-1} \Big(\Lambda (I - \gamma P^\pi) Q^\pi - \lambda_w \Lambda w^\pi \Big) \\
=& Q^\pi - \Big(\lambda_w \lambda_Q (I - \gamma P_*^\pi)^{-1} \Lambda + \Lambda (I - \gamma P^\pi) \Big)^{-1} \Big(\lambda_w \lambda_Q (I - \gamma P_*^\pi)^{-1} \Lambda Q^\pi + \lambda_w \Lambda w^\pi \Big) \\
=& Q^\pi - \Big(\lambda_w \lambda_Q I + \Lambda^{-1} (I - \gamma P_*^\pi) \Lambda (I - \gamma P^\pi) \Big)^{-1} \Big(\lambda_w \lambda_Q Q^\pi + \lambda_w (1-\gamma) \Lambda^{-1} \nu_0^\pi) \Big)
\end{aligned}$$

$\square$

Lemma B.2. *Under Assumption B:*

$$\begin{aligned}
\|w^\pi - w_{\mathcal{L}}^\pi\|_\Lambda^2 \leq & \frac{C^2 (\lambda_Q + \lambda_Q \lambda_w C)^2}{(1-\gamma)^2} \\
\|Q^\pi - Q_{\mathcal{L}}^\pi\|_\Lambda^2 \leq & \frac{C^2}{(1-\gamma)^2} (\frac{\lambda_w \lambda_Q}{1-\gamma} + \lambda_w)^2
\end{aligned}$$

where (w^π, Q^π) and $(w_{\mathcal{L}}^\pi, Q_{\mathcal{L}}^\pi)$ are defined in Theorem B.1. $\|x\|_\Lambda = x^\top \Lambda x$ denotes the norm of column vector x weighted by Λ.

Proof. From Theorem B.1, we have

$$\begin{aligned}
w_{\mathcal{L}}^\pi =& w^\pi + \Big(\lambda_w \lambda_Q I + (I - \gamma P^\pi) \Lambda^{-1} (I - \gamma P_*^\pi) \Lambda \Big)^{-1} \Big(\lambda_Q R - \lambda_Q \lambda_w w^\pi \Big) \\
Q_{\mathcal{L}}^\pi =& Q^\pi - \Big(\lambda_w \lambda_Q I + \Lambda^{-1} (I - \gamma P_*^\pi) \Lambda (I - \gamma P^\pi) \Big)^{-1} \Big(\lambda_w \lambda_Q Q^\pi + \lambda_w (1-\gamma) \Lambda^{-1} \nu_0^\pi) \Big)
\end{aligned}$$

We use $\mathbf{1} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times 1}$ to denote a vector whose all elements are 1. Then, we have

$$
\begin{aligned}
\|w^\pi - w^\pi_{\mathcal{L}}\|^2_\Lambda \leq & \|\Big(\lambda_w\lambda_Q I + (I-\gamma P^\pi)\Lambda^{-1}(I-\gamma P^\pi_*)\Lambda\Big)^{-1}\Big(\lambda_Q R - \lambda_Q\lambda_w w^\pi\Big)\|^2_\Lambda \\
= & \|\Big(\lambda_w\lambda_Q I + \Lambda^{1/2}(I-\gamma P^\pi)\Lambda^{-1}(I-\gamma P^\pi_*)\Lambda^{1/2}\Big)^{-1}\Lambda^{1/2}\Big(\lambda_Q R - \lambda_Q\lambda_w w^\pi\Big)\|^2 \\
\leq & \|\Lambda^{-1/2}(I-\gamma P^\pi_*)^{-1}\Lambda(I-\gamma P^\pi)^{-1}\Big(\lambda_Q R - \lambda_Q\lambda_w w^\pi\Big)\|^2 \\
= & \|\Lambda^{-1/2}(I-\gamma P^\pi_*)^{-1}\Lambda\widetilde{Q}^\pi\|^2 \\
\leq & \frac{(\lambda_Q + \lambda_Q\lambda_w C)^2}{(1-\gamma)^2}\|\Lambda^{-1}(I-\gamma P^\pi_*)^{-1}\Lambda\mathbf{1}\|^2_\Lambda \\
= & \frac{(\lambda_Q + \lambda_Q\lambda_w C)^2}{(1-\gamma)^2}\|\Lambda^{-1}(I-\gamma P^\pi_*)^{-1}d^\mu\|^2_\Lambda \\
= & \frac{(\lambda_Q + \lambda_Q\lambda_w C)^2}{(1-\gamma)^2}\|w^\pi_{d^\mu}\|^2_\Lambda \leq \frac{C^2(\lambda_Q + \lambda_Q\lambda_w C)^2}{(1-\gamma)^2}
\end{aligned}
$$

where in the second inequality, we use Lemma A.4; in the second equality, we use $\widetilde{Q}^\pi$ to denote the Q function after replacing true rewards with $\lambda_Q R - \lambda_Q\lambda_w w^\pi$; in the third inequality, we use Lemma A.5 and the result that $|\lambda_Q R - \lambda_Q\lambda_w w^\pi| \leq \lambda_Q + \lambda_Q\lambda_w C$ given Assumption B; in the last inequality, we use Assumption B again. Similarly,

$$
\begin{aligned}
\|Q^\pi - Q^\pi_{\mathcal{L}}\|^2_\Lambda \leq & \|\Big(\lambda_w\lambda_Q I + \Lambda^{-1}(I-\gamma P^\pi_*)\Lambda(I-\gamma P^\pi)\Big)^{-1}\Big(\lambda_w\lambda_Q Q^\pi + \lambda_w(1-\gamma)\Lambda^{-1}\nu_0^\pi)\Big)\|^2_\Lambda \\
= & \|\Big(\lambda_Q\lambda_w I + \Lambda^{-1/2}(I-\gamma P^\pi_*)\Lambda(I-\gamma P^\pi)\Lambda^{-1/2}\Big)^{-1}\Lambda^{1/2}\Big(\lambda_Q\lambda_w Q^\pi + \lambda_w(1-\gamma)\Lambda^{-1}\nu_0^\pi)\Big)\|^2 \\
\leq & \|\Lambda^{1/2}(I-\gamma P^\pi)^{-1}\Lambda^{-1}(I-\gamma P^\pi_*)^{-1}\Big(\lambda_w\lambda_Q\Lambda Q^\pi + \lambda_w(1-\gamma)\nu_0^\pi)\Big)\|^2 \\
= & \|\lambda_w\lambda_Q\Lambda^{1/2}(I-\gamma P^\pi)^{-1}\Lambda^{-1}(I-\gamma P^\pi_*)^{-1}\Lambda Q^\pi + \lambda_w\Lambda^{1/2}(I-\gamma P^\pi)^{-1}w^\pi)\|^2 \\
\leq & \|\frac{\lambda_w\lambda_Q}{1-\gamma}\Lambda^{1/2}(I-\gamma P^\pi)^{-1}\Lambda^{-1}(I-\gamma P^\pi_*)^{-1}\Lambda\mathbf{1} + \lambda_w\Lambda^{1/2}(I-\gamma P^\pi)^{-1}w^\pi)\|^2 \\
\leq & \|(I-\gamma P^\pi)^{-1}(\frac{\lambda_w\lambda_Q}{1-\gamma}w^\pi_{d^\mu} + \lambda_w w^\pi)\|^2_\Lambda \\
\leq & \frac{C^2}{(1-\gamma)^2}(\frac{\lambda_w\lambda_Q}{1-\gamma} + \lambda_w)^2
\end{aligned}
$$

where in the last but third inequality, we use Lemma A.5 and the fact that w^π is also non-negative. $\square$

Lemma B.3. *Under Assumption B, for arbitrary function $f(s,a)$,*

$$(1-\gamma)\mathbb{E}_{s_0\sim\nu_0,a_0\sim\pi}[f(s_0,a_0)] + \gamma\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi}[w^\pi(s,a)f(s',a')] = \mathbb{E}_{d^\mu}[w^\pi(s,a)f(s,a)] \tag{15}$$

$$\gamma\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi}[f^2(s',a')] \leq \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^\pi_{d^\mu}}[f^2(s,a)] \leq \frac{C}{1-\gamma}\mathbb{E}_{s,a\sim d^\mu}[f^2(s,a)] \tag{16}$$

where $d^\pi_{d^\mu} := (1-\gamma)\mathbb{E}_{\tau\sim\pi,s_0,a_0\sim d^\pi(\cdot,\cdot)}[\sum_{t=0}^\infty \gamma^t p(s_t=s,a_t=a)]$ is the normalized discounted state-action occupancy by treating d^μ as initial distribution; $s,a,s'\sim d^\mu, a'\sim\pi$ is a short note of $s,a\sim d^\mu, s'\sim P(s'|s,a), a'\sim\pi(\cdot|s')$.

Proof. Eq.(15) can be proved by the equation:

$$d^\pi(s,a) = (1-\gamma)\nu_0(s)\pi(a|s) + \gamma\sum_{s',a'}p(s|s',a')d^\pi(s',a')\pi(a|s)$$

For Eq.(16), the first step is because $\gamma\sum_{s',a'}d^\mu(s',a')p(s|s',a')\pi(a|s) \leq \frac{1}{1-\gamma}d^\pi_{d^\mu}(s,a)$, and the second step is the result of Assumption B. $\square$

Theorem 2.6. *[Bias] Under Assumption A, B, C, given arbitrary $\theta \in \Theta$, we have*

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

where ε_{data} is defined in Definition 2.3, and

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

($\varepsilon_\mathcal{W}$ and $\varepsilon_\mathcal{Q}$ defined in Prop. 2.5)

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

Proof. Firstly, by applying the triangle inequality:

$$\begin{aligned}\|\nabla_\theta \max_{w\in\mathcal{W}} \min_{Q\in\mathcal{Q}} \mathcal{L}^D(\pi_\theta, w, Q) - \nabla_\theta J(\pi_\theta)\| \le & \underbrace{\|\nabla_\theta \max_{w\in\mathcal{W}} \min_{Q\in\mathcal{Q}} \mathcal{L}^D(\pi_\theta, w, Q) - \nabla_\theta \max_{w\in\mathcal{W}} \min_{Q\in\mathcal{Q}} \mathcal{L}(\pi_\theta, w, Q)\|}_{Bounded\ in\ Assumption\ 2.3} \\ & + \underbrace{\|\nabla_\theta \max_{w} \min_{Q} \mathcal{L}(\pi_\theta, w, Q) - \nabla_\theta \max_{w\in\mathcal{W}} \min_{Q\in\mathcal{Q}} \mathcal{L}(\pi_\theta, w, Q)\|}_{t_1} \\ & + \underbrace{\|\nabla_\theta J(\pi_\theta) - \nabla_\theta \max_{w} \min_{Q} \mathcal{L}(\pi_\theta, w, Q)\|}_{t_2}\end{aligned}$$

where we use $\max_w \min_Q$ as a short note of $\max_{w\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \min_{Q\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}}$.

In the following, we again use $(w_\mathcal{L}^{\pi_\theta}, Q_\mathcal{L}^{\pi_\theta})$ to denote the saddle point of $\mathcal{L}(\pi_\theta, w, Q)$ without any constraint on w and Q, and use (w_μ^*, Q_μ^*) to denote the saddle point of $\mathcal{L}(\pi_\theta, w, Q)$. Next, we upper bound t_1 and t_2 one by one. For simplicity, we use $s, a, s' \sim d^\mu, a' \sim \pi_\theta$ as a short note of $s, a \sim d^\mu, s' \sim P(s'|s, a), a' \sim \pi_\theta(\cdot|s')$.

Upper bound t_1 With misspecification Definition 2.4, we can easily bound t_1:

$$\begin{aligned} t_1 =& \|\nabla_\theta \mathcal{L}(\pi_\theta, w_\mu^*, Q_\mu^*) - \nabla_\theta \mathcal{L}(\pi_\theta, w_\mathcal{L}^{\pi_\theta}, Q_\mathcal{L}^{\pi_\theta})\| \\ \le& \frac{1}{1-\gamma}\|(1-\gamma)\mathbb{E}_{\nu_0^{\pi_\theta}}[\Big(Q_\mu^*(s_0, a_0) - Q_\mathcal{L}^{\pi_\theta}(s_0, a_0)\Big)\nabla_\theta \log\pi_\theta(a_0|s_0)]\| \\ & + \frac{\gamma}{1-\gamma}\|\mathbb{E}_{s,a,s'\sim d^\mu, a'\sim\pi}[w_\mu^*(s, a)\Big(Q_\mu^*(s', a') - Q_\mathcal{L}^{\pi_\theta}(s', a')\Big)\nabla_\theta \log\pi(a'|s')]\| \\ & + \frac{\gamma}{1-\gamma}\|\mathbb{E}_{s,a,s'\sim d^\mu, a'\sim\pi}[(w_\mu^*(s, a) - w_\mathcal{L}^{\pi_\theta}(s, a))\Big(Q_\mu^*(s', a') - Q_\mathcal{L}^{\pi_\theta}(s', a')\Big)\nabla_\theta \log\pi(a'|s')]\| \\ & + \frac{\gamma}{1-\gamma}\|\mathbb{E}_{s,a,s'\sim d^\mu, a'\sim\pi_\theta}[(w_\mu^*(s, a) - w_\mathcal{L}^{\pi_\theta}(s, a))Q_\mu^*(s', a')\nabla_\theta \log\pi(a'|s')]\| \\ \le& \frac{CG}{1-\gamma}\mathbb{E}_{d^\mu}[|Q_\mu^*(s, a) - Q_\mathcal{L}^{\pi_\theta}(s, a)|] + \frac{\gamma C_\mathcal{W} G}{1-\gamma}\mathbb{E}_{s,a,s'\sim d^\mu, a'\sim\pi_\theta}[|Q_\mu^*(s', a') - Q_\mathcal{L}^{\pi_\theta}(s', a')|] \\ & \qquad ((1-\gamma)\nu_0^\pi(s, a) \le d^\pi(s, a) \le Cd^\mu(s, a)) \\ & + \frac{\gamma G}{1-\gamma}\mathbb{E}_{s,a,s'\sim d^\mu, a'\sim\pi_\theta}[|(w_\mu^*(s, a) - w_\mathcal{L}^{\pi_\theta}(s, a))\Big(Q_\mu^*(s', a') - Q_\mathcal{L}^{\pi_\theta}(s', a')\Big)|] \\ & + \frac{\gamma C_\mathcal{Q} G}{1-\gamma}\mathbb{E}_{s,a,s'\sim d^\mu, a'\sim\pi_\theta}[|w_\mu^*(s, a) - w_\mathcal{L}^{\pi_\theta}(s, a)|] \\ \le& \frac{CG}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|Q_\mu^*(s, a) - Q_\mathcal{L}^{\pi_\theta}(s, a)|^2]} + \frac{\gamma C_\mathcal{W} G}{1-\gamma}\sqrt{\mathbb{E}_{s,a,s'\sim d^\mu, a'\sim\pi_\theta}[|Q_\mu^*(s', a') - Q_\mathcal{L}^{\pi_\theta}(s', a')|^2]} \\ & + \frac{\gamma G}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|w_\mathcal{L}^{\pi_\theta}(s, a) - w_\mu^*(s, a)|^2]\mathbb{E}_{s,a,s'\sim d^\mu, a'\sim\pi_\theta}[|Q^{\pi_\theta}(s', a') - Q_\mathcal{L}^{\pi_\theta}(s', a')|^2|]} \\ & + \frac{\gamma C_\mathcal{Q} G}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|w_\mu^*(s, a) - w_\mathcal{L}^{\pi_\theta}(s, a))|^2]} \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{CG}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|Q^*_\mu(s,a)-Q^{\pi_\theta}_{\mathcal{L}}(s,a)|^2]} + \frac{C_{\mathcal{W}}G}{1-\gamma}\sqrt{\frac{\gamma C}{1-\gamma}\mathbb{E}_{d^\mu}[|Q^*_\mu(s,a)-Q^{\pi_\theta}_{\mathcal{L}}(s,a)|^2]} \\
&\quad + \frac{G}{1-\gamma}\sqrt{\frac{\gamma C}{1-\gamma}\mathbb{E}_{d^\mu}[|w^{\pi_\theta}_{\mathcal{L}}(s,a)-w^*_\mu(s,a)|^2]\mathbb{E}_{d^\mu}[|Q^{\pi_\theta}(s,a)-Q^{\pi_\theta}_{\mathcal{L}}(s,a)|^2|]} \\
&\quad + \frac{\gamma C_{\mathcal{Q}}G}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|w^*_\mu(s,a)-w^{\pi_\theta}_{\mathcal{L}}(s,a))|^2]} \\
&\leq \frac{G}{1-\gamma}\Big(C\sqrt{\varepsilon_{\mathcal{Q}}} + C_{\mathcal{W}}\sqrt{\frac{\gamma\varepsilon_{\mathcal{Q}}C}{1-\gamma}} + \sqrt{\frac{\gamma\varepsilon_{\mathcal{Q}}\varepsilon_{\mathcal{W}}C}{1-\gamma}} + \gamma C_{\mathcal{Q}}\sqrt{\varepsilon_{\mathcal{W}}}\Big)
\end{aligned}
$$

In the last equation, we first use Eq.(16) in Lemma B.3, and then apply Proposition 2.5.

Upper bound t_2 Similarly, we can give a bound for t_2:

$$
\begin{aligned}
t_2 =& \|\nabla_\theta J(\pi_\theta) - \nabla_\theta \mathcal{L}(\pi_\theta, w^{\pi_\theta}_{\mathcal{L}}, Q^{\pi_\theta}_{\mathcal{L}}))\| \\
\leq& \frac{1}{1-\gamma}\|(1-\gamma)\mathbb{E}_{\nu_0^{\pi_\theta}}[\Big(Q^{\pi_\theta}(s_0,a_0)-Q^{\pi_\theta}_{\mathcal{L}}(s_0,a_0)\Big)\nabla_\theta \log\pi_\theta(a_0|s_0)] \\
&+ \gamma\mathbb{E}_{d^\mu}[w^{\pi_\theta}(s,a)\Big(Q^{\pi_\theta}(s',a')-Q^{\pi_\theta}_{\mathcal{L}}(s',a')\Big)\nabla_\theta\log\pi(a'|s')]\| \\
&+ \frac{\gamma}{1-\gamma}\|\mathbb{E}_{d^\mu}[(w^{\pi_\theta}(s,a)-w^{\pi_\theta}_{\mathcal{L}}(s,a))\Big(Q^{\pi_\theta}(s',a')-Q^{\pi_\theta}_{\mathcal{L}}(s',a')\Big)\nabla_\theta\log\pi(a'|s')]\| \\
&+ \frac{\gamma}{1-\gamma}\|\mathbb{E}_{d^\mu}[(w^{\pi_\theta}(s,a)-w^{\pi_\theta}_{\mathcal{L}}(s,a))Q^{\pi_\theta}(s',a')\nabla_\theta\log\pi(a'|s')]\| \\
=& \frac{1}{1-\gamma}\|\mathbb{E}_{d^\mu}[w^{\pi_\theta}(s,a)\Big(Q^{\pi_\theta}(s,a)-Q^{\pi_\theta}_{\mathcal{L}}(s,a)\Big)\nabla_\theta\log\pi(a|s)]\| \qquad \text{(Eq.(15) in Lemma B.3)} \\
&+ \frac{\gamma}{1-\gamma}\|\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[(w^{\pi_\theta}(s,a)-w^{\pi_\theta}_{\mathcal{L}}(s,a))\Big(Q^{\pi_\theta}(s',a')-Q^{\pi_\theta}_{\mathcal{L}}(s',a')\Big)\nabla_\theta\log\pi(a'|s')]\| \\
&+ \frac{\gamma}{1-\gamma}\|\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[(w^{\pi_\theta}(s,a)-w^{\pi_\theta}_{\mathcal{L}}(s,a))Q^{\pi_\theta}(s',a')\nabla_\theta\log\pi(a'|s')]\| \\
\leq& \frac{CG}{1-\gamma}\mathbb{E}_{d^\mu}[|Q^{\pi_\theta}(s,a)-Q^{\pi_\theta}_{\mathcal{L}}(s,a)|] \\
&+ \frac{\gamma G}{1-\gamma}\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[|(w^{\pi_\theta}(s,a)-w^{\pi_\theta}_{\mathcal{L}}(s,a))\Big(Q^{\pi_\theta}(s',a')-Q^{\pi_\theta}_{\mathcal{L}}(s',a')\Big)|] \\
&+ \frac{\gamma G}{(1-\gamma)^2}\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[|w^{\pi_\theta}(s,a)-w^{\pi_\theta}_{\mathcal{L}}(s,a)|] \\
\leq& \frac{CG}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|Q^{\pi_\theta}-Q^{\pi_\theta}_{\mathcal{L}}|^2]} + \frac{\gamma G}{(1-\gamma)^2}\sqrt{\mathbb{E}_{d^\mu}[|(w^{\pi_\theta}(s,a)-w^{\pi_\theta}_{\mathcal{L}}(s,a)|^2]} \\
&+ \frac{\gamma G}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|w^{\pi_\theta}_{\mathcal{L}}(s,a)-w^{\pi_\theta}(s,a)|^2]\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[|Q^{\pi_\theta}(s',a')-Q^{\pi_\theta}_{\mathcal{L}}(s',a')|^2|]} \\
\leq& \frac{CG}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|Q^{\pi_\theta}-Q^{\pi_\theta}_{\mathcal{L}}|^2]} + \frac{\gamma G}{(1-\gamma)^2}\sqrt{\mathbb{E}_{d^\mu}[|(w^{\pi_\theta}(s,a)-w^{\pi_\theta}_{\mathcal{L}}(s,a)|^2]} \\
&+ \frac{G}{1-\gamma}\sqrt{\frac{\gamma C}{1-\gamma}\mathbb{E}_{d^\mu}[|w^{\pi_\theta}_{\mathcal{L}}(s,a)-w^{\pi_\theta}(s,a)|^2]\mathbb{E}_{d^\mu}[|Q^{\pi_\theta}(s,a)-Q^{\pi_\theta}_{\mathcal{L}}(s,a)|^2|]} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Eq.16 in Lemma B.3)} \\
\leq& \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)^2} + \frac{C^2(\lambda_Q+\lambda_Q\lambda_w C)}{(1-\gamma)^2}(\frac{\lambda_w\lambda_Q}{1-\gamma}+\lambda_w)\sqrt{\frac{\gamma C}{1-\gamma}}\Big)
\end{aligned}
$$

$\square$

B.1 Importance of the Regularization

Here we want to highlight that the additional regularization terms on Q and w are crucial. For example, suppose $Q^\pi \in \mathcal{Q}$ and $w^\pi \in \mathcal{W}$ for some policy π, if $\lambda_w = \lambda_Q = 0$, we have

$$\forall \zeta \in Z, \quad \nabla_\zeta \mathcal{L}^D(\pi_\theta, w_\zeta, Q^\pi) = \nabla_\zeta (1-\gamma)\mathbb{E}_{s_0\sim\nu_0^D}[Q^\pi(s_0,\pi)] = 0$$
$$\forall \xi \in \Xi, \quad \nabla_\xi \mathcal{L}^D(\pi_\theta, w^\pi, Q_\xi) = \nabla_\xi \mathbb{E}_{w^{\pi/\mu}}[r] = 0$$

which means $Q = Q^\pi$ (or $w = w^{\pi/\mu}$) can result in that the gradient w.r.t. ζ (or ξ) vanishes to 0, and that the estimation for w^π (or Q^π) can be arbitrarily worse. Moreover, $\mathcal{L}^D$ is no longer a strongly-concave-strongly-convex function.

C Missing Examples and Proofs in Section 3

C.1 Missing proofs

Theorem 3.1. *[Equivalence Between Stationary Points] Under Assumption A, C and D, suppose an Algorithm provides us one stationary point $(\theta_T, \zeta_T, \xi_T)$ 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.*

$$\begin{aligned}\mathbb{E}[\|\nabla_{\theta,\zeta}\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\|] :=&\mathbb{E}[\|\nabla_\theta\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\| + \|\nabla_\zeta\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\|] \\ \le& \frac{\varepsilon}{(\kappa_\xi+1)^2}\end{aligned} \tag{6}$$

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

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

Proof. First of all, as a results of Assumption A, C and D, we know there must exists $\zeta \in Z$, s.t. if $\zeta_T = \zeta$, then ζ_T can satisfy Eq.(6). Therefore, it's possible for an algorithm to return us a (θ_T, ζ_T) satisfy Eq.(6).

Next, suppose we already have Eq.(6), it implies that

$$\max\{\mathbb{E}[\|\nabla_\theta\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\|], \mathbb{E}[\|\nabla_\zeta\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\|]\} \le \frac{\varepsilon}{(\kappa_\xi+1)^2} \tag{17}$$

We can upper bounded $\mathbb{E}[\|\nabla_\theta J(\pi_{\theta_T})\|]$ with the triangle inequality:

$$\begin{aligned}\mathbb{E}[\|\nabla_\theta J(\pi_{\theta_T})\|] \le& \underbrace{\mathbb{E}[\|\nabla_\theta\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\|]}_{Bounded\ in\ Eq.(17)} + \mathbb{E}[\|\nabla_\theta\mathcal{L}^D(\theta_T,\zeta^*,\xi^*) - \nabla_\theta\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T)))\|] \\ &+ \underbrace{\mathbb{E}[\|\nabla_\theta\mathcal{L}^D(\theta_T,\zeta^*,\xi^*) - \nabla_\theta J(\pi_{\theta_T})\|]}_{Bounded\ in\ Theorem 2.6} \\ \le& \frac{\varepsilon}{(\kappa_\xi+1)^2} + \varepsilon_{func} + \varepsilon_{reg} + \varepsilon_{data} \\ &+ \mathbb{E}[\|\nabla_\theta\mathcal{L}^D(\theta_T,\zeta^*,\xi^*) - \nabla_\theta\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T)))\|]\end{aligned}$$

where we use ζ^*, ξ^* to denote the saddle-point of $\max_{\zeta\in Z}\min_{\xi\in\Xi}\mathcal{L}^D(\theta_T,\zeta,\xi)$; in the last inequality we use Eq.17 and Theorem 2.6.

Next, we try to bound the last term. According to the definition, ζ^* is also the maximum of function $\Phi_{\theta_T}(\cdot) = \min_{\xi\in\Xi}\mathcal{L}^D(\theta_T,\cdot,\xi)$ defined in Lemma A.1. Applying Property (2) in Lemma A.1, (10) in Lemma A.3 and inequality (17), we obtain that

$$\|\zeta_T - \zeta^*\| \le \frac{1}{\mu_\zeta}\|\nabla_\zeta\mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\| \le \frac{\varepsilon}{\mu_\zeta(\kappa_\xi+1)^2}$$

Then we can bound:

$$\begin{aligned}&\|\nabla_\theta \mathcal{L}^D(\theta_T,\zeta^*,\xi^*) - \nabla_\theta \mathcal{L}^D(\theta_T,\zeta_T,\phi_{\theta_T}(\zeta_T))\|\\ \le & L\|\zeta_T-\zeta^*\| + L\|\xi^* - \phi_{\theta_T}(\zeta_T))\| = L\|\zeta_T-\zeta^*\| + L\|\phi_{\theta_T}(\zeta^*) - \phi_{\theta_T}(\zeta_T))\|\\ \le & (L+L\kappa_\xi)\|\zeta_T-\zeta^*\| \le \frac{\varepsilon\kappa_\xi}{1+\kappa_\xi}\end{aligned}$$

where in the first inequality we use the smoothness Assumption A, and in the second inequality we use (1) in Lemma A.1. As a result,

$$\begin{aligned}\mathbb{E}[\|\nabla_\theta J(\pi_{\theta_T})\|] \le & \frac{\varepsilon}{(\kappa_\xi+1)^2} + \frac{\varepsilon\kappa_\xi}{1+\kappa_\xi} + \varepsilon_{func} + \varepsilon_{reg} + \varepsilon_{data}\\ \le & \varepsilon + \varepsilon_{func} + \varepsilon_{reg} + \varepsilon_{data}\end{aligned}$$

□

C.2 Algorithm Examples

We first introduce a useful assumption:

Assumption G (Diameter; Replace Assump. F)**.** We use Ξ to denote the set of parameters ξ, we assume Ξ is a convex and bounded set with a diameter $d > 0$.

C.2.1 Example 1: Stochastic Gradient Descent Ascent [20]

Algorithm 2: Direct SGDA

```
Initialize θ_0, ζ_0, ξ_0
for t = 0, 1, 2, ...T do
    Sample N (s, a, r, s') ~ d^D, a' ~ π_{θ_{t+1}}(s') tuples and computing:
    θ_{t+1} ← θ_t + η_θ ∇̂_θ L^D(θ_t, ζ_t, ξ_t)
    ζ_{t+1} ← ζ_t + η_ζ ∇̂_ζ L^D(θ_t, ζ_t, ξ_t)
    ξ_{t+1} ← P_ξ(ξ_t − η_ξ ∇̂_ξ L^D(θ_t, ζ_t, ξ_t)) // P_ξ is the projection operator.
end
```

Adapting from Theorem 4.5 and Proposition 4.11 in [20], we have the following theorem

Theorem C.1. *Define $\Delta = \max_{\theta,\zeta}\min_{\xi\in\Xi}\mathcal{L}^D(\theta,\zeta,\xi) - \min_{\xi\in\Xi}\mathcal{L}^D(\theta_0,\zeta_0,\xi)$. Under Assumption A, C, E and G, with step sizes $\eta_\xi = \Theta(1/L), \eta_\zeta = \eta_\theta = \Theta(1/\kappa_\xi^2 L)$ and batch size $N = \Theta(\max\{1, \kappa_\xi(\kappa_\xi+1)^4\sigma^2\varepsilon^{-2}\})$, if $T = O(\frac{(\kappa_\xi+1)^4(\kappa_\xi^2 L\Delta + \kappa_\xi^2 L^2 D^2)}{\varepsilon^2}$, Algorithm 1 will return us (θ_T,ζ_T,ξ_T) satisfying the ε-stationary condition in Eq.*(6)*.*

Corollary C.2. *Under the same assumption as Theorem C.1, after consuming $O(\frac{(\kappa_\xi+1)^4(\kappa_\xi^2 L\Delta+\kappa_\xi^2 L^2 D^2)}{\varepsilon^2}\max\{1, \frac{(\kappa_\xi+1)^4(\kappa_\xi^2\sigma^2)}{\varepsilon^2}\})$ steps, Algorithm 2 will provide us a policy π_{θ_T} satisfying*

$$\mathbb{E}[\|J(\pi_{\theta_T})\|] \le \varepsilon + \varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg} \tag{18}$$

where $\Delta = \max_{\theta,\zeta}\min_{\xi\in\Xi}\mathcal{L}^D(\theta,\zeta,\xi) - \min_{\xi\in\Xi}\mathcal{L}^D(\theta_0,\zeta_0,\xi)$; ε_{data} is defined in Assumption 2.3, and ε_{func} and ε_{reg} are defined in Theorem 2.6.

C.3 Example 2: Stochastic Recursive Gradient Descent Ascent [8]

In [8], the author presented another algorithm has better dependence on ε. Similarly, we can adapt their algorithm and we ignore the details here.

D Missing details for Algorithm 1

D.1 The practicality of Oracle in Definition 4.1

In some previous literatures related to stochastic optimization on strongly-convex-strongly-concave problems, some algorithms can achieve exponential convergence rate. For example, in the Theorem

2 of [22], the author proved that the distance between the variables and the saddle point decays exponentially. Although the SVRE algorithm in [22] relies on the finite-sum structure, we may adapt it to our setting by dividing our entire dataset D to n sub datasets $\{D_1, D_2, ..., D_n\}$. Since $\mathcal{L}^D = \sum_{i=1}^n \mathcal{L}^{D_i}$, we have the finite-sum structure and we can run SVRE with the same convergence guarantee if some necessary assumptions are satisfied.

However, one of the drawback of such direct adaption is that, we may need to process the entire dataset D (see Line 3 and 4 in Algorithm 1 of [22]), which is quite expensive sometimes. Besides, the division of D need to be done carefully, and the additional assumptions we require can be very strict in some cases. It would be an interesting question to design a new algorithm to get rid of these cons, and we leave it to the future work.

D.2 Missing Proofs

In the following, we will use $\mathcal{L}_t^D$, $\mathcal{L}_t^B$ and $\mathcal{L}_t^{D*}$ as shortnotes of $\mathcal{L}^D(\theta_t, \zeta_t, \xi_t)$, $\mathcal{L}^B(\theta_t, \zeta_t, \xi_t)$ and $\mathcal{L}^D(\theta_t, \zeta_t^*, \xi_t^*)$, where ζ_t^*, ξ_t^* is the only one saddle point of $\mathcal{L}^D(\theta_t, \zeta, \xi)$. Besides, we use $\nabla_\theta \mathcal{L}_t^D$ and $\nabla_\theta \mathcal{L}_t^B$ as a shortnote of the gradient averaged over d^D and the gradient averaged over batch, respectively.

Lemma D.1. *Suppose we have two empirical gradient estimator $\nabla_\theta \mathcal{L}_{t+1}^B$ and $\nabla_\theta \mathcal{L}_t^B$ built with the same batch data B, under Assumption A, we have:*

$$
\begin{aligned}
&\mathbb{E}[\|\nabla_\theta \mathcal{L}_{t+1}^B - \nabla_\theta \mathcal{L}_t^B\|^2] \\
\le &\frac{3}{|B|}\Big(G^2 L_w^2 C_\mathcal{Q}^2 \mathbb{E}[\|\zeta_{t+1} - \zeta_t\|^2] + G^2 L_Q^2 C_\mathcal{W}^2 \mathbb{E}[\|\xi_{t+1} - \xi_t\|^2] + H^2 C_\mathcal{Q}^2 C_\mathcal{W}^2 \mathbb{E}[\|\theta_{t+1} - \theta_t\|^2]\Big)
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&\mathbb{E}[\|\nabla_\theta \mathcal{L}_{t+1}^B - \nabla_\theta \mathcal{L}_t^B\|^2] \\
\le &\frac{3}{|B|^2}\mathbb{E}\Big[\sum_B \|(1-\gamma)\mathbb{I}[s \in S_0]\Big(Q_{t+1}(s, a_0) - Q_t(s, a_0)\Big)\nabla_\theta \log \pi_t(a_0|s) \\
&\qquad + \gamma w_t(s,a)\Big(Q_{t+1}(s', a') - Q_t(s', a')\Big)\nabla_\theta \log \pi_{t+1}(a'|s')\|^2 \\
&\qquad + \|(1-\gamma)\mathbb{I}[s \in S_0]Q_{t+1}(s, a_0)\Big(\nabla_\theta \log \pi_{t+1}(a_0|s) - \nabla_\theta \log \pi_t(a_0|s)\Big) \\
&\qquad + \gamma w_t(s,a)Q_t(s',a')\Big(\nabla_\theta \log \pi_{t+1}(a'|s') - \nabla_\theta \log \pi_t(a'|s')\Big)\|^2 \\
&\qquad + \|\gamma(w_{t+1}(s,a) - w_t(s,a))Q_{t+1}(s',a')\nabla_\theta \log \pi_{t+1}(a'|s')\|^2\Big)\Big] \\
\le &\frac{3}{|B|}\Big(\gamma^2 G^2 L_w^2 C_\mathcal{Q}^2 \mathbb{E}[\|\zeta_{t+1} - \zeta_t\|^2] + G^2 L_Q^2 \Big((1-\gamma) + \gamma C_\mathcal{W}\Big)^2 \mathbb{E}[\|\xi_{t+1} - \xi_t\|^2] \\
&+ H^2 C_\mathcal{Q}^2 \Big((1-\gamma) + \gamma C_\mathcal{W}\Big)^2 \mathbb{E}[\|\theta_{t+1} - \theta_t\|^2]\Big) \\
\le &\frac{3}{|B|}\Big(G^2 L_w^2 C_\mathcal{Q}^2 \mathbb{E}[\|\zeta_{t+1} - \zeta_t\|^2] + G^2 L_Q^2 C_\mathcal{W}^2 \mathbb{E}[\|\xi_{t+1} - \xi_t\|^2] + H^2 C_\mathcal{Q}^2 C_\mathcal{W}^2 \mathbb{E}[\|\theta_{t+1} - \theta_t\|^2]\Big)
\end{aligned}
$$

where in the first inequality, we use Young's inequality; in the second one we use Assumption A; in the last one, we use $1 \le C_\mathcal{W}$. □

Lemma D.2. *Under Assumption A, C and D, consider $\pi_{\theta_1}, \pi_{\theta_2}$ parameterized by $\theta_1, \theta_2 \in \Theta$. Denote (ζ_1^*, ξ_1^*) and (ζ_2^*, ξ_2^*) as the saddle-point of $\max_{\zeta \in Z} \min_{\xi \in \Xi} \mathcal{L}^D(\theta_1, \zeta, \xi)$ and $\max_{\zeta \in Z} \min_{\xi \in \Xi} \mathcal{L}^D(\theta_2, \zeta, \xi)$ respectively, then we have*

$$
\begin{aligned}
\|\zeta_1^* - \zeta_2^*\| \le \kappa_\mu(\kappa_\xi + 1)\|\theta_1 - \theta_2\| \\
\|\xi_1^* - \xi_2^*\| \le \kappa_\xi(\kappa_\mu + 1)\|\theta_1 - \theta_2\|
\end{aligned}
$$

Proof. With Assumption A and Assumption D, we have

$$\|\nabla_\zeta \mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| = \|\nabla_\zeta \mathcal{L}^D(\theta_1, \zeta_1^*, \xi_1^*) - \nabla_\zeta \mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| \le L\|\theta_1 - \theta_2\| \tag{19}$$

$$\|\nabla_\xi \mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| = \|\nabla_\xi \mathcal{L}^D(\theta_1, \zeta_1^*, \xi_1^*) - \nabla_\xi \mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| \le L\|\theta_1 - \theta_2\| \tag{20}$$

Recall in Lemma A.1, we know $\Phi_{\theta_2}(\zeta)$ should be a μ_ζ-strongly-concave function. Then, we have

$$\begin{aligned}
\|\zeta_1^* - \zeta_2^*\| \leq & \frac{1}{\mu_\zeta}\|\nabla_\zeta \Phi_{\theta_2}(\zeta_1^*)\| = \frac{1}{\mu_\zeta}\|\nabla_\zeta \mathcal{L}^D(\theta_2, \zeta_1^*, \phi_{\theta_2}(\zeta_1^*))\| \\
\leq & \frac{1}{\mu_\zeta}\|\nabla_\zeta \mathcal{L}^D(\theta_2, \zeta_1^*, \phi_{\theta_2}(\zeta_1^*)) - \nabla_\zeta \mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| + \frac{1}{\mu_\zeta}\|\nabla_\zeta \mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*))\| \\
\leq & \frac{1}{\mu_\zeta}\|\nabla_\zeta \mathcal{L}^D(\theta_2, \zeta_1^*, \phi_{\theta_2}(\zeta_1^*)) - \nabla_\zeta \mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| + \frac{L}{\mu_\zeta}\|\theta_1 - \theta_2\| \\
\leq & \frac{L}{\mu_\zeta}\|\phi_{\theta_2}(\zeta_1^*) - \xi_1^*\| + \frac{L}{\mu_\zeta}\|\theta_1 - \theta_2\| \\
\leq & \frac{L}{\mu_\zeta \mu_\xi}\|\nabla_\xi \mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| + \frac{L}{\mu_\zeta}\|\theta_1 - \theta_2\| \\
\leq & \kappa_\mu(\kappa_\xi + 1)\|\theta_1 - \theta_2\|
\end{aligned}$$

where in the first step, we use Lemma A.3; in the fourth inequality, we use Assumption A; in the fifth inequality, we use the Assumption C that, given θ_2, ζ_1^*, $\mathcal{L}^D(\theta_2, \zeta_1^*, \xi)$ is μ_ξ-strongly-convex w.r.t. ξ and $\phi_{\theta_2}(\zeta_1^*)$ is the optimum of it; in the last inequality, we use Eq.(19) again.

We can give a similarly discussion for $\|\xi_1^* - \xi_2^*\|$:

$$\begin{aligned}
\|\xi_1^* - \xi_2^*\| \leq & \frac{1}{\mu_\xi}\|\nabla_\xi \Psi_{\theta_2}(\xi_1^*)\| = \frac{1}{\mu_\xi}\|\nabla_\xi \mathcal{L}^D(\theta_2, \psi_{\theta_2}(\xi_1^*), \xi_1^*)\| \\
\leq & \frac{1}{\mu_\xi}\|\nabla_\xi \mathcal{L}^D(\theta_2, \psi_{\theta_2}(\xi_1^*), \xi_1^*) - \nabla_\xi \mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| + \frac{1}{\mu_\xi}\|\nabla_\xi \mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*))\| \\
\leq & \frac{1}{\mu_\xi}\|\nabla_\xi \mathcal{L}^D(\theta_2, \psi_{\theta_2}(\xi_1^*), \xi_1^*) - \nabla_\xi \mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| + \frac{L}{\mu_\xi}\|\theta_1 - \theta_2\| \\
\leq & \frac{L}{\mu_\xi}\|\zeta_1^* - \psi_{\theta_2}(\xi_1^*)\| + \frac{L}{\mu_\xi}\|\theta_1 - \theta_2\| \\
\leq & \frac{L}{\mu_\xi \mu_\zeta}\|\nabla_\zeta \mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| + \frac{L}{\mu_\xi}\|\theta_1 - \theta_2\| \\
\leq & \kappa_\xi(\kappa_\mu + 1)\|\theta_1 - \theta_2\|
\end{aligned}$$

□

Lemma D.3 (Relate the shift of ζ_t and ξ_t with θ_t)**.** *We consider the Assumptions A, C, F and D. Denote $(\theta_t, \zeta_t, \xi_t)$ as the parameter value at the beginning at the step t in Algorithm 1, and denote $(\zeta_t^*, \xi_t^*) \in Z \times \Xi$ as the only saddle point for $\mathcal{L}^D(\theta_t, \zeta, \xi)$ given θ_t. Recall the Oracle in Definition 4.1 that, for arbitrary t iteration, it will return us ζ_{t+1}, ξ_{t+1} satisfying*

$$\mathbb{E}[\|\zeta_{t+1} - \zeta_{t+1}^*\|^2 + \|\xi_{t+1} - \xi_{t+1}^*\|^2] \leq \frac{\beta}{2}\mathbb{E}[\|\zeta_t - \zeta_{t+1}^*\|^2 + \|\xi_t - \xi_{t+1}^*\|^2]$$

where $0 < \beta/2 \leq 1$. Then, we have:

$$\mathbb{E}[\|\zeta_{t+1} - \zeta_t\|^2 + \|\xi_{t+1} - \xi_t\|^2] \leq 6\beta^{t+1}d^2 + 6\eta_\theta^2 C_{\zeta,\mu}\sum_{\tau=0}^{t}\beta^{t-\tau}\mathbb{E}[\|g_\theta^\tau\|^2]$$

where d is the diameter defined in Assumption F, and $C_{\zeta,\mu}$ is a short note of $\kappa_\mu^2(\kappa_\xi+1)^2+\kappa_\xi^2(\kappa_\mu+1)^2$.

Proof. We will use $\Delta_t(\zeta, \xi)$ to denote $\mathbb{E}[\|\zeta - \zeta_t^*\|^2 + \|\xi - \xi_t^*\|^2]$. We first study some useful properties of $\Delta_t(\zeta, \xi)$.

Property 1 For $t \geq 1$

$$\begin{aligned}
\Delta_t(\zeta_{t-1}^*, \xi_{t-1}^*) = & \mathbb{E}[\|\zeta_t^* - \zeta_{t-1}^*\|^2 + \|\xi_t^* - \xi_{t-1}^*\|^2] \\
\leq & C_{\zeta,\mu}\mathbb{E}[\|\theta_t - \theta_{t-1}\|^2] \\
= & \eta_\theta^2 C_{\zeta,\mu}\mathbb{E}[\|g_\theta^{t-1}\|^2]
\end{aligned}$$

where in the inequality, we use Lemma D.2; and the last equality results from the update rule $\theta_t = \theta_{t-1} + \eta_\theta g_\theta^{t-1}$

Property 2 For $t \geq 0$,

$$\begin{aligned}
\Delta_t(\zeta_t,\xi_t) \leq& \frac{\beta}{2}\Delta_t(\zeta_{t-1},\xi_{t-1}) = \frac{\beta}{2}\mathbb{E}[\|\zeta_{t-1}-\zeta_t^*\|^2 + \|\xi_{t-1}-\xi_t^*\|^2] \\
\leq& \beta\mathbb{E}[\|\zeta_{t-1}-\zeta_{t-1}^*\|^2 + \|\xi_{t-1}-\xi_{t-1}^*\|^2 + \|\zeta_t^*-\zeta_{t-1}^*\|^2 + \|\xi_t^*-\xi_{t-1}^*\|^2] \\
=& \beta\Delta_{t-1}(\zeta_{t-1},\xi_{t-1}) + \beta\Delta_t(\zeta_{t-1}^*,\xi_{t-1}^*) \\
\leq& \beta^t\Delta_0(\zeta_0,\xi_0) + \sum_{\tau=1}^{t}\beta^{t-\tau+1}\Delta_\tau^2(\zeta_{\tau-1}^*,\xi_{\tau-1}^*) \\
\leq& \beta^{t+1}d^2 + \eta_\theta^2 C_{\zeta,\mu}\sum_{\tau=0}^{t-1}\beta^{t-\tau}\mathbb{E}[\|g_\theta^\tau\|^2]
\end{aligned}$$

where the first inequality is because of the property of the Oracle; for the second inequality we use Young's inequality; In the last step, we use

$$\Delta_0^2(\zeta_0,\xi_0) = \mathbb{E}[\|\zeta_0-\zeta_0^*\|^2 + \|\xi_0-\xi_0^*\|^2] \leq \frac{\beta}{2}\mathbb{E}[\|\zeta_{-1}-\zeta_0^*\|^2 + \|\xi_{-1}-\xi_0^*\|^2] \leq \beta d^2$$

With the two properties above, we can bound:

$$\begin{aligned}
&\mathbb{E}[\|\zeta_{t+1}-\zeta_t\|^2 + \|\xi_{t+1}-\xi_t\|^2] \\
\leq& 3\mathbb{E}[\|\zeta_{t+1}-\zeta_{t+1}^*\|^2 + \|\xi_{t+1}-\xi_{t+1}^*\|^2 + \|\zeta_{t+1}^*-\zeta_t^*\|^2 + \|\xi_{t+1}^*-\xi_t^*\|^2 + \|\zeta_t^*-\zeta_t\|^2 + \|\xi_t^*-\xi_t\|^2] \\
=& 3\Delta_{t+1}(\zeta_{t+1},\xi_{t+1}) + 3\Delta_{t+1}(\zeta_t^*,\xi_t^*) + 3\Delta_t(\zeta_t,\xi_t) \\
\leq& 3\beta^{t+2}d^2 + 3\eta_\theta^2 C_{\zeta,\mu}\sum_{\tau=0}^{t}\beta^{t-\tau+1}\mathbb{E}[\|g_\theta^\tau\|^2] + 3\eta_\theta^2 C_{\zeta,\mu}\mathbb{E}[\|g_\theta^t\|^2] + 3\beta^{t+1}d^2 + 3\eta_\theta^2 C_{\zeta,\mu}\sum_{\tau=0}^{t-1}\beta^{t-\tau}\mathbb{E}[\|g_\theta^\tau\|^2] \\
=& 3(1+\beta)\beta^{t+1}d^2 + 3\eta_\theta^2 C_{\zeta,\mu}\sum_{\tau=0}^{t}(1+\beta)\beta^{t-\tau}\mathbb{E}[\|g_\theta^\tau\|^2] \\
\leq& 6\beta^{t+1}d^2 + 6\eta_\theta^2 C_{\zeta,\mu}\sum_{\tau=0}^{t}\beta^{t-\tau}\mathbb{E}[\|g_\theta^\tau\|^2]
\end{aligned}$$

where for the first one we use an extended version of Young's inequality $\|\sum_{i=1}^k x_i\|^2 \leq k\sum_{i=1}^k \|x_i\|^2$; in the second inequality, we use the Property 1 and 2 to give the upper bound; in the third inequality, we use the fact that $0 < \beta \leq 1$. □

Lemma D.4. *Under the same condition of Lemma D.3 above, with an additional constraint $\beta \leq (1-\alpha)^2/2$ and an additional Assumption E, for $t \geq 0$, we have:*

$$\begin{aligned}
&\mathbb{E}[\|g_\theta^{t+1} - \nabla_\theta J(\theta_{t+1})\|^2] \\
\leq& 3(\varepsilon_{data}+\varepsilon_{func}+\varepsilon_{reg})^2 + 3(1-\alpha)^{2t+2}\mathbb{E}[\|g_\theta^0 - \nabla_\theta\mathcal{L}_0^D\|^2] + \frac{6\alpha\sigma^2}{|B|} + \Big(6L^2\beta^{t+2} + \frac{108C_{w,Q}}{|B|}(1-\alpha)^{2(t+2)}\Big)d^2 \\
&+ \sum_{i=0}^{t}\Big(\frac{108\eta_\theta^2}{|B|}(1-\alpha)^{2(t-i+1)}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_\mathcal{Q}^2C_\mathcal{W}^2\Big) + 6L^2\eta_\theta^2C_{\zeta,\mu}\beta^{t-i+1}\Big)\mathbb{E}[\|g_\theta^i\|^2]
\end{aligned}$$

where $\varepsilon_{data},\varepsilon_{func},\varepsilon_{reg}$ are the same as those in Theorem 2.6, and

$$C_{w,Q} := G^2L_w^2C_\mathcal{Q}^2 + G^2L_Q^2C_\mathcal{W}^2$$

Proof. Recall that we will use $\nabla_\theta\mathcal{L}_t^B$, $\nabla_\theta\mathcal{L}_t^D$ and $\nabla_\theta\mathcal{L}_t^{D*}$ as a shortnote of $\nabla_\theta\mathcal{L}^B(\theta_t,\zeta_t,\xi_t)$, $\nabla_\theta\mathcal{L}^D(\theta_t,\zeta_t,\xi_t)$, $\nabla_\theta\mathcal{L}^D(\theta_t,\zeta_t^*,\xi_t^*)$ respectively. First we can use the Young's inequality to obtain

$$\begin{aligned}
&\mathbb{E}[\|g_\theta^{t+1} - \nabla_\theta J(\theta_{t+1})\|^2] \\
\leq& 3\underbrace{\mathbb{E}[\|\nabla_\theta\mathcal{L}_{t+1}^{D*} - \nabla_\theta J(\theta_{t+1})\|^2]}_{Bias\ (Bounded\ in\ Theorem\ 2.6)} + 3\underbrace{\mathbb{E}[\|g_\theta^{t+1} - \nabla_\theta\mathcal{L}_{t+1}^D\|^2]}_{p_1} + 3\underbrace{\mathbb{E}[\|\nabla_\theta\mathcal{L}_{t+1}^D - \nabla_\theta\mathcal{L}_{t+1}^{D*}\|^2]}_{p_2}
\end{aligned}$$

Since the first term has already been bounded in Theorem 2.6. Next, we bound p_1 and p_2:

Upper bound p_1 We again use $C_{\zeta,\xi}$ as a short note of $\kappa_\mu^2(\kappa_\xi+1)^2+\kappa_\xi^2(\kappa_\mu+1)^2$. From Lemma D.3, we know that,

$$\max\{\mathbb{E}[\|\zeta_{t+1}-\zeta_t\|^2],\mathbb{E}[\|\xi_{t+1}-\xi_t\|^2]\}\leq 6\beta^{t+1}d^2+6\eta_\theta^2 C_{\zeta,\mu}\sum_{\tau=0}^{t}\beta^{t-\tau}\mathbb{E}[\|g_\theta^\tau\|^2]$$

Then, we have

$$\begin{aligned}
p_1 =&\mathbb{E}[\|g_\theta^{t+1}-\nabla_\theta\mathcal{L}_{t+1}^D\|^2]\\
=&\mathbb{E}\Big[\Big\|(1-\alpha)(g_\theta^t-\nabla_\theta\mathcal{L}_t^B)+\nabla_\theta\mathcal{L}_{t+1}^B-\nabla_\theta\mathcal{L}_{t+1}^D\pm(1-\alpha)\nabla_\theta\mathcal{L}_t^D\Big\|^2\Big]\\
=&\mathbb{E}\Big[\Big\|(1-\alpha)(g_\theta^t-\nabla_\theta\mathcal{L}_t^D)+\alpha(\nabla_\theta\mathcal{L}_{t+1}^B-\nabla_\theta\mathcal{L}_{t+1}^D)+(1-\alpha)(\nabla_\theta\mathcal{L}_{t+1}^B-\nabla_\theta\mathcal{L}_t^B)\\
&\quad-(1-\alpha)(\nabla_\theta\mathcal{L}_{t+1}^D-\nabla_\theta\mathcal{L}_t^D)\Big\|^2\Big]\\
=&(1-\alpha)^2\mathbb{E}[\|g_\theta^t-\nabla_\theta\mathcal{L}_t^D\|^2]\\
&+\mathbb{E}[\|\alpha(\nabla_\theta\mathcal{L}_{t+1}^B-\nabla_\theta\mathcal{L}_{t+1}^D)+(1-\alpha)(\nabla_\theta\mathcal{L}_{t+1}^B-\nabla_\theta\mathcal{L}_t^B)-(1-\alpha)(\nabla_\theta\mathcal{L}_{t+1}^D-\nabla_\theta\mathcal{L}_t^D)\|^2\Big]\\
&\hfill\text{(Drop 0 expectation)}\\
\leq&(1-\alpha)^2\mathbb{E}[\|g_\theta^t-\nabla_\theta\mathcal{L}_t^D\|^2]+2\alpha^2\mathbb{E}[\|(\nabla_\theta\mathcal{L}_{t+1}^B-\nabla_\theta\mathcal{L}_{t+1}^D)\|^2\Big]\\
&+2(1-\alpha)^2\mathbb{E}\Big[\Big\|(\nabla_\theta\mathcal{L}_{t+1}^B-\nabla_\theta\mathcal{L}_t^B)-(\nabla_\theta\mathcal{L}_{t+1}^D-\nabla_\theta\mathcal{L}_t^D)\Big\|^2\Big] \qquad\text{(Young's Ineq.)}\\
\leq&(1-\alpha)^2\mathbb{E}[\|g_\theta^t-\nabla_\theta\mathcal{L}_t^D\|^2]+\frac{2\alpha^2\sigma^2}{|B|}+2(1-\alpha)^2\mathbb{E}\Big[\Big\|(\nabla_\theta\mathcal{L}_{t+1}^B-\nabla_\theta\mathcal{L}_t^B)\Big\|^2\Big]\\
&\hfill\text{(Assumption E)}\\
\leq&(1-\alpha)^2\mathbb{E}[\|g_\theta^t-\nabla_\theta\mathcal{L}_t^D\|^2]+\frac{2\alpha^2\sigma^2}{|B|}\\
&+\frac{6(1-\alpha)^2}{|B|}\Big(G^2L_w^2C_{\mathcal{Q}}^2\mathbb{E}[\|\zeta_{t+1}-\zeta_t\|^2]+G^2L_Q^2C_{\mathcal{W}}^2\mathbb{E}[\|\xi_{t+1}-\xi_t\|^2]+H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\mathbb{E}[\|\theta_{t+1}-\theta_t\|^2]\Big)\\
\leq&(1-\alpha)^{2t+2}\mathbb{E}[\|g_\theta^0-\nabla_\theta\mathcal{L}_0^D\|^2]+\frac{2\alpha^2\sigma^2}{|B|}\frac{1-(1-\alpha)^{2t+2}}{1-(1-\alpha)^2}\\
&+\frac{6}{|B|}\mathbb{E}\Big[\sum_{i=0}^{t}(1-\alpha)^{2(t-i+1)}\Big(G^2L_w^2C_{\mathcal{Q}}^2\|\zeta_{i+1}-\zeta_i\|^2+G^2L_Q^2C_{\mathcal{W}}^2\|\xi_{i+1}-\xi_i\|^2+H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\|\theta_{i+1}-\theta_i\|^2\Big)\Big]\\
\leq&(1-\alpha)^{2t+2}\mathbb{E}[\|g_\theta^0-\nabla_\theta\mathcal{L}_0^D\|^2]+\frac{2\alpha\sigma^2}{|B|}+\frac{36}{|B|}\sum_{i=0}^{t}(1-\alpha)^{2(t-i+1)}C_{w,Q}\beta^{i+1}d^2 \qquad(\alpha<1)\\
&+\frac{36\eta_\theta^2}{|B|}\sum_{i=0}^{t}\Big(C_{\zeta,\mu}C_{w,Q}\sum_{\tau=i}^{t}(1-\alpha)^{2(t-\tau+1)}\beta^{\tau-i}+(1-\alpha)^{2(t-i+1)}H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big)\mathbb{E}[\|g_\theta^i\|^2]\\
&\hfill\text{(Lemma D.3 and } ab+cd\leq(a+b)(c+d) \text{ for } a,b,c,d\geq 0)\\
\leq&(1-\alpha)^{2t+2}\mathbb{E}[\|g_\theta^0-\nabla_\theta\mathcal{L}_0^D\|^2]+\frac{2\alpha\sigma^2}{|B|}+\frac{36C_{w,Q}}{|B|}\frac{\beta(1-\alpha)^{2(t+2)}}{(1-\alpha)^2-\beta}d^2\\
&+\frac{36\eta_\theta^2}{|B|}\sum_{i=0}^{t}(1-\alpha)^{2(t-i+1)}\Big(C_{\zeta,\mu}C_{w,Q}\frac{(1-\alpha)^2}{(1-\alpha)^2-\beta}+H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big)\mathbb{E}[\|g_\theta^i\|^2]\\
\leq&(1-\alpha)^{2t+2}\mathbb{E}[\|g_\theta^0-\nabla_\theta\mathcal{L}_0^D\|^2]+\frac{2\alpha\sigma^2}{|B|}+\frac{36C_{w,Q}}{|B|}\frac{\beta(1-\alpha)^{2(t+2)}}{(1-\alpha)^2-\beta}d^2\\
&+\frac{36\eta_\theta^2}{|B|}\sum_{i=0}^{t}(1-\alpha)^{2(t-i+1)}\Big(2C_{\zeta,\mu}C_{w,Q}+H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big)\mathbb{E}[\|g_\theta^i\|^2] \qquad (21)
\end{aligned}$$

where the fourth equality because $\mathbb{E}[\nabla_\theta \mathcal{L}_t^B] = \nabla_\theta \mathcal{L}_t^D$ holds for all t and so the cross terms has 0 expectation; the first inequality is because variance is less than the second momentum; the second inequality we apply Lemma D.1 and Assumption A; in the last but two inequality, we apply the summation formula of equal ratio sequence and use the fact that $0 < \alpha \le 1, \beta \le 1$; in the last step, we use our condition $\beta \le (1-\alpha)^2/2$

Upper bound p_2 Next, we give an upper bound for p_2. From the Property 2 in Lemma D.3, we know that

$$\Delta_{t+1}(\zeta_{t+1}, \xi_{t+1}) = \mathbb{E}[\|\zeta_{t+1} - \zeta_{t+1}^*\|^2] + \mathbb{E}[\|\xi_{t+1} - \xi_{t+1}^*\|^2] \le \beta^{t+2} d^2 + \eta_\theta^2 C_{\zeta,\mu} \sum_{\tau=0}^{t} \beta^{t-\tau+1} \mathbb{E}[\|g_\theta^\tau\|^2]$$

As a result

$$\begin{aligned} p_2 =& \mathbb{E}[\|\nabla_\theta \mathcal{L}_{t+1}^D - \nabla_\theta \mathcal{L}_{t+1}^{D*}\|^2] \le 2L^2 \mathbb{E}[\|\zeta_{t+1} - \zeta_{t+1}^*\|^2 + \|\xi_{t+1} - \xi_{t+1}^*\|^2] \\ \le& 2L^2 \Big(\beta^{t+2} d^2 + \eta_\theta^2 C_{\zeta,\mu} \sum_{\tau=0}^{t} \beta^{t-\tau+1} \mathbb{E}[\|g_\theta^\tau\|^2]) \end{aligned}$$

Combine these two results we can finish the proof:

$$\begin{aligned} &\mathbb{E}[\|g_\theta^{t+1} - \nabla_\theta J(\theta_{t+1})\|^2] \le 3\mathbb{E}[\|\nabla_\theta \mathcal{L}_{t+1}^{D*} - \nabla_\theta J(\theta_{t+1})\|^2] + 3p_1 + 3p_2 \\ \le& 3(\varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg})^2 + 3(1-\alpha)^{2t+2} \mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2] + \frac{6\alpha\sigma^2}{|B|} + \frac{108 C_{w,Q}}{|B|} \frac{\beta(1-\alpha)^{2(t+2)}}{(1-\alpha)^2 - \beta} d^2 \\ &+ \frac{108\eta_\theta^2}{|B|} \sum_{i=0}^{t} (1-\alpha)^{2(t-i+1)} \Big(2C_{\zeta,\mu} C_{w,Q} + H^2 C_Q^2 C_{\mathcal{W}}^2\Big) \mathbb{E}[\|g_\theta^i\|^2] \\ &+ 6L^2 \Big(\beta^{t+2} d^2 + \eta_\theta^2 C_{\zeta,\mu} \sum_{\tau=0}^{t} \beta^{t-\tau+1} \mathbb{E}[\|g_\theta^\tau\|^2]) \\ \le& 3(\varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg})^2 + 3(1-\alpha)^{2t+2} \mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2] + \frac{6\alpha\sigma^2}{|B|} + \Big(6L^2 \beta^{t+2} + \frac{108 C_{w,Q}}{|B|} \frac{\beta(1-\alpha)^{2(t+2)}}{(1-\alpha)^2 - \beta}\Big) d^2 \\ &+ \sum_{i=0}^{t} \Big(\frac{108\eta_\theta^2}{|B|} (1-\alpha)^{2(t-i+1)} \Big(2C_{\zeta,\mu} C_{w,Q} + H^2 C_Q^2 C_{\mathcal{W}}^2\Big) + 6L^2 \eta_\theta^2 C_{\zeta,\mu} \beta^{t-i+1}\Big) \mathbb{E}[\|g_\theta^i\|^2] \end{aligned}$$

□

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*

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

Proof. Recall that,

$$\nabla_\theta J(\pi) = \int_\tau p(\tau|\theta) \sum_{i=0}^{\infty} \gamma^i r_i \sum_{j=0}^{i} \nabla_\theta \log \pi_\theta(a_j|s_j) d\tau$$

Therefore,

$$\begin{aligned} \nabla_\theta^2 J(\pi) =& \int_\tau p(\tau|\theta) \sum_{i=0}^{\infty} \gamma^i r_i \sum_{j=0}^{i} \nabla_\theta^2 \log \pi_\theta(a_j|s_j) d\tau \\ &+ \int_\tau p(\tau|\theta) \nabla_\theta \log p(\tau|\theta) \sum_{i=0}^{\infty} \gamma^i r_i \sum_{j=0}^{i} \nabla_\theta \log \pi_\theta(a_j|s_j) d\tau \\ =& \int_\tau p(\tau|\theta) \sum_{i=0}^{\infty} \gamma^i r_i \sum_{j=0}^{i} \nabla_\theta^2 \log \pi_\theta(a_j|s_j) d\tau \\ &+ \int_\tau p(\tau|\theta) \sum_{i=0}^{\infty} \gamma^i r_i \big(\sum_{j=0}^{i} \nabla_\theta \log \pi(a_t|s_t)\big) \big(\sum_{j=0}^{i} \nabla_\theta \log \pi(a_t|s_t)\big)^\top d\tau \end{aligned}$$

Therefore,

$$\begin{aligned}
\|\nabla_\theta^2 J(\pi)\|_{op} \le & \int_\tau p(\tau|\theta)\sum_{i=0}^{\infty}\gamma^i\sum_{j=0}^{i}\|\nabla_\theta^2\log\pi_\theta(a_j|s_j)\|_{op}d\tau \\
& + \int_\tau p(\tau|\theta)\sum_{i=0}^{\infty}\gamma^i\|\big(\sum_{j=0}^{i}\nabla_\theta\log\pi(a_t|s_t)\big)\big(\sum_{j=0}^{i}\nabla_\theta\log\pi(a_t|s_t)\big)^\top\|_{op}d\tau \\
\le & \sum_{i=0}^{\infty}\gamma^i(i+1)H + \sum_{i=0}^{\infty}\gamma^i(i+1)^2G^2 \\
= & \frac{H}{(1-\gamma)^2} + \frac{(1+\gamma)G^2}{(1-\gamma)^3}
\end{aligned}$$

□

Theorem 4.3. *Given arbitrary ε, suppose $|B|$ and T satisfy the following constraints:*

$$\begin{aligned}
T \approx & \max\{96, \frac{16L_J}{\varepsilon^2}\} = O(\varepsilon^{-2}) \\
|B|T \approx & \max\{\frac{576\sigma}{(1-\gamma)\varepsilon^3}\sqrt{2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{W}}^2C_{\mathcal{Q}}^2}, \frac{864C_{w,Q}d^2}{\varepsilon^2}\} = O(\varepsilon^{-3})
\end{aligned}$$

where $C_{w,Q} = G^2L_w^2C_{\mathcal{Q}}^2 + G^2C_{\mathcal{W}}^2L_Q^2$, $C_{\zeta,\mu} = \kappa_\mu^2(\kappa_\xi+1)^2 + \kappa_\xi^2(\kappa_\mu+1)^2$ and L_J is defined in Prop. 4.2, while other hyper-parameters satisfy:

$$\begin{aligned}
\alpha = & \frac{|B|\varepsilon^2}{12\sigma}; \quad \beta \le \min\{\frac{\varepsilon^2}{L^2}, \frac{(1-\gamma)^2\varepsilon^4}{C_{\zeta,\mu}L^2}, \frac{\alpha}{2}(1-\alpha)^2\}; \quad B_0 = \frac{4\sigma^2}{\varepsilon^2} \\
\eta_\theta \le & \min\{\frac{1}{2L_J}, \Big(108\Big[\frac{C_{\zeta,\mu}L^2\beta}{18(1-\beta)} + \frac{1}{\alpha|B|}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{W}}^2C_{\mathcal{Q}}^2\Big)\Big]\Big)^{-1/2}\}
\end{aligned}$$

The Algorithm 2 will return us a policy π_{θ_T} after T steps with batch size $|B|$, satisfying

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

The total gradient computation of Algorithm 1 (ignoring Oracle) is $|B_0| + |B|T = O(\varepsilon^{-3})$.

Proof.

$$\begin{aligned}
J(\theta_{T+1}) = & J(\theta_T + \eta_\theta g_\theta^T) \\
\ge & J(\theta_T) + \eta_\theta(g_\theta^T)^\top\nabla_\theta J(\theta_T) - \frac{\eta_\theta^2L_J}{2}\|g_\theta^T\|^2 \\
= & J(\theta_T) + \frac{\eta_\theta}{2}\|\nabla_\theta J(\theta_T)\|^2 - \frac{\eta_\theta}{2}\|g_\theta^T - \nabla_\theta J(\theta_T)\|^2 + (\frac{\eta_\theta}{2} - \frac{\eta_\theta^2L_J}{2})\|g_\theta^T\|^2 \\
\ge & J(\theta_T) + \frac{\eta_\theta}{2}\|\nabla_\theta J(\theta_T)\|^2 - \frac{\eta_\theta}{2}\|g_\theta^T - \nabla_\theta J(\theta_T)\|^2 + \frac{\eta_\theta}{4}\|g_\theta^T\|^2 \\
\ge & J(\theta_0) + \frac{\eta_\theta}{2}\sum_{t=0}^{T}\|\nabla_\theta J(\theta_t)\|^2 - \frac{\eta_\theta}{2}\underbrace{\Big(\sum_{t=0}^{T}\|g_\theta^t - \nabla_\theta J(\theta_t)\|^2 - \frac{1}{2}\|g_\theta^t\|^2\Big)}_{p}
\end{aligned}$$

where in the second equation, we use the fact that $(g_\theta^T)^\top\nabla_\theta J(\theta_T) = \frac{1}{2}\|\nabla_\theta J(\theta_T)\|^2 + \frac{1}{2}\|g_\theta^T\|^2 - \frac{1}{2}\|g_\theta^T - \nabla_\theta J(\theta_T)\|^2$; in the second inequality, we add a constraint for η_θ that $\eta_\theta \le \frac{1}{2L_J}$;

Next, we give a upper bound for p with Lemma D.4:

$$p = \sum_{t=0}^{T}\|g_\theta^\tau - \nabla_\theta J(\theta_t)\|^2 - \frac{1}{2}\|g_\theta^t\|^2$$

$$
\begin{aligned}
&\leq \sum_{t=0}^{T} \Big\{ 3(\varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg})^2 \\
&\quad + 3(1-\alpha)^{2t+2}\mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2] + \frac{6\alpha\sigma^2}{|B|} + \Big(6L^2\beta^{t+2} + \frac{108C_{w,Q}}{|B|}\frac{\beta(1-\alpha)^{2(t+2)}}{(1-\alpha)^2-\beta}\Big)d^2 \\
&\quad + \sum_{i=0}^{t}\Big(\frac{108\eta_\theta^2}{|B|}(1-\alpha)^{2(t-i+1)}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big) + 6L^2\eta_\theta^2 C_{\zeta,\mu}\beta^{t-i+1}\Big)\mathbb{E}[\|g_\theta^i\|^2] - \frac{1}{2}\mathbb{E}[\|g_\theta^t\|^2]\Big\} \\
&\leq 3T(\varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg})^2 \\
&\quad + \frac{3}{1-(1-\alpha)^2}\mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2] + \frac{6\alpha T\sigma^2}{|B|} + (\frac{6\beta L^2}{1-\beta} + \frac{108\beta(1-\alpha)^2C_{w,Q}}{|B|(1-(1-\alpha)^2)((1-\alpha)^2-\beta)})\Big)d^2 \\
&\quad + \sum_{t=0}^{T}\mathbb{E}[\|g_\theta^t\|^2]\Big\{-\frac{1}{2} + 108\eta_\theta^2\sum_{i=1}^{T-t+1}\Big[\frac{C_{\zeta,\mu}L^2\beta^i}{18} + \frac{(1-\alpha)^{2i}}{|B|}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big)\Big]\Big\} \\
&\leq 3T(\varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg})^2 + \frac{3}{\alpha}\mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2] + \frac{6\alpha T\sigma^2}{|B|} + (\frac{6\beta L^2}{1-\beta} + 108\frac{C_{w,Q}}{|B|}\frac{\beta}{\alpha((1-\alpha)^2-\beta)})\Big)d^2 \\
&\quad + \sum_{t=0}^{T}\mathbb{E}[\|g_\theta^i\|^2]\Big(-\frac{1}{2} + 108\eta_\theta^2\Big[\frac{C_{\zeta,\mu}L^2\beta}{18(1-\beta)} + \frac{1}{|B|}\frac{(1-\alpha)^2}{1-(1-\alpha)^2}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big)\Big] \\
&\leq 3T(\varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg})^2 \\
&\quad + \frac{3}{\alpha}\mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2] + \frac{6\alpha T\sigma^2}{|B|} + (\frac{6\beta L^2}{1-\beta} + 108\frac{C_{w,Q}}{|B|}\frac{\alpha(1-\alpha)^2/2}{\alpha((1-\alpha)^2-(1-\alpha)^2/2))})\Big)d^2 \\
&\quad + \sum_{t=0}^{T}\mathbb{E}[\|g_\theta^i\|^2]\Big(-\frac{1}{2} + 108\eta_\theta^2\Big[\frac{C_{\zeta,\mu}L^2\beta}{18(1-\beta)} + \frac{1}{\alpha|B|}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big)\Big] \\
&\leq 3T(\varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg})^2 + \frac{3}{\alpha}\mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2] + \frac{6\alpha T\sigma^2}{|B|} + (\frac{6\beta L^2}{1-\beta} + \frac{108C_{w,Q}}{|B|})d^2
\end{aligned}
$$

In the first, second and third inequality, we use the fact that $0 < (1-\alpha) \leq 1, 0 < \beta \leq \alpha(1-\alpha)^2/2 \leq (1-\alpha)^2/2\}$. In the fourth inequality, we add the following constraint to drop the terms containing $\|g_\theta\|$:

$$
\eta_\theta \leq \Big(108\Big[\frac{C_{\zeta,\mu}L^2\beta}{18(1-\beta)} + \frac{1}{\alpha|B|}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big)\Big]\Big)^{-1/2} \tag{22}
$$

Therefore,

$$
\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla_\theta J(\theta_\tau)\|^2 \leq& \frac{2}{(T+1)\eta_\theta}(J(\theta_T) - J(\theta_0)) + \frac{1}{T+1}\sum_{\tau=0}^{T}\Big(\|g_\theta^\tau - \nabla_\theta J(\theta_\tau)\|^2 - \frac{1}{2}\|g_\theta^\tau\|^2\Big) \\
\leq& 3(\varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg})^2 + \frac{2}{(T+1)\eta_\theta(1-\gamma)} + \frac{3}{\alpha(T+1)}\mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2] \\
&+ \frac{6\alpha\sigma^2}{|B|} + \frac{1}{T+1}(\frac{6\beta L^2}{1-\beta} + \frac{108C_{w,Q}}{|B|})d^2 \\
\leq& 3(\varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg})^2 + \underbrace{\frac{2}{T\eta_\theta(1-\gamma)}}_{p_1} + \underbrace{\frac{3}{\alpha T}\mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2]}_{p_2} \\
&+ \underbrace{\frac{6\alpha\sigma^2}{|B|}}_{p_3} + \underbrace{\frac{1}{T}(\frac{6\beta L^2}{1-\beta} + \frac{108C_{w,Q}}{|B|})d^2}_{p_4}
\end{aligned}
$$

Next, we want to carefully choose hyper-parameters to make sure $p_1, p_2, p_3, p_4 \leq \varepsilon^2/4$. We consider $\beta \leq \min\{\frac{\varepsilon^2}{L^2}, \frac{(1-\gamma)^2\varepsilon^4}{C_{\zeta,\mu}L^2}, \frac{1}{2}(1-\alpha)^2, \alpha(1-\alpha)^2\}$. Since $0 < \alpha \leq 1$, we have $\beta < \frac{1}{2}$.

Control p_1 Since we have two constrains on η_θ, first we need to make sure, if $\eta_\theta = \frac{1}{2L_J}$

$$p_1 = \frac{4L_J}{T(1-\gamma)} \le \frac{\varepsilon^2}{4}$$

Combining 4.2, the above implies that:

$$T \ge \frac{16L_J}{(1-\gamma)\varepsilon^2} \tag{23}$$

Secondly, to make sure constraint (22):

$$\begin{aligned}
p_1 =& \frac{2}{T(1-\gamma)}\Big(108\Big[\frac{C_{\zeta,\mu}L^2\beta}{18(1-\beta)} + \frac{1}{\alpha|B|}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big)\Big]\Big)^{1/2} \\
\le& \frac{2}{T(1-\gamma)}\sqrt{\frac{12C_{\zeta,\mu}L^2\beta}{1-\beta}} + \frac{2}{T(1-\gamma)}\sqrt{\frac{108}{\alpha|B|}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big)} \\
\le& \frac{2}{T(1-\gamma)}\sqrt{12L^2C_{\zeta,\mu}\frac{(1-\gamma)^2\varepsilon^4}{C_{\zeta,\mu}L^2}} + \frac{2}{T(1-\gamma)}\sqrt{\frac{108}{\alpha|B|}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big)} \\
=& \frac{4\sqrt{3}\varepsilon^2}{T} + \frac{2}{T(1-\gamma)}\sqrt{\frac{108}{\alpha|B|}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big)}
\end{aligned}$$

To make sure $p_1 \le \frac{\varepsilon^2}{4}$, we need the above two terms less than $\frac{\varepsilon^2}{8}$ at the same time, which implies

$$T \ge 32\sqrt{3}; \quad |B|T \ge \frac{16}{(1-\gamma)\varepsilon^2}\sqrt{\frac{108|B|}{\alpha}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big)} \tag{24}$$

Control p_2 In fact, at the beginning step, $\mathbb{E}_{B_0}[g_\theta^0] = \nabla_\theta \mathcal{L}_0^D$. Therefore,

$$p_2 = \frac{\sigma^2}{|B_0|}$$

To make sure $|B_0| \ge \frac{4\sigma^2}{\varepsilon^2}$, we just set

$$|B_0| = \frac{4\sigma^2}{\varepsilon^2}. \tag{25}$$

Control p_3 We want $p_3 \le \frac{\varepsilon^2}{4}$. To do that, we add the following constraint

$$\frac{|B|}{\alpha} \ge \frac{12\sigma^2}{\varepsilon^2} \tag{26}$$

Control p_4 Since $\beta \le \{1/2, \varepsilon^2/L^2\}$, we have

$$p_4 = \frac{1}{T}\Big(\frac{6\beta L^2}{1-\beta} + \frac{108C_{w,Q}}{|B|}\Big)d^2 \le \frac{1}{T}\Big(\frac{\varepsilon^2}{L^2}\frac{6L^2}{1-1/2} + \frac{108C_{w,Q}}{|B|}d^2 = \frac{12\varepsilon^2}{T} + 108\frac{C_{w,Q}d^2}{|B|T}$$

To make sure $p_4 \le \frac{\varepsilon^2}{4}$, we need the above two terms individually smaller than $\frac{\varepsilon^2}{8}$

$$T \ge 96; \quad |B|T \ge \frac{864C_{w,Q}d^2}{\varepsilon^2} \tag{27}$$

Combine (23)-(27), we need

$$\begin{aligned}
&|B_0| + |B|T \ge \frac{4\sigma^2}{\varepsilon^2} + \max\{\frac{16}{(1-\gamma)\varepsilon^2}\sqrt{\frac{108|B|}{\alpha}\Big(2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2\Big)}, \frac{864C_{w,Q}d^2}{\varepsilon^2}\} \\
&subject\ to \qquad \frac{|B|}{\alpha} \ge \frac{12\sigma^2}{\varepsilon^2}; \quad T \ge \max\{96, \frac{16L_J}{\varepsilon^2}\};
\end{aligned}$$

To minimize $|B_0| + |B|T$, we may choose $\frac{|B|}{\alpha} = \frac{12\sigma^2}{\varepsilon^2}$. As a result,

$$\begin{aligned}
&|B_0| + |B|T = \frac{4\sigma^2}{\varepsilon^2} + \max\{\frac{576\sigma}{(1-\gamma)\varepsilon^3}\sqrt{2C_{\zeta,\mu}C_{w,Q} + H^2C_{\mathcal{Q}}^2C_{\mathcal{W}}^2}, \frac{864C_{w,Q}d^2}{\varepsilon^2}\} = O(\varepsilon^{-3}) \\
&subject\ to \qquad T \ge \max\{96, \frac{16L_J}{\varepsilon^2}\} = O(\varepsilon^{-2})
\end{aligned}$$

$\square$

E Practicality of the Assumptions in Section 2.2

First, it is common to use policy classes whose first and second order derivatives are bounded [15, 16], so the Assumption A-(1) is a reasonable one. Also, Assumption B is a common assumption in batch RL that guarantees exploratory dataset [23], and the smoothness Assumption A-(c) is frequently considered in optimization literatures.

The remaining assumptions are indeed quite strong. That said, below we show that when $\mathcal{W}$ and $\mathcal{Q}$ are the same linear class, we can satisfy these assumptions relatively easily. Indeed, Uehara et al. [4] showed that MIS-based OPE reduce to the familiar off-policy LSTD algorithms with linear classes [24, 25], and we show that Assumptions A-(b), C, D, E, F, G can be satisfied in this case if we simply assume Assumption H, which is standard in the off-policy LSTD literature.

Definition E.1 (Linear function classes)**.** We have a feature class $\{\phi(s,a) \in \mathbb{R}^{n\times 1} | \forall s,a \in \mathcal{S}\times\mathcal{A}\}$ subject to $\|\phi(s,a)\| = 1$, and two parameter spaces $Z, \Xi \in \mathbb{R}^{n\times 1}$. The approximated value function Q_ξ and density ratio w_ζ are represented by

$$w(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \zeta, \quad Q(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \xi$$

Remark E.2. *Since $\|\phi(\cdot,\cdot)\| \leq 1$, the matrix $\mathbb{E}_{s,a\sim d^D}[\phi(s,a)\phi(s,a)^\top]$ is semi-positive definite and its largest eigenvalue is less than 1.*

Assumption H. There exists a positive constant $\sigma_{\min}$ that, the matrix $\mathbb{E}_{s,a\sim d^D}[\phi(s,a)\phi(s,a)^\top]$ is full-rank, and all its eigenvalues are no less than $\sigma_{\min}$; besides, the matrix $\mathbb{E}_{s,a\sim d^D}[\phi(s,a)\phi(s,a)^\top - \gamma\phi(s,a)\phi(s',a)]$ is invertible, and its minimal sigular value is no less than $\sigma_{\min}$.

Remark E.3. *In Assumption H, we only add requirement on the smallest singular value of M and do not care about whether all its eigenvalues are positive or not.*

For simplicity, we choose $\lambda_w = \lambda_Q = \lambda > 0$. We use $\Phi \in \mathbb{R}^{|S||A|\times n}$ to denote the matrix concatenated by all features, use K to denote $\Phi^\top \Lambda^D \Phi$ and use M to denote $\Phi^\top\Lambda^D(I-\gamma P^\pi)\Phi$, where Λ^D is a diagonal matrix whose diagonal elements are $d^D(\cdot,\cdot)$. By choosing linear function classes, we can rewrite $\mathcal{L}^D$ to:

$$\begin{aligned}
\mathcal{L}^D(\pi,\zeta,\xi) =&(1-\gamma)\mathbb{E}_{s_0}[Q(s_0,\pi)] + \mathbb{E}_w[r+\gamma Q(s',\pi) - Q(s,a)] + \frac{\lambda}{2}\mathbb{E}_{d^D}[Q^2(s,a)] - \frac{\lambda}{2}\mathbb{E}_{d^D}[w^2(s,a)]\\
=&(1-\gamma)\nu_D^\pi\Phi\xi + \zeta^\top\Phi^\top\Lambda^D(R-(I-\gamma P^\pi)\Phi\xi) + \frac{\lambda}{2}\xi^\top K\xi - \frac{\lambda}{2}\zeta^\top K\zeta\\
=&(1-\gamma)\nu_D^\pi\Phi\xi + \zeta^\top\Phi^\top\Lambda^D R - \zeta^\top M\xi + \frac{\lambda}{2}\xi^\top K\xi - \frac{\lambda}{2}\zeta^\top K\zeta
\end{aligned}$$

Since $\mathcal{L}^D$ is quadratic, under Assumption H, matrix K is full-rank with minimal eigenvalue larger than $\sigma_{\min}$ and maximal eigenvalue smaller than 1, then $\mathcal{L}^D(\pi,\zeta,\xi)$ is $\lambda\sigma_{\min}$-strongly-concave-$\lambda\sigma_{\min}$-strongly-convex, and λ smooth. Combining bounded second order derivatives of $\log\pi$, $\mathcal{L}$ is also smooth w.r.t. θ. Therefore, we know Assumption C holds.

Next, we try to give a bound for the norm of the saddle point of $\mathcal{L}^D(\pi, w_\zeta, Q_\xi)$ denotes as (ζ^*,ξ^*), to testify the other assumptions. By taking derivatives w.r.t. ξ, we have:

$$\xi = \frac{1}{\lambda}K^{-1}\Big(M^\top\zeta - (1-\gamma)\Phi^\top(\nu_D^\pi)^\top\Big)$$

Plug it into $\mathcal{L}^D$:

$$-\frac{\lambda}{2}\zeta^\top K\zeta - \frac{1}{2\lambda}\Big(M^\top\zeta-(1-\gamma)\Phi^\top(\nu_D^\pi)^\top\Big)^\top K^{-1}\Big(M^\top\zeta - (1-\gamma)\Phi^\top(\nu_D^\pi)^\top\Big) + \zeta^\top\Phi^\top\Lambda^D R$$

Taking the derivative of ζ, we have:

$$\zeta^* = \Big(\lambda^2 K + MK^{-1}M^\top\Big)^{-1}\Big(-(1-\gamma)MK^{-1}\Phi^\top(\nu_D^\pi)^\top + \lambda\Phi^\top\Lambda^D R\Big)$$

and therefore,

$$\begin{aligned}
\xi^* =& \frac{1}{\lambda} K^{-1}\Big(M^\top \zeta^* - (1-\gamma)\Phi^\top (\nu_D^\pi)^\top\Big) \\
=& \frac{1}{\lambda} K^{-1} M^\top \Big(\lambda^2 K + M K^{-1} M^\top\Big)^{-1} \cdot \Big(-(1-\gamma) M K^{-1} \Phi^\top (\nu_D^\pi)^\top + \lambda \Phi^\top \Lambda^D R\Big) \\
& + (1-\gamma)\frac{1}{\lambda} K^{-1} \Phi^\top (\nu_D^\pi)^\top \\
=& (1-\gamma)\lambda \Big(\lambda^2 K + M^\top K^{-1} M\Big)^{-1} \cdot \Phi^\top (\nu_D^\pi)^\top + K^{-1} M^\top \Big(\lambda^2 K + M K^{-1} M^\top\Big)^{-1} \Phi^\top \Lambda^D R
\end{aligned}$$

where in the last step, we use the inverse matrix lemma:

$$(\lambda^2 K + M^\top K^{-1} M)^{-1} = \frac{1}{\lambda^2} K^{-1} - \frac{1}{\lambda^2} K^{-1} M^\top (\lambda^2 K + M K^{-1} M^\top) M K^{-1}$$

Because $\|\phi(\cdot,\cdot)\| \leq 1$, it's easy to prove that, for arbitrary vector $x \in \mathbb{R}^d$,

$$\max\{\|Mx\|, \|M^\top x\|\} \leq (1+\gamma)\|x\|$$

Therefore,

$$\begin{aligned}
\|\zeta^*\| \leq & \Big\|\Big(\lambda^2 K + M K^{-1} M^\top\Big)^{-1}\Big\| \cdot \|-(1-\gamma) M K^{-1} \Phi^\top (\nu_D^\pi)^\top + \lambda \Phi^\top \Lambda^D R\| \\
\leq & \Big\|\Big(M K^{-1} M^\top\Big)^{-1}\Big\| \cdot \Big(\|M\| \cdot \|K^{-1}\| \cdot (1-\gamma)\mathbb{E}_{\nu_D}[\|\phi(s,\pi)\|] + \lambda \mathbb{E}_{d^D}[\|\phi(s,a) r(s,a)\|]\Big) \\
\leq & \frac{1}{\sigma_{\min}^2}((1-\gamma)\frac{1+\gamma}{\sigma_{\min}} + \lambda) := D_\zeta \\
\|\xi^*\| \leq & (1-\gamma)\lambda \Big\|\Big(\lambda^2 K + M^\top K^{-1} M\Big)^{-1}\Big\| \cdot \mathbb{E}_{\nu_D}[\|\phi(s,\pi)\|] \\
& + \|K^{-1} M^\top\| \Big\|\Big(\lambda^2 K + M K^{-1} M^\top\Big)^{-1}\Big\| \mathbb{E}_{d^D}[\|\phi(s,a) r(s,a)\|] \\
\leq & \frac{1}{\sigma_{\min}^2}\Big((1-\gamma)\lambda + \frac{1+\gamma}{\sigma_{\min}}\Big) := D_\xi
\end{aligned}$$

By choosing $Z = \{\zeta | \|\zeta\| \leq D_\zeta + 1\}$ and $\Xi = \{\xi | \|\xi\| \leq D_\xi + 1\}$, Assumptions D and F, G can be satisfied when $d = 2\max\{D_\zeta, D_\xi\} + 2$. Moreover,

$$\begin{aligned}
w_\zeta(s,a) = \phi(s,a)^\top \zeta \leq & \|\phi(s,a)\| \|\zeta\| \leq D_\zeta \\
Q_\xi(s,a) = \phi(s,a)^\top \xi \leq & \|\phi(x,a)\| \|\xi\| \leq D_\xi \\
\|w_{\zeta_1}(s,a) - w_{\zeta_2}(s,a)\| \leq & \|\phi(s,a)\| \|\zeta_1 - \zeta_2\| \leq \|\zeta_1 - \zeta_2\| \\
\|Q_{\xi_1}(s,a) - Q_{\xi_2}(s,a)\| \leq & \|\phi(s,a)\| \|\xi_1 - \xi_2\| \leq \|\xi_1 - \xi_2\|
\end{aligned}$$

which means Assumption A-(b) is satisfied by setting $C_{\mathcal{W}} = D_\zeta, C_{\mathcal{Q}} = D_\xi$ and $L_{\mathcal{W}} = L_{\mathcal{Q}} = 1$. Besides, D_ζ and D_ξ are finite also implies that σ in Assumption E is finite.