On the Convergence Rate of Off-Policy Policy Optimization Methods with Density-Ratio Correction

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

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

Abstract

In this paper, we study the convergence properties of off-policy policy optimization algorithms with state-action density ratio correction under function approximation setting, where the objective function is formulated as a max-max-min problem. We first clearly characterize the bias of the learning objective, and then present two strategies with finite-time convergence guarantees. In our first strategy, we propose an algorithm called P-SREDA with convergence rate $O(\varepsilon^{-3})$, whose dependency on ε is optimal. Besides, in our second strategy, we design a new off-policy actor-critic style algorithm named O-SPIM. We prove that O-SPIM converges to a stationary point with total complexity $O(\varepsilon^{-4})$, which matches the convergence rate of some recent actor-critic algorithms under on-policy setting.

1 INTRODUCTION

Policy improvement is a popular class of methods in empirical reinforcement-learning (RL) research, and has attracted significant attention from the theoretical community recently [Agarwal et al., 2019], with many results analyzing the convergence properties of policy gradient-style algorithms [Xu et al., 2019b,a, Yuan et al., 2020, Huang et al., 2020b] and actor-critic algorithms [Fu et al., 2020, Wu

et al., 2020]. Most existing results require on-policy roll-outs, which are not available in offline RL, a paradigm considered crucial to applying RL in real-world problems [Levine et al., 2020]. While offline policy optimization algorithms also exist [Liu et al., 2019, Imani et al., 2018, Zhang et al., 2019c], the understanding on their finite-time convergence property is still limited.

To close this gap, in this paper, we propose two strategies for off-policy policy optimization with convergence guarantess, based on recent development in marginalized importance sampling (MIS) methods [Liu et al., 2018, Zhang et al., 2019a, Uehara et al., 2019, Yang et al., 2020]. We assume the agent can only get access to a fixed dataset D collected by some unknown policies, and formulate the off-policy learning problem as a max-max-min objective function below (similar objectives have been considered by Nachum et al. [2019b], Jiang and Huang [2020], Yang et al. [2020] without convergence rate analyses):

$$\max_{\pi \in \Pi} \max_{w \in \mathcal{W}} \min_{Q \in \mathcal{Q}} \mathcal{L}^D(\pi, w, Q)$$

$$:= \max_{\theta \in \Theta} \max_{\zeta \in Z} \min_{\xi \in \Xi} \mathcal{L}^D(\pi_\theta, w_\zeta, Q_\xi)$$

$$:= (1 - \gamma) \mathbb{E}_{s_0 \sim \nu_D}[Q_\xi(s_0, \pi_\theta)]$$

$$+ \mathbb{E}_{d^D}[w_\zeta(s, a)\Big(r + \gamma Q_\xi(s', \pi_\theta) - Q_\xi(s, a)\Big)]$$

$$+ \frac{\lambda_Q}{2} \mathbb{E}_{d^D}[Q_\xi^2(s, a)] - \frac{\lambda_w}{2} \mathbb{E}_{d^D}[w_\zeta^2(s, a)]. \quad (1)$$

In this objective, we optimize a parameterized policy $\pi_\theta \in \Pi$ with $\theta \in \Theta$ being its parameters, and the policy class Π can be non-convex. We do so with the help of linearly parameterized functions $w_\zeta \in \mathcal{W}$ and $Q_\xi \in \mathcal{Q}$, which are respectively parameterized by $(\zeta, \xi) \in Z \times \Xi$ and serve as approximators of the density ratio and the value functions. We assume the parameter spaces $\Theta \in \mathbb{R}^{\dim(\Theta)}, Z \subset \mathbb{R}^{\dim(Z)}$ and

$\Xi \subset \mathbb{R}^{\dim(\Xi)}$ are all convex sets. ν_D and d^D are the empirical approximations of the initial state distribution and the state-action distribution in the data; see Sec. 2.1 for a formal definition. $Q_\xi(s, \pi_\theta)$ is short for $\mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[Q_\xi(s, a)]$. [1]

The main contributions of this paper is three-fold, which we summarize below:

A Detailed Analysis of Bias We identify the inevitable bias between the stationary points of $\mathcal{L}^D(\pi_\theta, w_\zeta, Q_\xi)$ and $J(\pi_\theta)$, where $J(\pi_\theta)$ is the expected return of π_θ. We separate out and characterize the bias terms due to regularization (ε_{reg}), generalization (ε_{func}), and mis-specification (ε_{data}), as will be detailed in Sec. 3. Besides, our analysis also reveals how λ_Q and λ_w affect the trade-off between the convergence speed and the magnitude of bias. Since the bias is unavoidable, we focus on the convergence to the points satisfying the following inequality, which we call the *biased stationary points* of $J(\pi_\theta)$ up to ε error.

$$\mathbb{E}[\|\nabla_\theta J(\pi_\theta)\|] \leq \varepsilon + \varepsilon_{data} + \varepsilon_{func} + \varepsilon_{reg} \qquad (2)$$

where $O(\cdot)$ only suppresses absolute constants. [2] The rest of the paper then presents two optimization strategies for solving Eq.(1) and provides their convergence analyses.

First Strategy: P-SREDA In our first strategy, we convert the original max-max-min problem to a standard non-convex-strongly-concave problem $\min_{(\theta, \zeta) \in \Theta \times Z} \max_{\xi \in \Xi} -\mathcal{L}(\pi_\theta, w_\zeta, Q_\xi)$ by simultaneously optimizing θ and ζ in the outer min. Unfortunately, most of the existing non-convex-strongly-concave optimization algorithms can not be adapted directly here, because they focus on $\min_{x \in \mathbb{R}^d} \max_{y \in \mathcal{Y}} f(x, y)$ where the first player can play an arbitrary vector in $\mathbb{R}^d$, while our objective function requires boundedness of ζ and ξ in order to guarantee the smoothness of $\mathcal{L}^D$, as we will show in the proof of Property B.1 in Appendix B. To

tackle this challenge, we propose P-SREDA, which is adapted from the SREDA algorithm [Luo et al., 2020] by including a projection step every time after updating ζ. The proof of P-SREDA is non-trivial because the projection step will incur additional error and some crucial steps in the original proof no longer hold in our case. We overcome these difficulties by leveraging the properties of $\mathcal{L}^D$ and carefully choosing the projection sets for Z and Ξ, and prove that its convergence rate remains $O(\varepsilon^{-3})$, which matches the lower bound in non-convex optimization [Arjevani et al., 2019].

Second Strategy: O-SPIM In our second strategy for solving Eq.(1), we study a novel actor-critic style framework called O-SPIM (**O**racle-based **S**tochastic **P**olicy **I**mprovement with **M**omentum). In O-SPIM, we alternate between updating ζ and ξ (the critic update) and updating θ with momentum to achieve a better trade-off (the actor update). The main technique difficulty is that different from policy gradient and actor-critic algorithms in the on-policy setting, here we need to coordinate the updates of three objects θ, ζ and ξ, and the loss $\mathcal{L}^D(\pi_{(\cdot)}, w_{\zeta_t}, Q_{\xi_t})$—when viewed solely as a function of the policy parameter θ with fixed parameters for w and Q—varies across iterations because ζ_t, ξ_t are updated in the critic step. We handle this difficulty by considering a family of critic update rules that satisfy a general condition (Condition 5.1), which enables us to relate the variations $\|\zeta_{t+1} - \zeta_t\|$ and $\|\xi_{t+1} - \xi_t\|$ with $\|\theta_{t+1} - \theta_t\|$ and is crucial to establishing convergence guarantee. We use ORACLE to refer to all critic update subroutines satisfying such a condition, and present two concrete instantiations and analyze their convergence properties in Appendices E.2 and E.3. The first one is a least-square algorithm, and the second one is a first-order algorithm by extending the SVRE [Chavdarova et al., 2019] from finite-sum setting to stochastic setting. Using either instantiation (or possibly other subroutines that satisfy our conditions), the convergence rate of our second strategy is $O(\varepsilon^{-4})$, which matches the rate of recent actor-critic algorithms in the on-policy setting [Xu et al., 2020, Fu et al., 2020]

1.1 Related works

Along with the progress of variance reduction techniques for non-convex optimization, there are emerging works analyzing convergence rates for on-policy policy gradient methods [Papini et al., 2018, Xu et al., 2019a,b, Yuan et al., 2020, Huang et al., 2020a]. Besides, in another line of work [Fu et al.,

---

[1] In this paper, we only give analysis for $\max_{\pi \in \Pi} \max_{w \in \mathcal{W}} \min_{Q \in \mathcal{Q}} \mathcal{L}_D(\pi, w, Q)$. In fact, we may swap the role of w and Q function and obtain another objective function $\max_{\pi \in \Pi} \max_{Q \in \mathcal{Q}} \min_{w \in \mathcal{W}} \mathcal{L}'_D(\pi, w, Q)$, where $\mathcal{L}'_D$ is the same as $\mathcal{L}$ except that both λ_Q and λ_w are negative. In general, the solutions of Eq.(6) and those of this new objective are not the same (see [Jiang and Huang, 2020] for their connections), but we can give a similar anlysis for $\mathcal{L}'$ with the techniques in this paper and establish similar convergence results.

[2] All norms $\|\cdot\|$ in this paper is ℓ_2 norm for vectors and operator norm for matrices unless specified otherwise. The expectation is over the randomness of the algorithm (e.g., the randomness in SGD) and not that of the data.

2020, Wu et al., 2020, Hong et al., 2020, Xu et al., 2020], finite-time guarantees are established for actor-critic style algorithms. However, all of these works require online interaction with the environment, whereas our focus is the offline setting.

Turning to the offline setting, there have been substantial empirical developments and success for off-policy PG or actor-critic algorithms [Degris et al., 2012, Lillicrap et al., 2015, Haarnoja et al., 2018, Fujimoto et al., 2018], but most of these works ignore the distribution-shift issue and do not provide convergence guarantees. Recently, there has been a lot of interest in MIS methods for off-policy evaluation [Liu et al., 2018, Uehara et al., 2019, Zhang et al., 2019a, Nachum et al., 2019a] and turning them into off-policy policy-optimization algorithms. Among them, Liu et al. [2019] 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. Nachum et al. [2019b], Jiang and Huang [2020] discussed policy optimization given arbitrary off-policy dataset, but no convergence analysis was performed. Besides, [Zhang et al., 2019c] presented a provably convergent algorithm under a similar linear setting with emphatic weighting [Imani et al., 2018], but only showed asymptotic convergence behavior and did not establish finite convergence rate.

In another work concurrent to ours [Xu et al., 2021], the authors designed DR-Off-PAC, which is motivated by the doubly-robust estimator and their objective function is similar to ours. However, they fixed the coefficient of regularization terms and their bias analysis is much coarser than ours. As a result, their analysis cannot distinguish between the errors resulting from regularization and mis-specification, and do not characterize how the regularization weights affect the trade-off between bias and convergence speed. Besides, the convergence rate of their algorithm is $O(\varepsilon^{-4})$, which matches our second strategy but worse than our first strategy. Recently, [Lyu et al., 2020] developed VOMPS/ACE-STORM based on Geoff-PAC [Zhang et al., 2019b] and the STORM estimator [Cutkosky and Orabona, 2020]. However, as also pointed by [Xu et al., 2021], the unbiasness of their estimator only holds asymptotically, and the convergence property of their algorithms is still unclear.

We summarize the comparison to closely related works in Table 1 in Appendix.

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.

Fixing an 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 $J(\pi) = \mathbb{E}_{s \sim \nu_0, a \sim \pi}[Q^\pi(s, a)] = \frac{1}{1-\gamma}\mathbb{E}_{s,a \sim d^\pi}[r(s, a)]$ as the expected return of policy π. If π is parameterized by θ and differentiable, we have

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

where the first step is the policy-gradient theorem [Sutton et al., 2000], and in the second step, we replace d^π with distribution μ, a state-action distribution generated by some behavior policies, and introduce the density ratio $w^\pi(s, a) := \frac{d^\pi(s,a)}{\mu(s,a)}$.

In the rest of the paper, we assume we are only provided with a fixed off-line dataset $D = \{(s_i, a_i, r_i, s_i')\}_{i=1}^{|D|}$, where each tuple is sampled according to $s_i, a_i \sim \mu, r_i \sim R(s_i, a_i), s_i' \sim P(\cdot|s_i, a_i)$. Besides, we do not require the knowledge of behavior policies that generate μ. As we also mentioned in the introduction, we will use d^D to denote the empirical state-action distribution induced from the dataset, which is defined by $d^D(s, a) = \frac{1}{|D|}\sum_{(s_i, a_i, r_i, s_i') \in D} \mathbb{I}[s_i = s, a_i = a]$.

Since the state-action space can be very large, we consider generalization via linear function approximation:

Definition 2.1 (Linear function classes)**.** Suppose we have two feature maps $\{\phi_W : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{\dim(Z)}\}$ and $\{\phi_Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{\dim(\Xi)}\}$ subject to $\|\phi_w(\cdot,\cdot)\| \leq 1, \|\phi_Q(\cdot,\cdot)\| \leq 1$. Let $Z \subset \mathbb{R}^{\dim(Z)}, \Xi \subset \mathbb{R}^{\dim(\Xi)}$ be the corresponding parameter spaces, respectively, satisfying $C_{\mathcal{W}} := \max\{1, \max_{\zeta \in Z} \|\zeta\|\} < \infty$ and $C_{\mathcal{Q}} := \max_{\xi \in \Xi} \|\xi\| < \infty$. The approximated value function Q_ξ and density ratio w_ζ are represented by

$$w(\cdot,\cdot) = \phi_w(\cdot,\cdot)^\top \zeta, \qquad Q(\cdot,\cdot) = \phi_Q(\cdot,\cdot)^\top \xi.$$

We use $\mathbf{\Phi}_w \in \mathbb{R}^{|S||A| \times \dim(Z)}$ to denote the matrix whose (s,a)-th row is $\phi_w(s,a)^\top$, and use $\mathbf{K}_w$ to denote $\mathbf{\Phi}_w^\top \Lambda^D \mathbf{\Phi}_w$, where Λ^D is a diagonal matrix whose diagonal elements are $d^D(\cdot,\cdot)$. Similarly, we define $\mathbf{\Phi}_Q$ and $\mathbf{K}_Q$ for $\phi_Q(\cdot,\cdot)$. Besides, $\mathbf{M}_\pi$ denotes $\mathbf{\Phi}_w^\top \Lambda^D (I - \gamma \mathbf{P}_D^\pi) \mathbf{\Phi}_Q$, where $\mathbf{P}_D^\pi$ is the empirical transition matrix induced from the data distribution. Notes that if we never see some s in dataset, then the corresponding element in Λ^D should be 0, and therefore, the corresponding row in $\mathbf{P}_D^\pi$ can just be set arbitrarily without having any effects on the loss function. Under linear function classes, we can rewrite $\mathcal{L}^D$ in Eq.(1) as:

$$\mathcal{L}^D(\pi, \zeta, \xi) = (1-\gamma)(\nu_D^\pi)^\top \mathbf{\Phi}_Q \xi + \zeta^\top \mathbf{\Phi}_w^\top \Lambda^D R \\ - \zeta^\top \mathbf{M}_\pi \xi + \frac{\lambda_Q}{2} \xi^\top \mathbf{K}_Q \xi - \frac{\lambda_w}{2} \zeta^\top \mathbf{K}_w \zeta. \tag{3}$$

We will also denote Λ as the diagonal matrix whose diagonal elements are $\mu(\cdot,\cdot)$, and denote P^π as the transition matrix of π.

2.2 Assumptions

Our first assumption is about the policy class, which is practical and frequently considered in the literature [Zhang et al., 2019c, Wu et al., 2020]. In this paper, we choose $\Theta = \mathbb{R}^{\dim(\Theta)}$ and we will interchangeably use Θ and $\mathbb{R}^{\dim(\Theta)}$ to denote the space of policy parameter θ.

Assumption A (Smoothness of policy function)**.** For any $s, a \in \mathcal{S} \times \mathcal{A}$ and $\theta \in \Theta = \mathbb{R}^{\dim(\Theta)}$, $\pi_\theta(s,a)$ is second-order differentiable w.r.t. θ, and there exist constants L_Π, G and H, s.t. for arbitrary $\theta_1, \theta_2, \theta \in \Theta$:

$$\|\pi_{\theta_1}(\cdot|s) - \pi_{\theta_2}(\cdot|s)\|_1 \leq L_\Pi \|\theta_1 - \theta_2\|, \tag{4}$$

$$\|\nabla_\theta \log \pi_\theta(a|s)\| \leq G, \quad \|\nabla_\theta^2 \log \pi_\theta(a|s)\| \leq H. \tag{5}$$

Note that this assumption is equivalent to Assump. 4 in [Xu et al., 2021] and Assump. 3 in [Zhang et al., 2020].

We also assume $\mathbf{K}_w, \mathbf{K}_Q$ and $\mathbf{M}_\pi$ are non-singular.

Assumption B. Denote $v_{\min}(\cdot)$ as the function to return the minimum singular value. We assume that $v_w := v_{\min}(\mathbf{K}_w) > 0$ and $v_Q := v_{\min}(\mathbf{K}_Q) > 0$. Beisdes, there exists a positive constant v_M that, for arbitrary $\pi \in \Pi$ (i.e. $\forall \theta \in \Theta$), the matrix $\mathbf{M}_\pi$ is invertible, and $v_{\min}(\mathbf{M}_\pi) \geq v_{\mathbf{M}} > 0$.

Our assumption on v_Q and v_w is frequently considered in the literature, see e.g. Assump. 2 in Zhang et al. [2020]. Besides, v_w and v_Q can be computed from a holdout dataset. Comparing with Xu et al. [2021], our M_π is equivalent to matrix A in their Assump. 3 if we use the same feature to approximate Q and w. Although Xu et al. [2021] do not consider v_Q and v_w, they have additional constraints on another two matrices in their Assump. 3.

The next assumption characterizes the coverage of μ using a constant C, often known as the concentrability coefficient in the literature [Szepesvári and Munos, 2005, Chen and Jiang, 2019].

Assumption C (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

$$w^\pi(s,a) := \frac{d^\pi(s,a)}{\mu(s,a)} \leq C, \quad w_\mu^\pi(s,a) := \frac{d_\mu^\pi(s,a)}{\mu(s,a)} \leq C$$

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

Our Assump. C is slightly stronger than Assump. 1 in Xu et al. [2021] because we have additional constraints on $\|\frac{d_\mu^\pi}{\mu}\|_\infty$, but it is still weaker than assuming the behavior policy is exploration enough, which is used by Zhang et al. [2020] as their Assump. 1.

Finally, since ϕ_w and ϕ_Q has bounded norm, the following variance terms are therefore bounded, and we use $\sigma_{(\cdot)}$ to denote their minimal upper bounds:

Assumption D. [Variance] There are constants $\sigma_{\mathbf{K}}, \sigma_{\mathbf{M}}, \sigma_R$ and σ_ν satisfying

$$\mathbb{E}_{s,a \sim d^D}[\|\mathbf{K}_w - \phi_w(s,a)\phi_w(s,a)^\top\|^2] \leq \sigma_{\mathbf{K}}^2,$$
$$\mathbb{E}_{s,a \sim d^D}[\|\mathbf{K}_Q - \phi_Q(s,a)\phi_Q(s,a)^\top\|^2] \leq \sigma_{\mathbf{K}}^2,$$
$$\mathbb{E}_{s,a \sim d^D}[\|\mathbf{\Phi}_w^\top \Lambda^D R - \phi_w(s,a)r(s,a)\|^2] \leq \sigma_R^2,$$
$$\mathbb{E}_{s \sim \nu_0^D, a \sim \pi(\cdot|s)}[\|\mathbf{\Phi}_Q^\top \nu_D^\pi - \phi_Q(s,a)\|] \leq \sigma_\nu^2,$$
$$\mathbb{E}_{s,a,s' \sim d^D, a' \sim \pi(\cdot|s')}[\|\mathbf{M}_\pi - \phi_w(s,a)(\phi_Q(s,a) \\ - \gamma \phi_Q(s',a'))^\top\|] \leq \sigma_{\mathbf{M}}^2.$$

2.3 Useful Properties

Under Assumptions above, we have the following important property regarding the convexity, concavity and smoothness of our objective function, which will be central to our proofs. Due to space limits, we defer a more detailed version (Property B.1) in Appendix B, in addition to some other useful properties and their proofs.

Property 2.2 (Convexity, Concavity and Smoothness). *For arbitrary $\theta \in \Theta, \zeta \in Z$, $\mathcal{L}^D$ is μ_ξ-strongly convex w.r.t. $\xi \in \Xi$, and for arbitrary $\theta \in \Theta, \xi \in \Xi$, $\mathcal{L}^D$ is μ_ζ-strongly concave w.r.t. $\zeta \in Z$, where $\mu_\xi = \lambda_Q v_Q$ and $\mu_\zeta = \lambda_w v_w$. Besides, given that $C_\mathcal{Q}$ and $C_\mathcal{W}$ in Definition 2.1 is finite, $\mathcal{L}^D$ is L-smooth with finite L.*

In the rest of paper, we will use $\kappa_\zeta := \lambda_\zeta / L$ and $\kappa_\xi := \lambda_\xi / L$ as a short note of the condition number.

3 THE BIASED STATIONARY POINTS

As alluded to in the introduction, even if we optimize Eq.(1), we do not expect that it converges to the a true stationary point w.r.t. $J(\pi_\theta)$ due to regularization, function-class mis-specification, and finite-sample effects, which result in the three error terms in Eq.(2). More concretely, by the triangular inequality,

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

Optimizing our loss function $\mathcal{L}^D(\pi, w, Q)$ reduces the first term but leaves the second term intact. In the rest of this section we bound the second term by breaking it down to more basic quantities that can be bounded under further standard assumptions. To do so, we need some additional definitions: We use $\mathcal{L}(w, \pi, Q)$ (as well as $\mathcal{L}(\pi_\theta, w_\zeta, Q_\xi)$) to denote the asymptotic version of $\mathcal{L}^D$ as $|D| \to \infty$:

$$\max_{\pi \in \Pi} \max_{w \in \mathcal{W}} \min_{Q \in \mathcal{Q}} \mathcal{L}(\pi, w, Q)$$
$$:= \max_{\theta \in \Theta} \max_{\zeta \in Z} \min_{\xi \in \Xi} \mathcal{L}(\pi_\theta, w_\zeta, Q_\xi)$$
$$:= (1 - \gamma) \mathbb{E}_{s_0 \sim \nu_0}[Q_\xi(s_0, \pi_\theta)]$$
$$+ \mathbb{E}_\mu[w_\zeta(s, a)\big(r + \gamma Q_\xi(s', \pi_\theta) - Q_\xi(s, a)\big)]$$
$$+ \frac{\lambda_Q}{2} \mathbb{E}_\mu[Q_\xi^2(s, a)] - \frac{\lambda_w}{2} \mathbb{E}_\mu[w_\zeta^2(s, a)]. \quad (6)$$

Definition 3.1 (Generalization Error). We define $\bar{\varepsilon}_{data}$ to be the minimal value satisfying that, for arbitrary $\pi_\theta, w_\zeta, Q_\xi \in \Pi \times \mathcal{W} \times \mathcal{Q}$, we have:

$$|\mathcal{L}(\pi_\theta, w_\zeta, Q_\xi) - \mathcal{L}^D(\pi_\theta, w_\zeta, Q_\xi)| \leq \bar{\varepsilon}_{data},$$
$$\|\nabla_\theta \mathcal{L}(\pi_\theta, w_\mu^*, Q_\mu^*) - \nabla_\theta \mathcal{L}^D(\pi_\theta, w_\mu^*, Q_\mu^*)\|^2 \leq \bar{\varepsilon}_{data}$$

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

$\bar{\varepsilon}_{data}$ characterizes the uniform deviation between $\mathcal{L}$ and $\mathcal{L}^D$ ($\nabla_\theta \mathcal{L}$ and $\nabla_\theta \mathcal{L}^D$, resp.) and can be bounded by the sample size and the complexity of the function classes. Such uniform convergence bounds are standard and we do not further analyze it in this paper.

Definition 3.2 (Mis-specification Error). Denote $\| \cdot \|_\Lambda$ as the ℓ_2-norm weighted by Λ, and define $w_\mathcal{L}^\pi = \arg\max_{w \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \min_{Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \mathcal{L}(\pi, w, Q)$ given arbitrary $\pi \in \Pi$. We define

$$\varepsilon_1 := \max_{\pi \in \Pi} \min_{w \in \mathcal{W}} \|w - w_\mathcal{L}^\pi\|_\Lambda^2,$$
$$\varepsilon_2 := \max_{w \in \mathcal{W}, \pi \in \Pi} \min_{Q \in \mathcal{Q}} \|Q - \arg\min_{Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \mathcal{L}(\pi, w, Q)\|_\Lambda^2$$

Based on these definitions and Assumptions in Section 2.2, in the following theorem, we provide an upper bound for the bias between the stationary points of our objective function $\mathcal{L}^D$ and $J(\pi)$, which is one of our main contributions. The proofs are deferred to Appendix C.

Theorem 3.3. *[Bias] Based Assumptions in Section 2.2 and Definitions above, for arbitrary $\theta \in \Theta$, we have:*

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

The bias terms $\varepsilon_{reg}, \varepsilon_{func}$ and ε_{data} are defined by

$$\varepsilon_{func} := \frac{G}{1 - \gamma}\left(\sqrt{C\varepsilon_\mathcal{Q}} + C_\mathcal{W}\sqrt{\frac{\gamma\varepsilon_\mathcal{Q}C}{1 - \gamma}}\right.$$
$$\left. + \sqrt{\frac{\gamma\varepsilon_\mathcal{Q}\varepsilon_\mathcal{W}C}{1 - \gamma}} + \gamma C_\mathcal{Q}\sqrt{\varepsilon_\mathcal{W}}\right)$$
$$\varepsilon_{reg} := \frac{G}{1 - \gamma}\left(\frac{C^2}{(1 - \gamma)}(\frac{\lambda_w\lambda_Q}{1 - \gamma} + \lambda_w)\right.$$
$$+ \frac{\gamma C(\lambda_Q + \lambda_Q\lambda_w C)}{(1 - \gamma)^3}$$
$$\left. + \frac{C^2(\lambda_Q + \lambda_Q\lambda_w C)}{(1 - \gamma)^3}(\frac{\lambda_w\lambda_Q}{1 - \gamma} + \lambda_w)\sqrt{\frac{\gamma C}{1 - \gamma}}\right)$$
$$\varepsilon_{data} := (2\kappa_\zeta\kappa_\xi + 2\kappa_\zeta + 2\kappa_\xi + \sqrt{2}/2)\sqrt{2\bar{\varepsilon}_{data}}$$

where κ_ζ and κ_ξ is the condition number, and

$$\varepsilon_{\mathcal{W}} := 4\frac{\lambda_{\max}^2}{\lambda_Q \lambda_w}\varepsilon_1 + 2\frac{\lambda_{\max}}{\mu_\zeta}\varepsilon_2,$$

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

with $\lambda_{\max} := \max\{\lambda_Q, \lambda_w\}$.

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 decreases with sample size (assuming all function classes have bounded capacity and allow uniform convergence). ε_{reg} depends on the magnitude of regularization, and will decrease if λ_w and λ_Q decreases. As for ε_{func}, it depends on the approximation error $\varepsilon_{\mathcal{W}}$ and $\varepsilon_{\mathcal{Q}}$, which are proportional to ε_1 and ε_2. Besides, because μ_ζ should be proportional to λ_w, the coefficients before ε_1 and ε_2 will not vary significant as we change λ_w and λ_Q, as long as $\lambda_w \approx \lambda_Q$. ($\varepsilon_1$ and ε_2 themselves may change with λ_w and λ_Q). Overall, a large dataset, well-specified function classes, and small λ_w and λ_Q will result in a small total bias, while small λ_w and λ_Q can lead to weaker strong-concavity or strong-convexity of the loss function, resulting in slower convergence.

Based on the discussion above, our goal is to find stochastic optimization algorithms, which return a policy π_θ after consuming $\text{poly}(\varepsilon^{-1})$ samples from dataset (we omit the dependence on quantities such as μ_ζ and μ_ξ), satisfying the biased stationary condition in Eq.(2)

Global Convergence Under Polyak-Łojasiewicz Condition If we assume the policy function class satisfies certain additional conditions, we can establish the global convergence guarantee. Take the Polyak-Łojasiewicz condition as an example, which requires that, there exists a constant $c_{P\mathrm{L}} > 0$, s.t., for arbitrary $\theta \in \Theta$

$$\frac{1}{2}\|\nabla_\theta J(\pi_\theta)\|^2 \geq c_{P\mathrm{L}}(J^* - J(\pi_\theta)). \qquad (7)$$

where J^* is the value of the optimal policy. As a result, if the policy class satisfies the PL inequality, a biased stationary point θ in Eq.(2) indicates that,

$$\mathbb{E}[J(\pi_\theta)] \geq J^* - \frac{1}{c_{P\mathrm{L}}}(\varepsilon^2 + \varepsilon_{data}^2 + \varepsilon_{func}^2 + \varepsilon_{reg}^2).$$

Such condition is provably weaker than the strongly convexity/concavity [Karimi et al., 2020].

4 Strategy 1: Projected-Stochastic Recursive Gradient Descent Ascent

In our first optimization strategy, we rewrite the original max-max-min problem as a max-min:

$$\arg\max_{\theta \in \mathbb{R}^{d_\theta}} \max_{\zeta \in Z} \min_{\xi \in \Xi} \mathcal{L}^D(\theta, \zeta, \xi)$$

$$\to \arg\max_{\theta,\zeta \in \mathbb{R}^{d_\theta} \times Z} \min_{\xi \in \Xi} \mathcal{L}^D(\theta, \zeta, \xi) = \arg\min_{\theta,\zeta \in \mathbb{R}^{d_\theta} \times Z} \max_{\xi \in \Xi} \mathcal{L}_-^D(\theta, \zeta, \xi)$$

where we use $\mathcal{L}_-^D$ as a shorthand of $-\mathcal{L}^D$. Given Assumptions in Sec. 2.2, we know $\arg\min_{\theta,\zeta \in \mathbb{R}^{d_\theta} \times Z} \max_{\xi \in \Xi} \mathcal{L}_-^D(\theta, \zeta, \xi)$ is a non-convex-strongly-concave problem, which has been the focus of some prior works [Lin et al., 2019, Luo et al., 2020]. That said, most of them target at $\min_{x \in \mathbb{R}^d} \max_{y \in \mathcal{Y}} f(x, y)$ where the first player can play an arbitrary vector in $\mathbb{R}^d$, while in our setting, ζ and ξ have to be constrained in bounded sets to guarantee that $\mathcal{L}^D$ (and $\mathcal{L}_-^D$) is L-smooth with finite L (see the proof of Property B.1 in Appendix B for details).

Therefore, we introduce Projected-SREDA (Algorithm 1), where we project ζ back to the convex set Z every time after update. However, the proofs for original SREDA [Luo et al., 2020] cannot be adapted directly because the projection step will incur extra error. To overcome this difficulty, we first study the following property of our objective function, which illustrate that the saddle-points of $\mathcal{L}^D$ given arbitrary $\pi \in \Pi$ have bounded l_2-norm.

Property 4.1. *Denote $Z_0 = \{\zeta|\|\zeta\| \leq R_\zeta\}$ and $\Xi_0 = \{\xi|\|\xi\| \leq R_\xi\}$ with $R_\zeta := \frac{1}{\lambda_w \lambda_Q v_w + v_M^2}(\frac{1-\gamma^2}{v_Q} + \lambda_Q)$ and $R_\xi := \frac{1}{\lambda_w \lambda_Q v_Q + v_M^2}((1-\gamma)\lambda_w + \frac{1+\gamma}{v_w})$. For all $\pi \in \Pi$, we have $(\zeta_\pi^*, \xi_\pi^*) := \arg\max_{\zeta \in \mathbb{R}^{\dim(Z)}} \min_{\xi \in \mathbb{R}^{\dim(\Xi)}} \mathcal{L}^D(\pi_\theta, w_\zeta, Q_\xi) \in Z_0 \times \Xi_0$.*

As a result, we have Lemma 4.2 below, which is a key step to control the projection error. We carefully choose the projection set Z and Ξ; as we increase the diameter of Z (i.e. R') and fix the diameter of Ξ as R_0, the angle between $\nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k)$ and the projection direction $\zeta_{k+1} - \zeta_{k+1}^+$ increases, which reflects the decrease of the relative projection error.

Lemma 4.2. *According to Alg 1, at iteration k, we have $\zeta_k \in Z$, $\zeta_{k+1}^+ = \zeta_k - \eta_k \mathbf{v}_k^\zeta$ and $\zeta_{k+1} = P_Z(\zeta_{k+1}^+)$. Denote $R_0 := \frac{1}{\lambda_w v_w}(1 + (1+\gamma)R_\xi)$ where R_ξ is defined in Property 4.1. If we choose $\Xi = \Xi_0$ and $Z = \{\zeta|\|\zeta\| \leq R'\}$ with $R' \geq R_0$, we have:*

$$\langle \nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k), \zeta_{k+1} - \zeta_{k+1}^+ \rangle$$

$$\leq \eta_k \frac{R_0 + \eta_k \|\mathbf{v}_k^\zeta\|}{R' + \eta_k \|\mathbf{v}_k^\zeta\|} \|\nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k)\| \|\mathbf{v}_k^\zeta\|.$$

Based on this lemma and an appropriate choice of R', we can show that Algorithm 1 converges to the biased stationary points. We omit the detail of the hyper-parameter choices due to space limits, and a detailed version of the following theorem and its proof are deferred to Appendix D.

Theorem 4.3 (Informal). *For $\varepsilon < 1$, under Assumptions in Section 2.2, by choosing $\Xi = \Xi_0$ and $Z = \{\zeta | \|\zeta\| \leq R' := 8 \max\{R_0, 1\}\}$, Algorithm 1 will return $\widehat{\theta}$ satisfy the following with $O(\varepsilon^{-3})$ stochastic gradient evaluations.*

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

Algorithm 1: Projected SREDA

---

1 **Input**: initial point (θ_0, ζ_0), learning rates η_k, $\lambda > 0$, batch size $S_1, S_2 > 0$; periods $q, m > 0$, number of initial iterations K_0; Convex Sets Ξ and Z; Two functions PiSARAH and ConcaveMaximizer (Alg. 4) in [Luo et al., 2020].

2 $\xi_0 = \text{PiSARAH}(-\mathcal{L}_-^D(\theta_0, \zeta_0, \cdot), K_0)$

3 **for** $k = 0, ..., K - 1$ **do**

4   **if** $\text{mod}(k, q) = 0$ **then**

5   draw S_1 samples and compute:

6 $\mathbf{v}_k := (\mathbf{v}_k^\theta, \mathbf{v}_k^\zeta) = \nabla_{\theta, \zeta} \mathcal{L}_-^{S_1}(\theta_k, \zeta_k, \xi_k),$

7 $\mathbf{u}_k = \nabla_\xi \mathcal{L}_-^{S_1}(\theta_k, \zeta_k, \xi_k)$

8   **end**

9   **else**

10 $\mathbf{v}_k = \mathbf{v}_k', \quad \mathbf{u}_k = \mathbf{u}_k'$

11   **end**

12 $\theta_{k+1} = \theta_k - \eta_k \mathbf{v}_k^\theta, \quad \zeta_{k+1}^+ = \zeta_k - \eta_k \mathbf{v}_k^\zeta,$

13 $\zeta_{k+1} = P_Z(\zeta_{k+1}^+)$

14 $(\xi_{k+1}, \mathbf{v}_{k+1}', \mathbf{u}_{k+1}') = $
     ConcaveMaximizer(k, m,
     $S_2, (\theta_k, \zeta_k), (\theta_{k+1}, \zeta_{k+1}), \xi_k, \mathbf{u}_k, \mathbf{v}_k, \lambda)$

15 **end**

16 **Output**: $(\widehat{\theta}, \widehat{\zeta})$ chosen uniformly at random from $\{(\theta_k, \zeta_k)\}_{k=0}^{K-1}$

---

5 Strategy 2: Oracle-based Stochastic Policy Improvement with Momentum

A natural question is that, can we solve max-max-min objective directly without merging $\max_\theta \max_\zeta$

together? We answer it firmly in this section by proposing an off-policy actor-critic style algorithm named O-SPIM in Algorithm 2, where we solve ζ and ξ with an abstract subroutine ORACLE in the critic step, and update θ with momentum in the actor step. Our analysis of O-SPIM also contributes to the understanding of convergence properties of actor-critic algorithms in the off-policy setting.

Different from the on-policy setting, where distribution shift is not a problem, our objective function has three variables to optimize, which causes difficulty in analyzing convergence property. Therefore, we enforce the following condition on ORACLE to coordinate the actor and critic steps.

Condition 5.1. For any strongly-concave-strongly-convex objective $\max_{\zeta \in Z} \min_{\xi \in \Xi} \mathcal{L}^D(\theta, \zeta, \xi)$ with saddle point $(\zeta^*, \xi^*) \in Z \times \Xi$, and arbitrary $0 \leq \beta \leq 1$ and $c > 0$, starting from a random initializer $(\bar{\zeta}, \bar{\xi}) \in Z \times \Xi$, ORACLE can return a solution $(\widehat{\zeta}, \widehat{\xi})$ satisfying

$$\mathbb{E}[\|\widehat{\zeta} - \zeta^*\|^2 + \|\widehat{\xi} - \xi^*\|^2]$$
$$\leq \frac{\beta}{2} \mathbb{E}[\|\bar{\zeta} - \zeta^*\|^2 + \|\bar{\xi} - \xi^*\|^2] + c. \quad (8)$$

Algorithm 2: Stochastic Momentum Policy Improvement with Oracle

---

1 **Input**: Total number of iteration T; Learning rate $\eta_\theta, \eta_\zeta, \eta_\xi$; Dataset distribution d^D; ORACLE parameter β.

2 Set $Z = Z_0$ and $\Xi = \Xi_0$ with Z_0 and Ξ_0 defined in Property 4.1.

3 Initialize $\theta_0, \zeta_{-1}, \xi_{-1}$.

4 $\zeta_0, \xi_0 \leftarrow \text{Oracle.init}(T_1, \eta_\zeta, \eta_\xi, \theta_0, \zeta_{-1}, \xi_{-1}, d^D)$.

5 Sample $B_0 \sim d^D$ with batch size $|B_0|$ and estimate batch gradient $g_\theta^0 = \nabla_\theta \mathcal{L}^{B_0}(\theta_0, \zeta_0, \xi_0)$.

6 **for** $t = 0, 1, 2, ...T - 1$ **do**

7 $\theta_{t+1} \leftarrow \theta_t + \eta_\theta g_\theta^t;$

8 $\zeta_{t+1}, \xi_{t+1} \leftarrow \text{Oracle}(\beta, \theta_{t+1}, \zeta_t, \xi_t, d^D);$

9   Sample $B \sim d^D;$

10 $g_\theta^{t+1} \leftarrow (1 - \alpha) g_\theta^t + \alpha \nabla_\theta \mathcal{L}^B(\theta_{t+1}, \zeta_{t+1}, \xi_{t+1})$
     // update g_θ with batch gradient

11 **end**

12 **Output**: Sample $\widehat{\theta} \sim \text{Unif}\{\theta_0, \theta_1, ..., \theta_T\}$ and output $\pi_{\widehat{\theta}}$.

5.1 Analysis of O-SPIM

The key insight of Condition 5.1 is the following lemma, in which we make a connection between the actor step and the critic step. As a result, in addition

to the smoothness of $\mathcal{L}^D$, we can control the shift between $\nabla \mathcal{L}^D(\theta_t, \zeta_t, \xi_t)$ and $\nabla \mathcal{L}^D(\theta_{t+1}, \zeta_{t+1}, \xi_{t+1})$ only with the shift of θ (i.e. $\eta_\theta g_\theta^\tau$) and hyperparameters β and c, which paves the way to comparing $\nabla_\theta \mathcal{L}^D(\theta, \zeta_t, \xi_t)$ with $\nabla_\theta J(\pi_\theta)$ and finally establishing the convergence guarantee.

Lemma 5.2. *[Relate the shift of ζ_t and ξ_t with θ_t] Denote $(\theta_t, \zeta_t, \xi_t)$ as the parameter value at the beginning at the step t in Algorithm 2, and denote $(\zeta_t^*, \xi_t^*) = \arg\max_{\zeta \in Z} \min_{\xi \in \Xi} \mathcal{L}^D(\theta_t, \zeta, \xi)$ as the only saddle point given θ_t. Under Assumptions in Section 2.2, we have:*

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

where $d := \max\{C_{\mathcal{W}}, C_{\mathcal{Q}}\}$ is the maximum of diameters of Z and Ξ, $C_{\zeta,\xi}$ is a short note of $\kappa_\mu^2(\kappa_\xi + 1)^2 + \kappa_\xi^2(\kappa_\mu + 1)^2$, $g_\theta^\tau := \frac{1}{\eta_\theta}(\theta^{\tau+1} - \theta^\tau)$, and η_θ is the step size of θ.

In the actor step, we introduce another hyperparameter α and adopt a momentum-based update rule, aiming at a better trade-off between the variance of the gradient estimation in the current step and the bias of using accumulative gradient in previous steps. As we will show in the proof of our main theorem (Theorem F.4 in Appendix F), by choosing α appropriately, although the trade-off cannot improve the dependence of ε in the convergence rate, it can indeed reduce the upper bound of $\mathbb{E}[\|\nabla_\theta J(\widehat{\theta})\|]$ comparing with the case without momentum (i.e. $\alpha = 1$)

Now, we are ready to state the main theorem of our second strategy. We defer the formal version including the hyper-parameter choices and its proof to Theorem F.4 in Appendix F.

Theorem 5.3 (Informal). *Under Assumptions in Section 2.2, given arbitrary ε, with appropriate hyper-parameter choices, by using either Algorithm 3 or Algorithm 4 as Oracle, Algorithm 2 will return us a policy $\pi_{\widehat{\theta}}$ with total complexity $O(\varepsilon^{-4})$, satisfying*

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

5.2 Concrete Examples of Oracle

We provide two concrete examples of ORACLE, and defer the algorithm details and the discussion about related works to Appendix E.

5.2.1 The Least-Square Oracle

In the linear setting, the saddle-point (ζ^*, ξ^*) has a closed form, which can be regarded as the regularized LSTD-Q solution [Kolter and Ng, 2009] and can be represented by $\mathbf{K}_w, \mathbf{K}_Q, \mathbf{M}_\pi$ and so on. Therefore, a natural solution is to estimate these matrices from (a subsample of) the dataset and plug them into the closed-form solution. We describe this idea in Alg. 3 and show that it satisfies our definition of ORACLE in Appendix E.2.

In Alg. 3, we use $\mathbf{u}_R$ to denote $\mathbf{\Phi}_w^\top \Lambda^D R$ and use $\mathbf{u}_\nu^\pi$ to denote $\mathbf{\Phi}_Q^\top \nu_D^\pi$. Besides, we use $\widehat{(\cdot)}$ to denote the empirical version of the matrices/vectors estimated via samples. Besides, $P_\Omega(x)$ means projected x into set Ω.

Algorithm 3: Projected Least-Square Oracle

1 **Input**: Distribution d^D; Batch Size N_{all};

2 $\zeta \leftarrow \left(\lambda_w \lambda_Q \widehat{\mathbf{K}}_w + \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1} \widehat{\mathbf{M}}_\pi^\top \right)^{-1} \left(-(1 - \gamma)\widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1} \widehat{\mathbf{u}}_\nu^\pi + \lambda_Q \widehat{\mathbf{u}}_R \right).$

3 $\xi \leftarrow \left(\lambda_w \lambda_Q \widehat{\mathbf{K}}_Q + \widehat{\mathbf{M}}_\pi^\top \widehat{\mathbf{K}}_w^{-1} \widehat{\mathbf{M}}_\pi \right)^{-1} \left((1 - \gamma)\lambda_w \widehat{\mathbf{u}}_\nu^\pi + \widehat{\mathbf{M}}_\pi^\top \widehat{\mathbf{K}}_w^{-1} \widehat{\mathbf{u}}_R \right).$

4 **Output**: $P_Z(\zeta), P_\Xi(\xi)$

5.2.2 Stochastic Variance-Reduced Extragradient with Batch Data

Similar to the LSTD, the per-step computational complexity of the first solver is quadratic in the dimension of the feature, which can be expensive for high dimensional features. Therefore, in Alg. 4, we present a first-order algorithm inspired by Stochastic Variance Reduced Extra-gradient [Chavdarova et al., 2019], which reduces the per-step complexity to $O(d)$ and also satisfies the ORACLE condition. Besides, Alg. 4 can also handle general strongly-convex-strongly-concave problem beyond the linear setting.

In the algorithm, $\mathcal{P}_\zeta$ and $\mathcal{P}_\xi$ are projection operators; $\nabla \mathcal{L}^N(\theta, \zeta, \xi)$ denotes the average gradient over samples from mini batch data N. We also define:

$$d_\zeta^N(\zeta_1, \xi_1, \zeta_2, \xi_2) = \nabla_\zeta \mathcal{L}^N(\theta, \zeta_1, \xi_1) - \nabla_\zeta \mathcal{L}^N(\theta, \zeta_2, \xi_2)$$
$$d_\xi^N(\zeta_1, \xi_1, \zeta_2, \xi_2) = \nabla_\xi \mathcal{L}^N(\theta, \zeta_1, \xi_1) - \nabla_\xi \mathcal{L}^N(\theta, \zeta_2, \xi_2).$$

It is clear that

$$\mathbb{E}[g_k^\zeta] = \nabla_\zeta \mathcal{L}^D(\theta, \zeta_k, \xi_k),$$
$$\mathbb{E}[g_{k+1/2}^\zeta] = \nabla_\zeta \mathcal{L}^D(\theta, \zeta_{k+1/2}, \xi_{k+1/2}),$$

Algorithm 4: Stochastic Variance-Reduced Extragradient with Mini Batch Data (SVREB)

---

1 **Input**: Stopping time K; learning rates η_ζ, η_ξ; Initial wegihts ζ_0, ξ_0; Distribution d^D; Batch size $|N|$.

2 Sample mini batch $N_\zeta, N_\xi \sim d^D$ with batch size $|N|$

3 $g_0^\zeta \leftarrow \nabla_\zeta \mathcal{L}^{N_\zeta}(\theta, \zeta_0, \xi_0), \quad g_0^\xi \leftarrow \nabla_\xi \mathcal{L}^{N_\xi}(\theta, \zeta_0, \xi_0).$

4 $\zeta_1 \leftarrow \mathcal{P}_\zeta(\zeta_0 + \eta_\zeta g_0^\zeta).$

5 $\xi_1 \leftarrow \mathcal{P}_\xi(\xi_0 - \eta_\xi g_0^\xi).$

6 $m_1^\zeta, m_1^\xi \leftarrow \nabla_\zeta \mathcal{L}^{N_\zeta}(\theta, \zeta_0, \xi_0), \nabla_\xi \mathcal{L}^{N_\xi}(\theta, \zeta_0, \xi_0).$

7 **for** $k = 1, 2, ...K+1$ **do**

8   Sample mini batch $N_\zeta, N_\xi \sim d^D$ with batch size $|N|$.

9     $g_k^\zeta = m_k^\zeta + d_\zeta^{N_\zeta}(\zeta_k, \xi_k, \zeta_{k-1}, \xi_{k-1}).$

10     $g_k^\xi = m_k^\xi + d_\xi^{N_\xi}(\zeta_k, \xi_k, \zeta_{k-1}, \xi_{k-1}).$

11     $\zeta_{k+1/2} = \mathcal{P}_\zeta(\zeta_k + \eta_\zeta g_k^\zeta).$

12     $\xi_{k+1/2} = \mathcal{P}_\xi(\xi_k - \eta_\xi g_k^\xi).$

13   Sample mini batch $N_\zeta', N_\xi' \sim d^D$ with batch size $|N|$.

14     $g_{k+1/2}^\zeta = $
    $m_k^\zeta + d_\zeta^{N_\zeta'}(\zeta_{k+1/2}, \xi_{k+1/2}, \zeta_{k-1}, \xi_{k-1}).$

15     $g_{k+1/2}^\xi = $
    $m_k^\xi + d_\xi^{N_\xi'}(\zeta_{k+1/2}, \xi_{k+1/2}, \zeta_{k-1}, \xi_{k-1}).$

16     $\zeta_{k+1} = \mathcal{P}_\zeta(\zeta_k + \eta_\zeta g_{k+1/2}^\zeta).$

17     $\xi_{k+1} = \mathcal{P}_\xi(\xi_k - \eta_\xi g_{k+1/2}^\xi).$

18   // The following has been computed in step 9 and 10

19     $m_{k+1}^\zeta, m_{k+1}^\xi \leftarrow$
    $\nabla_\zeta \mathcal{L}^{N_\zeta}(\theta, \zeta_k, \xi_k), \nabla_\xi \mathcal{L}^{N_\xi}(\theta, \zeta_k, \xi_k)$

20     $k \leftarrow k+1.$

21 **end**

22 **Output**: ζ_K, ξ_K

---

where the expectation only concerns the randomness of sample when computing g. The above relationship also holds if we consider gradient w.r.t. ξ.

For this Algorithm 4, we have the following theorem:

Theorem 5.4. *Under Assumptions in Section 2.2, in Algorithm 4, if step sizes satisfy*

$$\eta_\zeta \leq \frac{1}{50 \max\{\bar{L}_\zeta, \mu_\zeta\}}, \qquad \eta_\xi \leq \frac{1}{50 \max\{\bar{L}_\xi, \mu_\xi\}}$$

after K iterations, the algorithm will output (ζ_K, ξ_K):

$$\mathbb{E}[\|\zeta_K - \zeta^*\|^2 + \|\xi_K - \xi^*\|^2]$$
$$\leq \frac{201}{100}\Big(1 - \frac{\mu\eta}{4}\Big)^K \mathbb{E}[\|\zeta_0 - \zeta^*\|^2 + \|\xi_0 - \xi^*\|^2]$$
$$+ \frac{8\sigma^2}{\min\{\frac{\mu_\zeta \eta_\zeta}{4}, \frac{\mu_\xi \eta_\xi}{4}\}|N|}\Big(\frac{\eta_\zeta}{\mu_\zeta} + \frac{\eta_\xi}{\mu_\xi}\Big)$$

where (ζ^*, ξ^*) is the saddle point of $\mathcal{L}^D(\theta, \zeta, \xi)$ given input θ.

The theorem implies that Alg. 4 can be used as an oracle for arbitrary β and c as long as K and $|N|$ are chosen appropriately. We defer the proofs to Appx. E.3.1.

6 Conclusion

In this paper, we study two natural optimization strategies for density-ratio based off-policy policy optimization, establish their convergence rates, and characterize the quality of the output policies. In the future, there are several potentially interesting directions to study. (**i**) It would be interesting to investigate the possibility of improving the dependence on ε on the convergence rate of our second strategy. (**ii**) Some of our assumptions are still quite strong, such as Assump. B and C, where we assume that some condition holds for *any* policy $\pi \in \Pi$. An open question is whether we can derive convergence rate when those conditions holds only for the optimal policy. (**iii**) In this paper, we consider the linear approximation for $\mathcal{W}$ and $\mathcal{Q}$ in order to ensure the strongly concavity/convexity property of the loss function w.r.t. ζ and ξ so that the analysis is more tractable. It is still an open problem whether we can show some convergence property if $\mathcal{W}$ and $\mathcal{Q}$ are more complicated function classes.

Acknowledgements

References

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.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

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.

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.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in nonconvex sgd, 2020.

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

Simon S. Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. volume 70 of *Proceedings of Machine Learning Research*, pages 1049–1058, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/du17a.html.

Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. Single-timescale actor-critic provably finds globally optimal policy, 2020.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic, 2020.

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

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4422–4433. PMLR, 13–18 Jul 2020b. URL http://proceedings.mlr.press/v119/huang20a.html.

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.

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

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition, 2020.

J. Z. Kolter and A. Ng. Regularization and feature selection in least-squares temporal difference learning. In *ICML '09*, 2009.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning, 2015.

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

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

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient td algorithms, 2020.

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.

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.

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

Daoming Lyu, Qi Qi, Mohammad Ghavamzadeh, Hengshuai Yao, Tianbao Yang, and Bo Liu. Variance-reduced off-policy memory-efficient policy search, 2020.

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, 2019a.

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

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

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.

Csaba Szepesvári and Rémi Munos. Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd international conference on Machine learning*, pages 880–887, 2005.

Joel A. Tropp. An introduction to matrix concentration inequalities, 2015.

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

Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite time analysis of two time-scale actor critic methods, 2020.

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

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

Tengyu Xu, Zhe Wang, and Yingbin Liang. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms, 2020.

Tengyu Xu, Zhuoran Yang, Zhaoran Wang, and Yingbin Liang. Doubly robust off-policy actor-critic: Convergence and optimality, 2021.

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

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

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

Shangtong Zhang, Wendelin Boehmer, and Shimon Whiteson. Generalized off-policy actor-critic, 2019b.

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

Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation, 2020.

Table 1: Comparison of Convergence Guarantee of Recent Methods

Algorithms	Convergence Rate	Off-Policy?	Detailed Bias Analysis?
SVRPG [Xu et al., 2019a]	$O(\varepsilon^{-10/3})$	×	
SRVR-PG [Xu et al., 2019b]	$O(\varepsilon^{-3})$	×	
STORM-PG [Yuan et al., 2020]	$O(\varepsilon^{-3})$	×	N.A.
MBPG [Huang et al., 2020b]	$O(\varepsilon^{-3})$	×	
On Policy AC/NAC [Fu et al., 2020, Xu et al., 2020]	$O(\varepsilon^{-4})$	×	
DR-Off-PAC [Xu et al., 2021]	$O(\varepsilon^{-4})$	✓	×
P-SREDA (Ours)	$O(\varepsilon^{-3})$	✓	✓
O-SPIM (Ours)	$O(\varepsilon^{-4})$	✓	✓

A   Useful Lemma

Lemma A.1 (Lemma B.2 in [Lin et al., 2020]). *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 B, 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). *According to Danskin's Theorem,*

$$\nabla \Phi_\theta(\cdot) := \nabla_\zeta \mathcal{L}^D(\theta, \zeta, \phi_\theta(\zeta)) = \nabla_\zeta \mathcal{L}^D(\theta, \zeta, \xi)|_{\xi=\phi_\theta(\zeta)}$$

Therefore, when we compute $\nabla_\zeta \mathcal{L}^D(\theta, \zeta, \phi_\theta(\zeta))$, we can treat $\phi_\theta(\zeta)$ as a constant. Then, 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, we have*

$$\|x - x_f^*\| \le \frac{1}{\alpha}\|\nabla_x f(x)\|, \qquad \frac{\alpha}{2}\|x - x_f^*\|^2 \le f(x) - f(x_f^*) \tag{9}$$

$$\|x - x_g^*\| \le \frac{1}{\beta}\|\nabla_x g(x)\|, \qquad \frac{\beta}{2}\|x - x_f^*\|^2 \le g(x_g^*) - g(x) \tag{10}$$

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

$$(\nabla_x f(x) - \nabla_x f(x_f^*))^\top (x - x_f^*) \geq \alpha \|x - x_f^*\|^2, \qquad f(x) \geq f(x_f^*) + \nabla_x f(x_f^*)^\top (x - x_f^*) + \frac{\alpha}{2} \|x - x_f^*\|^2.$$

Since x_f^* is the minimizer of $f(x)$, we know that $\nabla_x f(x_f^*)^\top (x - x_f^*) \geq 0, \forall x \in X$. Combining all the above inequalities together and we obtain

$$\|x - x_f^*\|^2 \leq \frac{1}{\alpha} \nabla_x f(x)^\top (x - x_f^*) \leq \frac{1}{\alpha} \|\nabla_x f(x)\| \|x - x_f^*\|, \qquad f(x) \geq f(x_f^*) + \frac{\alpha}{2} \|x - x_f^*\|^2.$$

which implies

$$\|x - x_f^*\| \leq \frac{1}{\alpha} \|\nabla_x f(x)\|, \qquad \frac{\alpha}{2} \|x - x_f^*\|^2 \leq f(x) - f(x_f^*).$$

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

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

$$(\mathbf{A}^\top \mathbf{A})^{-1} \succ \left((\alpha \mathbf{I} + \mathbf{A})^\top (\alpha \mathbf{I} + \mathbf{A}) \right)^{-1}$$

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

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

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

Then, we only need to prove

$$(\alpha \mathbf{I} + \mathbf{A})^\top (\alpha \mathbf{I} + \mathbf{A}) \succ \mathbf{A}^\top \mathbf{A}$$

We have

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

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

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

Proof. For arbitrary initial state-action distribution vector $\mu_0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times 1}$, $(\mathbf{I} - \gamma \mathbf{P}_*^\pi)^{-1} \mu_0$ 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. $\qquad \square$

B Useful Properties Implied by Assumptions in 2.2

In this section, we first prove several properties implied by our basic assumptions in Section 2.2.

Property B.1. *[A detailed version of Property 2.2] Under Assumption A and B, given that $C_{\mathcal{W}} := \max\{1, \max_{\zeta \in Z} \|\zeta\|\}$ and $C_{\mathcal{Q}} := \max_{\xi \in \Xi} \|\xi\|$ are finite, we have:*

(a) For arbitrary $\theta \in \Theta, \zeta \in Z$, $\mathcal{L}^D$ is μ_ξ-strongly convex w.r.t. $\xi \in \Xi$, and for arbitrary $\theta \in \Theta, \xi \in \Xi$, $\mathcal{L}^D$ is μ_ζ-strongly concave w.r.t. $\zeta \in Z$, where $\mu_\xi = \lambda_Q v_Q$ and $\mu_\zeta = \lambda_w v_w$.

(b) For any $\xi, \xi_1, \xi_2 \in \Xi, \zeta, \zeta_1, \zeta_2 \in Z, (s,a) \in \mathcal{S} \times \mathcal{A}$,

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

(c) For any $\zeta_1, \zeta_2 \in Z, \xi_1, \xi_2 \in \Xi, \theta_1, \theta_2 \in \Theta$, $\mathcal{L}^D$ defined in Eq.(1) is differentiable, and there exists constant L s.t.

$$\|\nabla_\theta \mathcal{L}^D(\theta_1, \zeta_1, \xi_1) - \nabla_\theta \mathcal{L}^D(\theta_2, \zeta_2, \xi_2)\| + \|\nabla_\zeta \mathcal{L}^D(\theta_1, \zeta_1, \xi_1) - \nabla_\zeta \mathcal{L}^D(\theta_2, \zeta_2, \xi_2)\|$$
$$+ \|\nabla_\xi \mathcal{L}^D(\theta_1, \zeta_1, \xi_1) - \nabla_\xi \mathcal{L}^D(\theta_2, \zeta_2, \xi_2)$$
$$\leq L\|\theta_1 - \theta_2\| + L\|\zeta_1 - \zeta_2\| + L\|\xi_1 - \xi_2\|.$$

In other words, $\mathcal{L}^D$ is L-smooth when $\zeta \in Z, \xi \in \Xi, \theta \in \Theta$.

Proof.

Proof of (a) Since $\mathcal{L}^D$ is second-order differentible w.r.t. arbitrary $\zeta \in \mathbb{R}^{d_\zeta}$ and $\xi \in \mathbb{R}^{d_\xi}$, under Assumption B, we have:

$$\nabla_\zeta^2 \mathcal{L}^D = -\lambda_w K_w \prec -\lambda_w v_w \mathbf{I}, \qquad \nabla_\xi^2 \mathcal{L}^D = \lambda_Q K_Q \succ \lambda_Q v_Q \mathbf{I}.$$

where $\mathbf{I}$ is the identity matrix.

Proof of (b) Because w_ζ and Q_ξ are linear and features has bounded l_2-norm, and Z and Ξ are all convex sets with bounded radius, we have:

$$|Q_\xi(s,a)| \leq \|\phi_Q(s,a)^\top \xi\| \leq C_\mathcal{Q}, \qquad |Q_{\xi_1}(s,a) - Q_{\xi_2}(s,a)| \leq \|\xi_1 - \xi_2\|;$$
$$|w_\zeta(s,a)| \leq \|\phi_w(s,a)^\top \zeta\| \leq C_\mathcal{W}, \qquad |w_{\zeta_1}(s,a) - w_{\zeta_2}(s,a)| \leq \|\zeta_1 - \zeta_2\|;$$

Therefore, Property B.1-(b) holds.

Proof of (c) We will use $, w_1, Q_1, \pi_1$ and $\mathcal{L}_1^D$ as shortnotes of $w_{\zeta_1}, Q_{\xi_1}, \pi_{\theta_1}$ and $\mathcal{L}^D(\theta_1, \zeta_1, \xi_1)$, and the meaning of w_2, Q_2, π_2 and $\mathcal{L}_2^D$ are similar.

$$\|\nabla_\theta \mathcal{L}_1^D - \nabla_\theta \mathcal{L}_2^D\|$$
$$= \|(1-\gamma)\mathbb{E}_{\nu_D, a \sim \pi_1(\cdot|s_0)}[Q_1(s_0,a)\nabla_\theta \log \pi_1(a|s_0)] - (1-\gamma)\mathbb{E}_{\nu_D, a \sim \pi_2(\cdot|s_0)}[Q_2(s_0,a)\nabla_\theta \log \pi_2(a|s_0)]$$
$$+ \gamma \mathbb{E}_{d^D, a' \sim \pi_1(\cdot|s')}[w_1(s,a)Q_1(s',a')\nabla_\theta \log \pi_1(a'|s')] - \gamma \mathbb{E}_{d^D, a' \sim \pi_2(\cdot|s')}[w_2(s,a)Q_2(s',a')\nabla_\theta \log \pi_2(a'|s')]\|$$
$$\leq (1-\gamma)\|\mathbb{E}_{\nu_D, a \sim \pi_1(\cdot|s_0)}[(Q_1(s_0,a) - Q_2(s_0,a))\nabla_\theta \log \pi_1(a|s_0)]$$
$$+ \mathbb{E}_{\nu_D, a \sim \pi_1(\cdot|s_0) - \pi_2(\cdot|s_0)}[Q_2(s_0,a)\nabla_\theta \log \pi_1(a|s_0)] + \mathbb{E}_{\nu_D, a \sim \pi_2(\cdot|s_0)}[Q_2(s_0,a)(\nabla_\theta \log \pi_1(a|s_0) - \nabla_\theta \log \pi_2(a|s_0))]\|$$
$$+ \gamma \|\mathbb{E}_{d^D, a' \sim \pi_1(\cdot|s')}[(w_1(s,a) - w_2(s,a))Q_1(s',a')\nabla_\theta \log \pi_1(a'|s')]$$
$$+ \mathbb{E}_{d^D, a' \sim \pi_1(\cdot|s')}[w_2(s,a)(Q_1(s',a') - Q_2(s',a'))\nabla_\theta \log \pi_1(a'|s')]$$
$$+ \mathbb{E}_{d^D, a' \sim \pi_1(\cdot|s') - \pi_2(\cdot|s')}[w_2(s,a)Q_2(s',a')\nabla_\theta \log \pi_1(a'|s')]$$
$$+ \mathbb{E}_{d^D, a' \sim \pi_2(\cdot|s')}[w_2(s,a)Q_2(s',a')(\nabla_\theta \log \pi_1(a'|s') - \nabla_\theta \log \pi_2(a'|s'))]\|$$
$$\leq (1-\gamma)\Big(G\|\xi_1 - \xi_2\| + GC_\mathcal{Q}\mathbb{E}_{\nu_D}[\|\pi(\cdot|s_0) - \pi(\cdot|s_0)\|_1] + HC_\mathcal{Q}\|\theta_1 - \theta_2\|\Big)$$
$$+ \gamma\Big(GC_\mathcal{Q}\|\zeta_1 - \zeta_2\| + GC_\mathcal{W}\|\xi_1 - \xi_2\| + GC_\mathcal{W}C_\mathcal{Q}\mathbb{E}_{d^D}[\|\pi_1(\cdot|s') - \pi_2(\cdot|s')\|_1] + HC_\mathcal{W}C_\mathcal{Q}\|\theta_1 - \theta_2\|\Big)$$
$$\leq C_\mathcal{W}C_\mathcal{Q}(GL_\Pi + H)\|\theta_1 - \theta_2\| + GC_\mathcal{Q}\|\zeta_1 - \zeta_2\| + GC_\mathcal{W}\|\xi_2 - \xi_2\|$$

In the last inequality, we use $C_\mathcal{W} \geq 1$ and $0 < \gamma \leq 1$. Besides,

$$\|\nabla_\zeta \mathcal{L}_1^D - \nabla_\zeta \mathcal{L}_2^D\| = \|(M_{\pi_1} - M_{\pi_2})\xi_1 + M_{\pi_2}(\xi_1 - \xi_2) - \lambda_w K_w(\zeta_1 - \zeta_2)\|$$

14

$$\leq \gamma C_{\mathcal{Q}} \mathbb{E}_{d^D}[\|\pi_1(\cdot|s') - \pi_2(\cdot|s')\|_1] + (1+\gamma)\|\xi_1 - \xi_2\| + \lambda_w\|\zeta_1 - \zeta_2\|$$
$$\leq \gamma C_{\mathcal{Q}} L_\Pi \|\theta_1 - \theta_2\| + (1+\gamma)\|\xi_1 - \xi_2\| + \lambda_w\|\zeta_1 - \zeta_2\|$$
$$\|\nabla_\xi \mathcal{L}_1^D - \nabla_\xi \mathcal{L}_2^D\| = \|(M_{\pi_1} - M_{\pi_2})^\top \zeta_1 + M_{\pi_2}^\top(\zeta_1 - \zeta_2) + \lambda_Q K_Q(\xi_1 - \xi_2)\|$$
$$\leq \gamma C_{\mathcal{W}} \mathbb{E}_{d^D}[\|\pi_1(\cdot|s') - \pi_2(\cdot|s')\|_1] + (1+\gamma)\|\zeta_1 - \zeta_2\| + \lambda_Q\|\xi_1 - \xi_2\|$$
$$\leq \gamma C_{\mathcal{W}} L_\Pi \|\theta_1 - \theta_2\| + (1+\gamma)\|\zeta_1 - \zeta_2\| + \lambda_Q\|\xi_1 - \xi_2\|$$

As a result,

$$\|\nabla_\theta \mathcal{L}^D(\theta_1, \zeta_1, \xi_1) - \nabla_\theta \mathcal{L}^D(\theta_2, \zeta_2, \xi_2)\| + \|\nabla_\zeta \mathcal{L}^D(\theta_1, \zeta_1, \xi_1) - \nabla_\zeta \mathcal{L}^D(\theta_2, \zeta_2, \xi_2)\|$$
$$+ \|\nabla_\xi \mathcal{L}^D(\theta_1, \zeta_1, \xi_1) - \nabla_\xi \mathcal{L}^D(\theta_2, \zeta_2, \xi_2)\|$$
$$\leq \Big(C_{\mathcal{W}} C_{\mathcal{Q}}(G L_\Pi + H) + \gamma(C_{\mathcal{Q}} + C_{\mathcal{W}})L_\Pi\Big)\|\theta_1 - \theta_2\| + \Big(G C_{\mathcal{Q}} + (1+\gamma) + \lambda_w\Big)\|\zeta_1 - \zeta_2\|$$
$$+ \Big(G C_{\mathcal{W}} + (1+\gamma) + \lambda_Q\Big)\|\xi_1 - \xi_2\|$$

Therefore, Condition A-(b) holds with

$$L = \max\{C_{\mathcal{W}} C_{\mathcal{Q}}(G L_\Pi + H) + \gamma(C_{\mathcal{Q}} + C_{\mathcal{W}})L_\Pi, G C_{\mathcal{Q}} + (1+\gamma) + \lambda_w, G C_{\mathcal{W}} + (1+\gamma) + \lambda_Q\} \quad (11)$$

$\square$

Property 4.1. *Denote* $Z_0 = \{\zeta | \|\zeta\| \leq R_\zeta\}$ *and* $\Xi_0 = \{\xi | \|\xi\| \leq R_\xi\}$ *with* $R_\zeta := \frac{1}{\lambda_w \lambda_Q v_w + v_M^2}(\frac{1-\gamma^2}{v_Q} + \lambda_Q)$ *and* $R_\xi := \frac{1}{\lambda_w \lambda_Q v_Q + v_M^2}((1-\gamma)\lambda_w + \frac{1+\gamma}{v_w})$. *For all* $\pi \in \Pi$, *we have* $(\zeta_\pi^*, \xi_\pi^*) := \arg\max_{\zeta \in \mathbb{R}^{\dim(Z)}} \min_{\xi \in \mathbb{R}^{\dim(\Xi)}} \mathcal{L}^D(\pi_\theta, w_\zeta, Q_\xi) \in Z_0 \times \Xi_0$.

Proof. Recall the definition of $\mathcal{L}^D$ in Eq.(3):

$$\mathcal{L}^D(\pi, \zeta, \xi) = (1-\gamma)(\nu_D^\pi)^\top \mathbf{\Phi}_Q \xi + \zeta^\top \mathbf{\Phi}_w^\top \Lambda^D R - \zeta^\top \mathbf{M}_\pi \xi + \frac{\lambda_Q}{2}\xi^\top \mathbf{K}_Q \xi - \frac{\lambda_w}{2}\zeta^\top \mathbf{K}_w \zeta.$$

by taking derivatives w.r.t. ξ and setting it to be zero, we have:

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

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

$$-\frac{\lambda_w}{2}\zeta^\top \mathbf{K}_w \zeta - \frac{1}{2\lambda_Q}\Big(\mathbf{M}_\pi^\top \zeta - (1-\gamma)\mathbf{\Phi}_Q^\top \nu_D^\pi\Big)^\top \mathbf{K}_Q^{-1}\Big(\mathbf{M}_\pi^\top \zeta - (1-\gamma)\mathbf{\Phi}_Q^\top \nu_D^\pi\Big) + \zeta^\top \mathbf{\Phi}_w^\top \Lambda^D R$$

Taking the derivative of ζ and set it to be zero, we have:

$$\zeta_\pi^* = \Big(\lambda_w \lambda_Q \mathbf{K}_w + \mathbf{M}_\pi \mathbf{K}_Q^{-1}\mathbf{M}_\pi^\top\Big)^{-1}\Big(-(1-\gamma)\mathbf{M}_\pi \mathbf{K}_Q^{-1}\mathbf{\Phi}_Q^\top \nu_D^\pi + \lambda_Q \mathbf{\Phi}_w^\top \Lambda^D R\Big)$$

and therefore,

$$\xi_\pi^* = \frac{1}{\lambda_Q}\mathbf{K}_Q^{-1}\Big(\mathbf{M}_\pi^\top \zeta_\pi^* - (1-\gamma)\mathbf{\Phi}_Q^\top \nu_D^\pi\Big)$$
$$= \frac{1}{\lambda_Q}\mathbf{K}_Q^{-1}\mathbf{M}_\pi^\top\Big(\lambda_w \lambda_Q \mathbf{K}_w + \mathbf{M}_\pi \mathbf{K}_Q^{-1}\mathbf{M}_\pi^\top\Big)^{-1} \cdot \Big(-(1-\gamma)\mathbf{M}_\pi \mathbf{K}_Q^{-1}\mathbf{\Phi}_Q^\top \nu_D^\pi + \lambda_Q \mathbf{\Phi}_w^\top \Lambda^D R\Big)$$
$$+ (1-\gamma)\frac{1}{\lambda_Q}\mathbf{K}_Q^{-1}\mathbf{\Phi}_Q^\top(\nu_D^\pi)^\top$$
$$= (1-\gamma)\lambda_w\Big(\lambda_w \lambda_Q \mathbf{K}_Q + \mathbf{M}_\pi^\top \mathbf{K}_w^{-1}\mathbf{M}_\pi\Big)^{-1}\mathbf{\Phi}_Q^\top \nu_D^\pi + \mathbf{K}_Q^{-1}\mathbf{M}_\pi^\top\Big(\lambda_w \lambda_Q \mathbf{K}_w + \mathbf{M}_\pi \mathbf{K}_Q^{-1}\mathbf{M}_\pi^\top\Big)^{-1}\mathbf{\Phi}_w^\top \Lambda^D R$$

$$=\left(\lambda_w\lambda_Q\mathbf{K}_Q + \mathbf{M}_\pi^\top\mathbf{K}_w^{-1}\mathbf{M}_\pi\right)^{-1}\left((1-\gamma)\lambda_w\mathbf{\Phi}_Q^\top\nu_D^\pi + \mathbf{M}_\pi^\top\mathbf{K}_w^{-1}\mathbf{\Phi}_w^\top\Lambda^D R\right)$$

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

$$(\lambda_w\lambda_Q\mathbf{K}_Q + \mathbf{M}_\pi^\top\mathbf{K}_w^{-1}\mathbf{M}_\pi)^{-1} = \frac{1}{\lambda_w\lambda_Q}\mathbf{K}_Q^{-1} - \frac{1}{\lambda_w\lambda_Q}\mathbf{K}_Q^{-1}\mathbf{M}_\pi^\top(\lambda_w\lambda_Q\mathbf{K}_w + \mathbf{M}_\pi\mathbf{K}_Q^{-1}\mathbf{M}_\pi^\top)\mathbf{M}_\pi\mathbf{K}_Q^{-1}$$

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

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

Therefore,

$$
\begin{aligned}
\|\zeta_\pi^*\| \leq &(1-\gamma)\|\left(\lambda_w\lambda_Q\mathbf{K}_w + \mathbf{M}_\pi\mathbf{K}_Q^{-1}\mathbf{M}_\pi^\top\right)^{-1}\mathbf{M}_\pi\mathbf{K}_Q^{-1}\| \cdot \|\mathbf{\Phi}_Q^\top\nu_D^\pi\| \\
&+ \|\left(\lambda_w\lambda_Q\mathbf{K}_w + \mathbf{M}_\pi\mathbf{K}_Q^{-1}\mathbf{M}_\pi^\top\right)^{-1}\| \cdot \|\lambda_Q\mathbf{\Phi}_w^\top\Lambda^D R\| \\
\leq &\frac{1}{\lambda_w\lambda_Q v_w + v_\mathbf{M}^2}(\frac{1-\gamma^2}{v_Q} + \lambda_Q) := R_\zeta \qquad (12)\\
\|\xi_\pi^*\| \leq &(1-\gamma)\lambda_w\|\left(\lambda_w\lambda_Q\mathbf{K}_Q + \mathbf{M}_\pi^\top\mathbf{K}_w^{-1}\mathbf{M}_\pi\right)^{-1}\| \cdot \|\mathbf{\Phi}_Q^\top\nu_D^\pi\| \\
&+ \|\left(\lambda_w\lambda_Q\mathbf{K}_Q + \mathbf{M}_\pi^\top\mathbf{K}_w^{-1}\mathbf{M}_\pi\right)^{-1}\mathbf{M}_\pi^\top\mathbf{K}_w^{-1}\|\|\mathbf{\Phi}_w^\top\Lambda^D R\| \\
\leq &\frac{1}{\lambda_w\lambda_Q v_Q + v_\mathbf{M}^2}((1-\gamma)\lambda_w + \frac{1+\gamma}{v_w}) := R_\xi \qquad (13)
\end{aligned}
$$

$\square$

Given the special property of Z_0 and Ξ_0, we force Z and Ξ satisfying the following condition:

Condition B.2. $Z_0 \subseteq Z, \Xi_0 \subseteq \Xi$.

As a direct result, we have:

$$\|\nabla_\zeta \max_{\zeta\in\mathbb{R}^u}\min_{\xi\in\mathbb{R}^u}\mathcal{L}^D(\theta,\zeta_\theta^*,\xi_\theta^*)\| = \|\nabla_\zeta\max_{\zeta\in Z}\min_{\xi\in\Xi}\mathcal{L}^D(\theta,\zeta_\theta^*,\xi_\theta^*)\| = 0,$$
$$\|\nabla_\xi \max_{\zeta\in\mathbb{R}^u}\min_{\xi\in\mathbb{R}^u}\mathcal{L}^D(\theta,\zeta_\theta^*,\xi_\theta^*)\| = \|\nabla_\xi\max_{\zeta\in Z}\min_{\xi\in\Xi}\mathcal{L}^D(\theta,\zeta_\theta^*,\xi_\theta^*)\| = 0.$$

Property B.3. *[Variance of Estimated Gradient] Under Assumption A, B and D, given convex sets Θ, Z, Ξ, where Z and Ξ have finite diameter $C_\mathcal{W}$ and $C_\mathcal{Q}$, then there exists constants $\sigma_\theta, \sigma_\zeta, \sigma_\xi$, such that, for arbitrary $\theta, \zeta, \xi \in \Theta \times Z \times \Xi$, we have:*

$$\mathbb{E}_{s,a,r,s',a',s_0,a_0}[\|\nabla_\theta\mathcal{L}^{(s,a,r,s',a',s_0,a_0)}(\theta,\zeta,\xi) - \nabla_\theta\mathcal{L}^D(\theta,\zeta,\xi)\|^2] \leq \sigma_\theta^2;$$
$$\mathbb{E}_{s,a,r,s',a',s_0,a_0}[\|\nabla_\zeta\mathcal{L}^{(s,a,r,s',a',s_0,a_0)}(\theta,\zeta,\xi) - \nabla_\zeta\mathcal{L}^D(\theta,\zeta,\xi)\|^2] \leq \sigma_\zeta^2;$$
$$\mathbb{E}_{s,a,r,s',a',s_0,a_0}[\|\nabla_\xi\mathcal{L}^{(s,a,r,s',a',s_0,a_0)}(\theta,\zeta,\xi) - \nabla_\xi\mathcal{L}^D(\theta,\zeta,\xi)\|^2] \leq \sigma_\xi^2.$$

Here we use $\mathbb{E}_{s,a,r,s',a',s_0,a_0}[\cdot]$ as a shorthand of $\mathbb{E}_{(s,a,r,s')\sim d^D, a'\sim\pi(\cdot|s'), s_0\sim\nu_0^D, a_0\sim\pi(\cdot|s_0)}[\cdot]$, and use $\nabla\mathcal{L}^{(s,a,r,s',a',s_0,a_0)}(\theta,\zeta,\xi)$ to denote the stochastic gradient estimated using a single data point:

$$\nabla_\theta\mathcal{L}^{(s,a,r,s',a',s_0,a_0)}(\theta,\zeta,\xi) = (1-\gamma)Q_\xi(s_0,a_0)\nabla_\theta\log\pi_\theta(a_0|s) + \gamma w_\zeta(s,a)Q_\xi(s',a')\nabla_\theta\log\pi_\theta(a'|s'),$$
$$\nabla_\zeta\mathcal{L}^{(s,a,r,s',a',s_0,a_0)}(\theta,\zeta,\xi) = \left(r+\gamma Q_\xi(s',a') - Q_\xi(s,a)\right)\nabla_\zeta w_\zeta(s,a) - \lambda_w w_\zeta(s,a)\nabla_\zeta w_\zeta(s,a),$$
$$\nabla_\xi\mathcal{L}^{(s,a,r,s',a',s_0,a_0)}(\theta,\zeta,\xi) = (1-\gamma)\nabla_\xi Q_\xi(s_0,a_0) + w_\zeta(s,a)\nabla_\xi\left(\gamma Q_\xi(s',a') - Q_\xi(s,a)\right) + \lambda_Q Q_\xi(s,a)\nabla_\xi Q_\xi(s,a).$$

$$(14)$$

16

Proof. Under Linear case and Assumption D, we should have

$$\mathbb{E}_{s,a,r,s',a',s_0,a_0}[\|\nabla_\theta \mathcal{L}^{(s,a,r,s',a',s_0,a_0)}(\theta,\zeta,\xi) - \nabla_\theta \mathcal{L}^D(\theta,\zeta,\xi)\|^2]$$

$$\leq 2(1-\gamma)^2 \mathbb{E}[\|Q_\xi(s_0,a_0)\nabla_\theta \log \pi_\theta(a_0|s_0) - \mathbb{E}[Q_\xi(s_0,a_0)\nabla_\theta \log \pi_\theta(a_0|s_0)]\|^2]$$

$$+ 2\gamma^2 \mathbb{E}[\|w_\zeta(s,a)Q_\xi(s',a')\nabla_\theta \log \pi_\theta(a'|s') - \mathbb{E}[w_\zeta(s,a)Q_\xi(s',a')\nabla_\theta \log \pi_\theta(a'|s')]\|^2]$$

$$\leq 2(1-\gamma)^2 \mathbb{E}[\|Q_\xi(s_0,a_0)\nabla_\theta \log \pi_\theta(a_0|s_0)\|^2] + 2\gamma^2 \mathbb{E}[\|w_\zeta(s,a)Q_\xi(s',a')\nabla_\theta \log \pi_\theta(a'|s')\|^2]$$

$$\leq 2(1-\gamma)^2 C_\mathcal{Q}^2 G^2 + 2\gamma^2 C_\mathcal{W}^2 C_\mathcal{Q}^2 G^2$$

$$\mathbb{E}_{s,a,r,s',a',s_0,a_0}[\|\nabla_\zeta \mathcal{L}^{(s,a,r,s',a',s_0,a_0)}(\theta,\zeta,\xi) - \nabla_\zeta \mathcal{L}^D(\theta,\zeta,\xi)\|^2]$$

$$\leq 3\mathbb{E}[\|\phi_w(s,a)(\gamma\phi_Q(s',a') - \phi_Q(s,a))^\top \xi - \mathbb{E}[\phi_w(s,a)(\gamma\phi_Q(s',a') - \phi_Q(s,a))^\top \xi]\|^2]$$

$$+ 3\mathbb{E}[\|\phi_w(s,a)r - \mathbb{E}[\phi_w(s,a)r]\|^2] + 3\lambda_w^2 \mathbb{E}[\|\phi_w(s,a)\phi_w(s,a)^\top \zeta - \mathbb{E}[\phi_w(s,a)\phi_w(s,a)^\top \zeta]\|^2]$$

$$\leq 3\sigma_R^2 + 3\sigma_\mathbf{M}^2 C_\mathcal{Q}^2 + 3\lambda_w^2 \sigma_\mathbf{K}^2 C_\mathcal{W}^2$$

$$\mathbb{E}_{s,a,r,s',a',s_0,a_0}[\|\nabla_\xi \mathcal{L}^{(s,a,r,s',a',s_0,a_0)}(\theta,\zeta,\xi) - \nabla_\xi \mathcal{L}^D(\theta,\zeta,\xi)\|^2]$$

$$\leq 3(1-\gamma)^2 \mathbb{E}[\|\phi_Q(s_0,a_0) - \mathbb{E}[\phi_Q(s_0,a_0)]\|^2] + 3\lambda_Q^2 \mathbb{E}[\|\phi_Q(s,a)\phi_Q(s,a)^\top \xi - \mathbb{E}[\phi_Q(s,a)\phi_Q(s,a)^\top \xi]\|^2]$$

$$+ 3\mathbb{E}[\|\zeta\phi_w(s,a)(\gamma\phi_Q(s',a') - \phi_Q(s,a))^\top - \mathbb{E}[\zeta\phi_w(s,a)(\gamma\phi_Q(s',a') - \phi_Q(s,a))^\top]\|^2]$$

$$\leq 3(1-\gamma)^2 \sigma_\nu^2 + 3\sigma_\mathbf{M}^2 C_\mathcal{W}^2 + 3\lambda_Q^2 \sigma_\mathbf{K}^2 C_\mathcal{Q}^2$$

which finishes the proof. $\qquad\square$

In Linear setting, $\sigma_\theta^2, \sigma_\zeta^2, \sigma_\xi^2$ can be chosen as:

$$\sigma_\theta^2 = 2(1-\gamma)^2 \sigma_\nu^2 G^2 C_\mathcal{Q}^2 + 2\gamma^2 \sigma_\mathbf{M}^2 G^2 C_\mathcal{W}^2 C_\mathcal{Q}^2,$$

$$\sigma_\zeta^2 = 3\sigma_R^2 + 3\sigma_\mathbf{M}^2 C_\mathcal{Q}^2 + 3\lambda_w^2 \sigma_\mathbf{K}^2 C_\mathcal{W}^2,$$

$$\sigma_\xi^2 = 3(1-\gamma)^2 \sigma_\nu^2 + 3\sigma_\mathbf{M}^2 C_\mathcal{W}^2 + 3\lambda_Q^2 \sigma_\mathbf{K}^2 C_\mathcal{Q}^2.$$

In the following, we will use σ to refer to the $\max\{\sigma_\theta, \sigma_\zeta, \sigma_\xi\}$ value satisfying Property B.3.

Finally, we prove a condition which is useful in the analysis of our second strategy. We first introduce some new notations. Suppose we have a mini batch data N sampled according to d^D whose batch size is constant $|N|$. Then, we denote the average batch gradients as

$$\nabla \mathcal{L}^N(\theta,\zeta,\xi) = \frac{1}{|N|} \sum_{i=1}^{|N|} \nabla \mathcal{L}^{(s^i,a^i,r^i,s'^i,a'^i,s_0^i,a_0^i)}(\theta,\zeta,\xi)$$

where $\nabla \mathcal{L}^{(s^i,a^i,r^i,s'^i,a'^i,s_0^i,a_0^i)}(\theta,\zeta,\xi)$ is defined in Eq.(14).

Property B.4. *Under Assumption A and B, there exists two constants two constants $\bar{L}_\zeta$ and $\bar{L}_\xi$, such that:*

$$\mathbb{E}_{N\sim d^D}[\|\nabla_\zeta \mathcal{L}^N(\theta,\zeta_1,\xi_1) - \nabla_\zeta \mathcal{L}^N(\theta,\zeta_2,\xi_2)\|^2 + \|\nabla_\xi \mathcal{L}^N(\theta,\zeta_1,\xi_1) - \nabla_\xi \mathcal{L}^N(\theta,\zeta_2,\xi_2)\|^2]$$

$$\leq \mathbb{E}_{N\sim d^D}\left[\bar{L}_\zeta \left(\nabla_\zeta \mathcal{L}^N(\theta,\zeta_1,\xi_1) - \nabla_\zeta \mathcal{L}^N(\theta,\zeta_2,\xi_2)\right)^\top (\zeta_2 - \zeta_1) + \bar{L}_\xi \left(\nabla_\xi \mathcal{L}^N(\theta,\zeta_1,\xi_1) - \nabla_\xi \mathcal{L}^N(\theta,\zeta_2,\xi_2)\right)^\top (\xi_1 - \xi_2)\right],$$

$$\mathbb{E}_{N\sim d^D}[\|\nabla_\zeta \mathcal{L}^N(\theta,\zeta_1,\xi_1) - \nabla_\zeta \mathcal{L}^N(\theta,\zeta_2,\xi_2)\|^2 + \|\nabla_\xi \mathcal{L}^N(\theta,\zeta_1,\xi_1) - \nabla_\xi \mathcal{L}^N(\theta,\zeta_2,\xi_2)\|^2]$$

$$\leq \bar{L}_\zeta^2 \|\zeta_1 - \zeta_2\|^2 + \bar{L}_\xi^2 \|\xi_1 - \xi_2\|^2.$$

Proof. For simplicity, we use $\mathbf{K}_w^N$ to denote matrix $\mathbb{E}_N[\phi_w(s,a)\phi_w(s,a)^\top]$ ($\mathbf{K}_Q^N$ is similar) and use $\mathbf{M}_\pi^N$ to denote $\mathbb{E}_N[\phi(s,a)\phi(s,a)^\top - \gamma\phi(s,a)\phi(s',\pi)^\top]$

$$\nabla_\zeta \mathcal{L}^N(\theta,\zeta_1,\xi_1) - \nabla_\zeta \mathcal{L}^N(\theta,\zeta_2,\xi_2) = -\lambda_w \mathbf{K}_w^N(\zeta_1 - \zeta_2) - \mathbf{M}_\pi^N(\xi_1 - \xi_2)$$

$$\nabla_\xi \mathcal{L}^N(\theta,\zeta_1,\xi_1) - \nabla_\xi \mathcal{L}^N(\theta,\zeta_2,\xi_2) = \lambda_Q \mathbf{K}_Q^N(\xi_1 - \xi_2) - \mathbf{M}_\pi^{N\top}(\zeta_1 - \zeta_2)$$

17

Therefore,

$$\mathbb{E}_{N\sim d^D}[\|\nabla_\zeta \mathcal{L}^N(\theta,\zeta_1,\xi_1) - \nabla_\zeta \mathcal{L}^N(\theta,\zeta_2,\xi_2)\|^2 + \|\nabla_\xi \mathcal{L}^N(\theta,\zeta_1,\xi_2) - \nabla_\xi \mathcal{L}^N(\theta,\zeta_2,\xi_2)\|^2]$$

$$\leq 2\mathbb{E}_{N\sim d^D}[(\zeta_1 - \zeta_2)^\top(\lambda_w^2(\mathbf{K}_w^N)^\top \mathbf{K}_w^N + {\mathbf{M}_\pi^N}^\top \mathbf{M}_\pi^N)(\zeta_1 - \zeta_2)]$$

$$\quad + 2\mathbb{E}_{N\sim d^D}[(\xi_1 - \xi_2)^\top(\lambda_Q^2(\mathbf{K}_Q^N)^\top \mathbf{K}_Q^N + {\mathbf{M}_\pi^N}^\top \mathbf{M}_\pi^N)(\xi_1 - \xi_2)]$$

$$\leq 2\mathbb{E}_{N\sim d^D}[(\zeta_1 - \zeta_2)^\top(\lambda_w^2(\mathbf{K}_w^N)^2 + (1+\gamma)^2 I)(\zeta_1 - \zeta_2)] + 2\mathbb{E}_{N\sim d^D}[(\xi_1 - \xi_2)^\top(\lambda_Q^2(\mathbf{K}_Q^N)^2 + (1+\gamma)^2 I)(\xi_1 - \xi_2)]$$

$$\leq 2\mathbb{E}_{N\sim d^D}[(\zeta_1 - \zeta_2)^\top(\lambda_w^2(\mathbf{K}_w^N)^2 + (1+\gamma)^2 I)(\zeta_1 - \zeta_2)] + 2\mathbb{E}_{N\sim d^D}[(\xi_1 - \xi_2)^\top(\lambda_Q^2(\mathbf{K}_Q^N)^2 + (1+\gamma)^2 I)(\xi_1 - \xi_2)]$$

$$= (\zeta_1 - \zeta_2)^\top(2\lambda_w^2 \mathbf{K}_w + 2(1+\gamma)^2 I)(\zeta_1 - \zeta_2) + (\xi_1 - \xi_2)^\top(2\lambda_Q^2 \mathbf{K}_Q + 2(1+\gamma)^2 I)(\xi_1 - \xi_2)$$

In the first inequality, we use Young's inequality; in the second one, we use the fact that the largest singular value of $\mathbf{M}_\pi^N$ is less than $(1+\gamma)$; the third one is because all eigenvalues of $\mathbf{K}_w^N$ and $\mathbf{K}_Q^N$ locate in $[0,1]$, and we should have $I \succ \mathbf{K}_w^N \succ (\mathbf{K}_w^N)^2$ and $I \succ \mathbf{K}_Q^N \succ (\mathbf{K}_Q^N)^2$. Notice that,

$$\mathbb{E}_{N\sim d^D}\Big[- \Big(\nabla_\zeta \mathcal{L}^N(\theta,\zeta_1,\xi_1) - \nabla_\zeta \mathcal{L}^N(\theta,\zeta_2,\xi_2)\Big)^\top (\zeta_1 - \zeta_2)$$

$$+ \Big(\nabla_\xi \mathcal{L}^N(\theta,\zeta_1,\xi_1) - \nabla_\xi \mathcal{L}^N(\theta,\zeta_2,\xi_2)\Big)^\top (\xi_1 - \xi_2)\Big]$$

$$= \lambda_w(\zeta_1 - \zeta_2)^\top \mathbf{K}_w(\zeta_1 - \zeta_2) + \lambda_Q(\xi_1 - \xi_2)^\top \mathbf{K}_Q(\xi_1 - \xi_2)$$

Therefore,

$$(\zeta_1 - \zeta_2)^\top(2\lambda_w^2 \mathbf{K}_w + 2(1+\gamma)^2 I)(\zeta_1 - \zeta_2) + (\xi_1 - \xi_2)^\top(2\lambda_Q^2 \mathbf{K}_Q + 2(1+\gamma)^2 I)(\xi_1 - \xi_2)$$

$$\leq \frac{2\max\{\lambda_w^2, \lambda_Q^2\} + 2(1+\gamma)^2}{\min\{\lambda_w v_w, \lambda_Q v_Q\}}\Big(\lambda_w(\zeta_1 - \zeta_2)^\top \mathbf{K}_w(\zeta_1 - \zeta_2) + \lambda_Q(\xi_1 - \xi_2)^\top \mathbf{K}_Q(\xi_1 - \xi_2)\Big)$$

Moreover,

$$(\zeta_1 - \zeta_2)^\top(2\lambda_w^2 \mathbf{K}_w + 2(1+\gamma)^2 I)(\zeta_1 - \zeta_2) + (\xi_1 - \xi_2)^\top(2\lambda_Q^2 \mathbf{K}_Q + 2(1+\gamma)^2 I)(\xi_1 - \xi_2)$$

$$\leq (2\max\{\lambda_w^2, \lambda_Q^2\} + 2(1+\gamma)^2)\Big((\zeta_1 - \zeta_2)^\top(\zeta_1 - \zeta_2) + (\xi_1 - \xi_2)^\top(\xi_1 - \xi_2)\Big)$$

As a result, Assumption B.4 holds with

$$\bar{L}_\zeta = \bar{L}_\xi = \max\{\frac{2\max\{\lambda_w^2, \lambda_Q^2\} + 2(1+\gamma)^2}{\min\{\lambda_w v_w, \lambda_Q v_Q\}}, \sqrt{2\max\{\lambda_w^2, \lambda_Q^2\} + 2(1+\gamma)^2}\}$$

$\square$

C   The Analysis of Bias

We first prove two propositions, which are crucial for analyzing the biases due to the finite dataset and mis-specified function classes.

Proposition C.1. *For abitrary $\pi \in \Theta$, we have:*

$$\|\nabla_\theta \max_{w\in\mathcal{W}} \min_{Q\in\mathcal{Q}} \mathcal{L}(\pi_\theta, w, Q) - \nabla_\theta \max_{w\in\mathcal{W}} \min_{Q\in\mathcal{Q}} \mathcal{L}^D(\pi_\theta, w, Q)\| \leq \underbrace{(2\kappa_\zeta \kappa_\xi + 2\kappa_\zeta + 2\kappa_\xi + \sqrt{2}/2)\sqrt{2\bar{\varepsilon}_{data}}}_{denoted\ as\ \varepsilon_{data}}$$

where $\bar{\varepsilon}_{data}$ is defined in Definition 3.1.

Proof. For the simplicity of notation, we give the proof for a fixed π.

Denote (w_μ^*, Q_μ^*) parameterized by (ζ_μ^*, ξ_μ^*) as $\arg\max_{w\in\mathcal{W}} \min_{Q\in\mathcal{Q}} \mathcal{L}(\pi, w, Q)$ and denote (w^*, Q^*) parameterized by (ζ^*, ξ^*) as $\arg\max_{w\in\mathcal{W}} \min_{Q\in\mathcal{Q}} \mathcal{L}^D(\pi, w, Q)$. First, we try to bound $\zeta^* - \zeta_\mu^*$. We use Q_w and

Q_w^D (parameterized by ξ_w and ξ_w^D) as the short notes of $\arg\min_{Q\in\mathcal{Q}}\mathcal{L}(\pi,w,Q)$ and $\arg\min_{Q\in\mathcal{Q}}\mathcal{L}^D(\pi,w,Q)$, respectively. Then,

$$|\mathcal{L}(\pi,w,Q_w) - \mathcal{L}^D(\pi,w,Q_w^D)| \leq \max\{\mathcal{L}(\pi,w,Q_w^D) - \mathcal{L}^D(\pi,w,Q_w^D), \mathcal{L}^D(\pi,w,Q_w) - \mathcal{L}(\pi,w,Q_w)\} \leq \bar{\varepsilon}_{data}$$

As a result,

$$
\begin{aligned}
0 \leq &\mathcal{L}^D(\pi,w^*,Q^*) - \min_{Q\in\mathcal{Q}}\mathcal{L}^D(\pi,w_\mu^*,Q)\\
\leq &\mathcal{L}^D(\pi,w^*,Q^*) - \min_{Q\in\mathcal{Q}}\mathcal{L}(\pi,w^*,Q) + \mathcal{L}(\pi,w_\mu^*,Q_\mu^*) - \min_{Q\in\mathcal{Q}}\mathcal{L}^D(\pi,w_\mu^*,Q)\\
\leq &2\bar{\varepsilon}_{data}
\end{aligned}
$$

According to Lemma A.1, $\min_{Q\in\mathcal{Q}}\mathcal{L}^D(\pi,w,Q)$ is μ_ζ-strongly concave. Therefore,

$$\|\zeta^* - \zeta_\mu^*\| \leq \frac{2}{\mu_\zeta}\sqrt{\mathcal{L}^D(\pi,w^*,Q^*) - \min_{Q\in\mathcal{Q}}\mathcal{L}^D(\pi,w_\mu^*,Q)} \leq \frac{2}{\mu_\zeta}\sqrt{2\bar{\varepsilon}_{data}}$$

Next, we bound $\|\xi^* - \xi_\mu^*\|$. For arbitrary $\pi\in\Pi$ and $w\in\mathcal{W}$, we have:

$$0 \leq \mathcal{L}^D(\pi,w,Q_w) - \mathcal{L}^D(\pi,w,Q_w^D) \leq \mathcal{L}^D(\pi,w,Q_w) - \mathcal{L}(\pi,w,Q_w) + \mathcal{L}(\pi,w,Q_w^D) - \mathcal{L}^D(\pi,w,Q_w^D) \leq 2\bar{\varepsilon}_{data}$$

Since L^D is μ_ξ strongly-convex, as a result of Lemma A.3, for arbitrary w,

$$\|\xi_w - \xi_w^D\| \leq \frac{2}{\mu_\xi}\sqrt{2\bar{\varepsilon}_{data}} \tag{15}$$

Then, we have

$$
\begin{aligned}
\|\xi^* - \xi_\mu^*\| \leq &\|\xi^* - \arg\min_{\xi\in\Xi}\mathcal{L}^D(\pi,w_\mu^*,Q_\xi)\| + \|\arg\min_{\xi\in\Xi}\mathcal{L}^D(\pi,w_\mu^*,Q_\xi) - \xi_\mu^*\|\\
= &\|\xi^* - \arg\min_{\xi\in\Xi}\mathcal{L}^D(\pi,w_\mu^*,Q_\xi)\| + \|\arg\min_{\xi\in\Xi}\mathcal{L}^D(\pi,w_\mu^*,Q_\xi) - \arg\min_{\xi\in\Xi}\mathcal{L}(\pi,w_\mu^*,Q_\xi)\|\\
\leq &\frac{L}{\mu_\xi}\|\zeta^* - \zeta_\mu^*\| + \frac{2}{\mu_\xi}\sqrt{2\bar{\varepsilon}_{data}}\\
\leq &(\frac{2L}{\mu_\xi\mu_\zeta} + \frac{2}{\mu_\xi})\sqrt{2\bar{\varepsilon}_{data}}
\end{aligned}
$$

where in the last but two step, we use Lemma A.1-(1).

As a directly application of Property B.1, we have:

$$
\begin{aligned}
&\|\nabla_\theta\max_{w\in\mathcal{W}}\min_{Q\in\mathcal{Q}}\mathcal{L}(\pi_\theta,w,Q) - \nabla_\theta\max_{w\in\mathcal{W}}\min_{Q\in\mathcal{Q}}\mathcal{L}^D(\pi_\theta,w,Q)\|\\
= &\|\nabla_\theta\mathcal{L}(\pi_\theta,w_\mu^*,Q_\mu^*) - \nabla_\theta\mathcal{L}^D(\pi_\theta,w_\mu^*,Q_\mu^*)\| + \|\nabla_\theta\mathcal{L}^D(\pi_\theta,w_\mu^*,Q_\mu^*) - \nabla_\theta\mathcal{L}^D(\pi_\theta,w^*,Q^*)\|\\
\leq &\sqrt{\bar{\varepsilon}_{data}} + L\|\zeta^* - \zeta_\mu^*\| + L\|\xi^* - \xi_\mu^*\|\\
\leq &(2\kappa_\zeta\kappa_\xi + 2\kappa_\zeta + 2\kappa_\xi + \sqrt{2}/2)\sqrt{2\bar{\varepsilon}_{data}}
\end{aligned}
$$

$\square$

Proposition C.2. *For arbitrary $\pi\in\Pi$, we have:*

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

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

where (w_μ^*,Q_μ^*) denotes the saddle point of $\mathcal{L}(\pi,w,Q)$ constrained by $w,Q\in\mathcal{W}\times\mathcal{Q}$ (i.e. $\zeta\in Z, \xi\in\Xi$), $(w_\mathcal{L}^\pi,Q_\mathcal{L}^\pi)$ denotes the saddle point of $\mathcal{L}(\pi,w,Q)$ without any constraint on w and Q (i.e. w and Q can be arbitrary vectors in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$), $\lambda_{\max} = \max\{\lambda_Q,\lambda_w\}$, μ_ζ is defined in Property 2.2.

Proof. In the following, we will frequently consider two loss functions. The first one is $\mathcal{L}(\pi, w, Q)$ defined in Eq.(6), 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:

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

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. $x, y \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

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,

$$\frac{\lambda_w}{2} \| \Lambda^{1/2} w_{\mathbb{R}}^* - \Lambda^{1/2} w_{\mathcal{L}}^\pi \|^2$$

$$\leq \mathcal{F}(\pi, \Lambda^{1/2} w_{\mathcal{L}}^\pi, \Lambda^{1/2} Q_{\mathcal{L}}^\pi) - \min_y \mathcal{F}(\pi, \Lambda^{1/2} w_{\mathbb{R}}^*, y) \qquad \text{(Strong concavity of } \min_y \mathcal{F}(\pi, x, y))$$

$$= \mathcal{F}(\pi, \Lambda^{1/2} w_{\mathcal{L}}^\pi, \Lambda^{1/2} Q_{\mathcal{L}}^\pi) - \max_{w \in \mathcal{W}} \min_y \mathcal{F}(\pi, \Lambda^{1/2} w, y)$$

$$\leq \mathcal{F}(\pi, \Lambda^{1/2} w_{\mathcal{L}}^\pi, \Lambda^{1/2} Q_{\mathcal{L}}^\pi) - \min_y \mathcal{F}(\pi, \Lambda^{1/2} w_{\zeta^\pi}, y) \qquad (w_{\zeta^\pi} \text{ is defined in Def. 3.2})$$

$$\leq \frac{\lambda_{\max}^2}{\lambda_Q} \| \Lambda^{1/2} w_{\zeta^\pi} - \Lambda^{1/2} w_{\mathcal{L}}^\pi \|^2 \qquad \text{(Smoothness of } \min_y \mathcal{F}(\pi, x, y))$$

$$= \frac{\lambda_{\max}^2}{\lambda_Q} \| w_{\zeta^\pi} - w_{\mathcal{L}}^\pi \|_\Lambda^2 = \frac{\lambda_{\max}^2}{\lambda_Q} \varepsilon_1 \qquad \text{(see definition of } \varepsilon_1 \text{ in Def.3.2)}$$

which implies

$$\| \Lambda^{1/2} w_{\mathbb{R}}^* - \Lambda^{1/2} w_{\mathcal{L}}^\pi \|^2 \leq 2 \frac{\lambda_{\max}^2}{\lambda_Q \lambda_w} \varepsilon_1 \qquad (16)$$

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

$$\frac{\mu_\zeta}{2} \| \zeta^* - \zeta_{\mathbb{R}}^* \|^2 \leq \mathcal{L}(\pi, w_\mu^*, Q_\mu^*) - \min_{Q \in \mathcal{Q}} \mathcal{L}(\pi, w_{\mathbb{R}}^*, Q) \qquad \text{(Strong concavity of } \min_{Q \in \mathcal{Q}} \mathcal{L}(\pi, w, Q); \text{ Lemma A.3)}$$

$$= \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, \Lambda^{1/2} Q_\mu^*) - \min_{Q \in \mathcal{Q}} \mathcal{F}(\pi, \Lambda^{1/2} w_{\mathbb{R}}^*, \Lambda^{1/2} Q)$$

$$\leq \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, \Lambda^{1/2} Q_\mu^*) - \min_y \mathcal{F}(\pi, \Lambda^{1/2} w_{\mathbb{R}}^*, y)$$

$$\leq \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, \Lambda^{1/2} Q_\mu^*) - \min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, y)$$

$$\qquad \text{(Because } w_{\mathbb{R}}^* = \arg\max_{w \in \mathcal{W}} \min_y \mathcal{F}(\pi, \Lambda^{1/2} w, y))$$

$$\leq \frac{\lambda_{\max}}{2} \| \Lambda^{1/2} Q_\mu^* - \arg\min_y \mathcal{F}(\pi, \Lambda^{1/2} w_\mu^*, y) \|^2$$

$$\qquad \text{(Smoothness of } \mathcal{F}(\pi, x, y) \text{ for fixed } x \text{ and } \nabla_y \min_y \mathcal{F} = 0)$$

$$\leq \frac{\lambda_{\max}}{2} \varepsilon_2$$

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 $\nabla_y \min_y \mathcal{F}(\pi, \Lambda^{1/2}w_\mu^*, Q) = 0$; in the last equality, we use the definition of ε_2 in Def. 3.2. Combing (b) in Property B.1 with $L_w = 1$, for arbitrary $s, a \in \mathcal{S} \times \mathcal{A}$, we have:

$$|w_\mu^*(s,a) - w_{\mathbb{R}}^*(s,a)|^2 \le \|\zeta^* - \zeta_{\mathbb{R}}^*\|^2 \le \frac{\lambda_{\max}}{\mu_\zeta}\varepsilon_2 \tag{17}$$

Therefore, as a result of Eq.(16) and Eq.(17):

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

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

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

As a result,

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

$\square$

Theorem C.3 (Bias resulting from regularization). *Let's rewrite Eq.(6) in a vector-matrix form:*

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

where ν_0^π and $\mathbf{P}^\pi$ denotes the initial state-action distribution and the transition matrix w.r.t. policy π, respectively; $\Lambda \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ denotes the diagonal matrix whose diagonal elements are $d^\mu(\cdot, \cdot)$. 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 (i.e. $\mathcal{W} = \mathcal{Q} = \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$), then we have:

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

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

Proof. Recall the loss function

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

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

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

21

Plug this result in, and we have

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

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

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

As a result,

$$w_{\mathcal{L}}^\pi = \Big(\lambda_w I + \frac{1}{\lambda_Q}(I - \gamma\mathbf{P}^\pi)\Lambda^{-1}(I - \gamma\mathbf{P}_*^\pi)\Lambda\Big)^{-1}\Big(\frac{1}{\lambda_Q}(I - \gamma\mathbf{P}^\pi)\Lambda^{-1}(1-\gamma)\nu_0^\pi + R\Big)$$

$$= \Big(\lambda_w\lambda_Q I + (I - \gamma\mathbf{P}^\pi)\Lambda^{-1}(I - \gamma\mathbf{P}_*^\pi)\Lambda\Big)^{-1}\Big((I - \gamma\mathbf{P}^\pi)\Lambda^{-1}(I - \gamma\mathbf{P}_*^\pi)\Lambda\Lambda^{-1}(I - \gamma\mathbf{P}_*^\pi)^{-1}(1-\gamma)\nu_0^\pi + \lambda_Q R\Big)$$

$$= w^\pi + \Big(\lambda_w\lambda_Q I + (I - \gamma\mathbf{P}^\pi)\Lambda^{-1}(I - \gamma\mathbf{P}_*^\pi)\Lambda\Big)^{-1}\Big(\lambda_Q R - \lambda_Q\lambda_w w^\pi\Big)$$

and

$$Q_{\mathcal{L}}^\pi = \frac{1}{\lambda_Q}\Lambda^{-1}\Big((I - \gamma\mathbf{P}_*^\pi)\Lambda w_{\mathcal{L}}^\pi - (1-\gamma)\nu_0^\pi\Big)$$

$$= \frac{1}{\lambda_Q}\Lambda^{-1}\Big((I - \gamma\mathbf{P}_*^\pi)\Lambda w_{\mathcal{L}}^\pi - (I - \gamma\mathbf{P}_*^\pi)\Lambda w^\pi\Big)$$

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

$$= \Big(\lambda_w\lambda_Q(I - \gamma\mathbf{P}_*^\pi)^{-1}\Lambda + \Lambda(I - \gamma\mathbf{P}^\pi)\Big)^{-1}\Big(\Lambda R - \lambda_w\Lambda w^\pi\Big)$$

$$= \Big(\lambda_w\lambda_Q(I - \gamma\mathbf{P}_*^\pi)^{-1}\Lambda + \Lambda(I - \gamma\mathbf{P}^\pi)\Big)^{-1}\Big(\Lambda(I - \gamma\mathbf{P}^\pi)Q^\pi - \lambda_w\Lambda w^\pi\Big)$$

$$= Q^\pi - \Big(\lambda_w\lambda_Q(I - \gamma\mathbf{P}_*^\pi)^{-1}\Lambda + \Lambda(I - \gamma\mathbf{P}^\pi)\Big)^{-1}\Big(\lambda_w\lambda_Q(I - \gamma\mathbf{P}_*^\pi)^{-1}\Lambda Q^\pi + \lambda_w\Lambda w^\pi\Big)$$

$$= Q^\pi - \Big(\lambda_w\lambda_Q I + \Lambda^{-1}(I - \gamma\mathbf{P}_*^\pi)\Lambda(I - \gamma\mathbf{P}^\pi)\Big)^{-1}\Big(\lambda_w\lambda_Q Q^\pi + \lambda_w(1-\gamma)\Lambda^{-1}\nu_0^\pi\Big)$$

$\square$

Lemma C.4. *Under Assumption C:*

$$\|w^\pi - w_{\mathcal{L}}^\pi\|_\Lambda^2 \le \frac{C^2(\lambda_Q + \lambda_Q\lambda_w C)^2}{(1-\gamma)^4}, \quad \|Q^\pi - Q_{\mathcal{L}}^\pi\|_\Lambda^2 \le \frac{C^2}{(1-\gamma)^2}\Big(\frac{\lambda_w\lambda_Q}{1-\gamma} + \lambda_w\Big)^2$$

where (w^π, Q^π) and $(w_{\mathcal{L}}^\pi, Q_{\mathcal{L}}^\pi)$ are defined in Theorem C.3. $\|x\|_\Lambda = x^\top\Lambda x$ denotes the norm of column vector x weighted by Λ.

Proof. From Theorem C.3, we have

$$w_{\mathcal{L}}^\pi = w^\pi + \Big(\lambda_w\lambda_Q I + (I - \gamma\mathbf{P}^\pi)\Lambda^{-1}(I - \gamma\mathbf{P}_*^\pi)\Lambda\Big)^{-1}\Big(\lambda_Q R - \lambda_Q\lambda_w w^\pi\Big)$$

$$Q_{\mathcal{L}}^\pi = Q^\pi - \Big(\lambda_w\lambda_Q I + \Lambda^{-1}(I - \gamma\mathbf{P}_*^\pi)\Lambda(I - \gamma\mathbf{P}^\pi)\Big)^{-1}\Big(\lambda_w\lambda_Q Q^\pi + \lambda_w(1-\gamma)\Lambda^{-1}\nu_0^\pi\Big)$$

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

$$\|w^\pi - w_{\mathcal{L}}^\pi\|_\Lambda^2 = \|\Big(\lambda_w\lambda_Q I + (I - \gamma\mathbf{P}^\pi)\Lambda^{-1}(I - \gamma\mathbf{P}_*^\pi)\Lambda\Big)^{-1}\Big(\lambda_Q R - \lambda_Q\lambda_w w^\pi\Big)\|_\Lambda^2$$

22

$$=\|\left(\lambda_w\lambda_Q I+\Lambda^{1/2}(I-\gamma\mathbf{P}^\pi)\Lambda^{-1}(I-\gamma\mathbf{P}^\pi_*)\Lambda^{1/2}\right)^{-1}\Lambda^{1/2}\left(\lambda_Q R-\lambda_Q\lambda_w w^\pi\right)\|^2$$

$$\leq\|\Lambda^{-1/2}(I-\gamma\mathbf{P}^\pi_*)^{-1}\Lambda(I-\gamma\mathbf{P}^\pi)^{-1}\left(\lambda_Q R-\lambda_Q\lambda_w w^\pi\right)\|^2$$

$$=\|\Lambda^{-1/2}(I-\gamma\mathbf{P}^\pi_*)^{-1}\Lambda\widetilde{Q}^\pi\|^2$$

$$\leq\frac{(\lambda_Q+\lambda_Q\lambda_w C)^2}{(1-\gamma)^2}\|\Lambda^{-1}(I-\gamma\mathbf{P}^\pi_*)^{-1}\Lambda\mathbf{1}\|^2_\Lambda$$

$$=\frac{(\lambda_Q+\lambda_Q\lambda_w C)^2}{(1-\gamma)^2}\|\Lambda^{-1}(I-\gamma\mathbf{P}^\pi_*)^{-1}d^\mu\|^2_\Lambda$$

$$=\frac{(\lambda_Q+\lambda_Q\lambda_w C)^2}{(1-\gamma)^4}\|w^\pi_{d^\mu}\|^2_\Lambda\leq\frac{C^2(\lambda_Q+\lambda_Q\lambda_w C)^2}{(1-\gamma)^4}$$

where in the first inequality, we use Lemma A.4; in the third 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 second 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 C; in the last inequality, we use Assumption C again. Similarly,

$$\|Q^\pi-Q^\pi_{\mathcal{L}}\|^2_\Lambda\leq\|\left(\lambda_w\lambda_Q I+\Lambda^{-1}(I-\gamma\mathbf{P}^\pi_*)\Lambda(I-\gamma\mathbf{P}^\pi)\right)^{-1}\left(\lambda_w\lambda_Q Q^\pi+\lambda_w(1-\gamma)\Lambda^{-1}\nu^\pi_0)\right)\|^2_\Lambda$$

$$=\|\left(\lambda_Q\lambda_w I+\Lambda^{-1/2}(I-\gamma\mathbf{P}^\pi_*)\Lambda(I-\gamma\mathbf{P}^\pi)\Lambda^{-1/2}\right)^{-1}\Lambda^{1/2}\left(\lambda_Q\lambda_w Q^\pi+\lambda_w(1-\gamma)\Lambda^{-1}\nu^\pi_0)\right)\|^2$$

$$\leq\|\Lambda^{1/2}(I-\gamma\mathbf{P}^\pi)^{-1}\Lambda^{-1}(I-\gamma\mathbf{P}^\pi_*)^{-1}\left(\lambda_w\lambda_Q\Lambda Q^\pi+\lambda_w(1-\gamma)\nu^\pi_0)\right)\|^2$$

$$=\|\lambda_w\lambda_Q\Lambda^{1/2}(I-\gamma\mathbf{P}^\pi)^{-1}\Lambda^{-1}(I-\gamma\mathbf{P}^\pi_*)^{-1}\Lambda Q^\pi+\lambda_w\Lambda^{1/2}(I-\gamma\mathbf{P}^\pi)^{-1}w^\pi\|^2$$

$$\leq\|\frac{\lambda_w\lambda_Q}{1-\gamma}\Lambda^{1/2}(I-\gamma\mathbf{P}^\pi)^{-1}\Lambda^{-1}(I-\gamma\mathbf{P}^\pi_*)^{-1}\Lambda\mathbf{1}+\lambda_w\Lambda^{1/2}(I-\gamma\mathbf{P}^\pi)^{-1}w^\pi\|^2$$

$$\leq\|(I-\gamma\mathbf{P}^\pi)^{-1}\left(\frac{\lambda_w\lambda_Q}{1-\gamma}w^\pi_{d^\mu}+\lambda_w w^\pi\right)\|^2_\Lambda$$

$$\leq\frac{C^2}{(1-\gamma)^2}\left(\frac{\lambda_w\lambda_Q}{1-\gamma}+\lambda_w\right)^2$$

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

Lemma C.5. *Under Assumption C, for arbitrary function $f(s,a)$,*

$$(1-\gamma)\mathbb{E}_{s_0\sim\nu_0,a_0\sim\pi}[f(s_0,a_0)]+\gamma\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi}[w^\pi(s,a)f(s',a')]=\mathbb{E}_{d^\mu}[w^\pi(s,a)f(s,a)]\qquad(18)$$

$$\gamma\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi}[f^2(s',a')]\leq\frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^\pi_{d^\mu}}[f^2(s,a)]\leq\frac{C}{1-\gamma}\mathbb{E}_{s,a\sim d^\mu}[f^2(s,a)]\qquad(19)$$

where $d^\pi_{d^\mu}:=(1-\gamma)\mathbb{E}_{\tau\sim\pi,s_0,a_0\sim d^\mu(\cdot,\cdot)}[\sum_{t=0}^\infty\gamma^t p(s_t=s,a_t=a)]$ is the normalized discounted state-action occupancy by treating $d^\mu(\cdot,\cdot)$ 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.(18) 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.(19), 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 C. $\qquad\square$

Theorem 3.3. *[Bias] Based Assumptions in Section 2.2 and Definitions above, for arbitrary $\theta\in\Theta$, we have:*

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

$$\le \varepsilon_{reg} + \varepsilon_{func} + \varepsilon_{data}$$

The bias terms $\varepsilon_{reg}, \varepsilon_{func}$ and ε_{data} are defined by

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

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

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

$$+ \frac{\gamma C(\lambda_Q + \lambda_Q\lambda_w C)}{(1-\gamma)^3}$$

$$+ \frac{C^2(\lambda_Q + \lambda_Q\lambda_w C)}{(1-\gamma)^3}\big(\frac{\lambda_w\lambda_Q}{1-\gamma} + \lambda_w\big)\sqrt{\frac{\gamma C}{1-\gamma}}\Big)$$

$$\varepsilon_{data} := (2\kappa_\zeta\kappa_\xi + 2\kappa_\zeta + 2\kappa_\xi + \sqrt{2}/2)\sqrt{2\bar{\varepsilon}_{data}}$$

where κ_ζ and κ_ξ is the condition number, and

$$\varepsilon_{\mathcal{W}} := 4\frac{\lambda_{\max}^2}{\lambda_Q\lambda_w}\varepsilon_1 + 2\frac{\lambda_{\max}}{\mu_\zeta}\varepsilon_2,$$

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

with $\lambda_{\max} := \max\{\lambda_Q, \lambda_w\}$.

Proof. Firstly, by applying the triangle inequality:

$$\|\nabla_\theta \max_{w \in \mathcal{W}} \min_{Q \in \mathcal{Q}} \mathcal{L}^D(\pi_\theta, w, Q) - \nabla_\theta J(\pi_\theta)\| \le \underbrace{\|\nabla_\theta \max_{w \in \mathcal{W}} \min_{Q \in \mathcal{Q}} \mathcal{L}^D(\pi_\theta, w, Q) - \nabla_\theta \max_{w \in \mathcal{W}} \min_{Q \in \mathcal{Q}} \mathcal{L}(\pi_\theta, w, Q)\|}_{Bounded~in~Proposition~C.1}$$

$$+ \underbrace{\|\nabla_\theta \max_w \min_Q \mathcal{L}(\pi_\theta, w, Q) - \nabla_\theta \max_{w \in \mathcal{W}} \min_{Q \in \mathcal{Q}} \mathcal{L}(\pi_\theta, w, Q)\|}_{t_1}$$

$$+ \underbrace{\|\nabla_\theta J(\pi_\theta) - \nabla_\theta \max_w \min_Q \mathcal{L}(\pi_\theta, w, Q)\|}_{t_2}$$

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 3.2, we can easily bound t_1:

$$t_1 = \|\nabla_\theta \mathcal{L}(\pi_\theta, w_\mu^*, Q_\mu^*) - \nabla_\theta \mathcal{L}(\pi_\theta, w_{\mathcal{L}}^{\pi_\theta}, Q_{\mathcal{L}}^{\pi_\theta})\|$$

$$\le \frac{1}{1-\gamma}\|(1-\gamma)\mathbb{E}_{\nu_0^{\pi_\theta}}[\big(Q_\mu^*(s_0, a_0) - Q_{\mathcal{L}}^{\pi_\theta}(s_0, a_0)\big)\nabla_\theta \log \pi_\theta(a_0|s_0)]\|$$

$$+ \frac{\gamma}{1-\gamma}\|\mathbb{E}_{s,a,s'\sim d^\mu, a'\sim \pi}[w_\mu^*(s,a)\big(Q_\mu^*(s',a') - Q_{\mathcal{L}}^{\pi_\theta}(s',a')\big)\nabla_\theta \log \pi(a'|s')]\|$$

24

$$+ \frac{\gamma}{1-\gamma}\|\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi}[(w^*_\mu(s,a) - w^{\pi_\theta}_\mathcal{L}(s,a))\Big(Q^*_\mu(s',a') - Q^{\pi_\theta}_\mathcal{L}(s',a')\Big)\nabla_\theta \log\pi(a'|s')]\|$$

$$+ \frac{\gamma}{1-\gamma}\|\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[(w^*_\mu(s,a) - w^{\pi_\theta}_\mathcal{L}(s,a))Q^*_\mu(s',a')\nabla_\theta \log\pi(a'|s')]\|$$

$$\leq \frac{G}{1-\gamma}\mathbb{E}_{\nu^{\pi_\theta}_0}[|Q^*_\mu(s,a) - Q^{\pi_\theta}_\mathcal{L}(s,a)|] + \frac{\gamma C_\mathcal{W}G}{1-\gamma}\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[|Q^*_\mu(s',a') - Q^{\pi_\theta}_\mathcal{L}(s',a')|]$$

$$((1-\gamma)\nu^\pi_0(s,a) \leq d^\pi(s,a) \leq Cd^\mu(s,a))$$

$$+ \frac{\gamma G}{1-\gamma}\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[|(w^*_\mu(s,a) - w^{\pi_\theta}_\mathcal{L}(s,a))\Big(Q^*_\mu(s',a') - Q^{\pi_\theta}_\mathcal{L}(s',a')\Big)|]$$

$$+ \frac{\gamma C_\mathcal{Q}G}{1-\gamma}\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[|w^*_\mu(s,a) - w^{\pi_\theta}_\mathcal{L}(s,a)|]$$

$$\leq \frac{G}{1-\gamma}\sqrt{\mathbb{E}_{\nu^{\pi_\theta}_0}[|Q^*_\mu(s,a) - Q^{\pi_\theta}_\mathcal{L}(s,a)|^2]} + \frac{\gamma C_\mathcal{W}G}{1-\gamma}\sqrt{\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[|Q^*_\mu(s',a') - Q^{\pi_\theta}_\mathcal{L}(s',a')|^2]}$$

$$+ \frac{\gamma G}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|w^{\pi_\theta}_\mathcal{L}(s,a) - w^*_\mu(s,a)|^2]\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[|Q^{\pi_\theta}(s',a') - Q^{\pi_\theta}_\mathcal{L}(s',a')|^2]|}$$

$$+ \frac{\gamma C_\mathcal{Q}G}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|w^*_\mu(s,a) - w^{\pi_\theta}_\mathcal{L}(s,a))|^2]}$$

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

$$+ \frac{G}{1-\gamma}\sqrt{\frac{\gamma C}{1-\gamma}\mathbb{E}_{d^\mu}[|w^{\pi_\theta}_\mathcal{L}(s,a) - w^*_\mu(s,a)|^2]\mathbb{E}_{d^\mu}[|Q^{\pi_\theta}(s,a) - Q^{\pi_\theta}_\mathcal{L}(s,a)|^2]|}$$

$$+ \frac{\gamma C_\mathcal{Q}G}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|w^*_\mu(s,a) - w^{\pi_\theta}_\mathcal{L}(s,a))|^2]}$$

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

In the last equation, we first use Eq.(19) in Lemma C.5, and then apply Proposition C.2.

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

$$t_2 = \|\nabla_\theta J(\pi_\theta) - \nabla_\theta \mathcal{L}(\pi_\theta, w^{\pi_\theta}_\mathcal{L}, Q^{\pi_\theta}_\mathcal{L}))\|$$

$$\leq \frac{1}{1-\gamma}\|(1-\gamma)\mathbb{E}_{\nu^{\pi_\theta}_0}[\Big(Q^{\pi_\theta}(s_0,a_0) - Q^{\pi_\theta}_\mathcal{L}(s_0,a_0)\Big)\nabla_\theta \log\pi_\theta(a_0|s_0)]$$

$$+ \gamma\mathbb{E}_{d^\mu}[w^{\pi_\theta}(s,a)\Big(Q^{\pi_\theta}(s',a') - Q^{\pi_\theta}_\mathcal{L}(s',a')\Big)\nabla_\theta \log\pi(a'|s')]\|$$

$$+ \frac{\gamma}{1-\gamma}\|\mathbb{E}_{d^\mu}[(w^{\pi_\theta}(s,a) - w^{\pi_\theta}_\mathcal{L}(s,a))\Big(Q^{\pi_\theta}(s',a') - Q^{\pi_\theta}_\mathcal{L}(s',a')\Big)\nabla_\theta \log\pi(a'|s')]\|$$

$$+ \frac{\gamma}{1-\gamma}\|\mathbb{E}_{d^\mu}[(w^{\pi_\theta}(s,a) - w^{\pi_\theta}_\mathcal{L}(s,a))Q^{\pi_\theta}(s',a')\nabla_\theta \log\pi(a'|s')]\|$$

$$= \frac{1}{1-\gamma}\|\mathbb{E}_{d^\mu}[w^{\pi_\theta}(s,a)\Big(Q^{\pi_\theta}(s,a) - Q^{\pi_\theta}_\mathcal{L}(s,a)\Big)\nabla_\theta \log\pi(a|s)]\| \qquad\text{(Eq.(18) in Lemma C.5)}$$

$$+ \frac{\gamma}{1-\gamma}\|\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[(w^{\pi_\theta}(s,a) - w^{\pi_\theta}_\mathcal{L}(s,a))\Big(Q^{\pi_\theta}(s',a') - Q^{\pi_\theta}_\mathcal{L}(s',a')\Big)\nabla_\theta \log\pi(a'|s')]\|$$

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

$$\leq \frac{CG}{1-\gamma}\mathbb{E}_{d^\mu}[|Q^{\pi_\theta}(s,a) - Q^{\pi_\theta}_\mathcal{L}(s,a)|]$$

$$+ \frac{\gamma G}{1-\gamma}\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[|(w^{\pi_\theta}(s,a) - w^{\pi_\theta}_\mathcal{L}(s,a))\Big(Q^{\pi_\theta}(s',a') - Q^{\pi_\theta}_\mathcal{L}(s',a')\Big)|]$$

$$+ \frac{\gamma G}{(1-\gamma)^2}\mathbb{E}_{s,a,s'\sim d^\mu,a'\sim\pi_\theta}[|w^{\pi_\theta}(s,a) - w^{\pi_\theta}_\mathcal{L}(s,a)|]$$

25

$$\leq \frac{CG}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|Q^{\pi_\theta} - Q_{\mathcal{L}}^{\pi_\theta}|^2]} + \frac{\gamma G}{(1-\gamma)^2}\sqrt{\mathbb{E}_{d^\mu}[|(w^{\pi_\theta}(s,a) - w_{\mathcal{L}}^{\pi_\theta}(s,a)|^2]}$$

$$+ \frac{\gamma G}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|w_{\mathcal{L}}^{\pi_\theta}(s,a) - w^{\pi_\theta}(s,a)|^2]\mathbb{E}_{s,a,s'\sim d^\mu, a'\sim \pi_\theta}[|Q^{\pi_\theta}(s',a') - Q_{\mathcal{L}}^{\pi_\theta}(s',a')|^2]]}$$

$$\leq \frac{CG}{1-\gamma}\sqrt{\mathbb{E}_{d^\mu}[|Q^{\pi_\theta} - Q_{\mathcal{L}}^{\pi_\theta}|^2]} + \frac{\gamma G}{(1-\gamma)^2}\sqrt{\mathbb{E}_{d^\mu}[|(w^{\pi_\theta}(s,a) - w_{\mathcal{L}}^{\pi_\theta}(s,a)|^2]}$$

$$+ \frac{G}{1-\gamma}\sqrt{\frac{\gamma C}{1-\gamma}\mathbb{E}_{d^\mu}[|w_{\mathcal{L}}^{\pi_\theta}(s,a) - w^{\pi_\theta}(s,a)|^2]\mathbb{E}_{d^\mu}[|Q^{\pi_\theta}(s,a) - Q_{\mathcal{L}}^{\pi_\theta}(s,a)|^2]} \quad \text{(Eq.19 in Lemma C.5)}$$

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

$\square$

D   Missing Examples and Proofs for Strategy 1

D.1   Equivalence between Stationary Points

Theorem D.1. *[Equivalence Between Stationary Points] Under Assumptions in Section 2.2, given Z and Ξ with finite $C_{\mathcal{Q}}$ and $C_{\mathcal{W}}$, suppose there is an Algorithm provides us with one stationary point $(\theta_T, \zeta_T, \xi_T)$ of the non-concave-strongly-convex objective $\max_{\theta,\zeta} \min_\xi \mathcal{L}^D(\theta, \zeta, \xi)$ after running T iterations, statisfying the following conditions in expectation over the randomness of algorithm.*

$$\mathbb{E}[\|\nabla_{\theta,\zeta}\mathcal{L}^D(\theta_T, \zeta_T, \phi_{\theta_T}(\zeta_T))\|] := \mathbb{E}[\|\nabla_\theta\mathcal{L}^D(\theta_T, \zeta_T, \phi_{\theta_T}(\zeta_T))\| + \|\nabla_\zeta\mathcal{L}^D(\theta_T, \zeta_T, \phi_{\theta_T}(\zeta_T))\|]$$
$$\leq \frac{\varepsilon}{(\kappa_\xi + 1)(\kappa_\zeta + 1)} \tag{20}$$

where $\phi_\theta(\zeta) = \arg\min_{\xi\in\Xi}\mathcal{L}^D(\theta,\zeta,\xi)$ and κ_ζ and κ_ξ are the condition numbers, then in expectation θ_T is a biased stationary point satisfying Eq.(2).

Proof. Eq.(20) implies that

$$\max\{\mathbb{E}[\|\nabla_\theta\mathcal{L}^D(\theta_T, \zeta_T, \phi_{\theta_T}(\zeta_T))\|], \mathbb{E}[\|\nabla_\zeta\mathcal{L}^D(\theta_T, \zeta_T, \phi_{\theta_T}(\zeta_T))\|]\} \leq \frac{\varepsilon}{(\kappa_\xi + 1)(\kappa_\zeta + 1)} \tag{21}$$

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

$$\mathbb{E}[\|\nabla_\theta J(\pi_{\theta_T})\|] \leq \underbrace{\mathbb{E}[\|\nabla_\theta\mathcal{L}^D(\theta_T, \zeta_T, \phi_{\theta_T}(\zeta_T))\|]}_{Bounded\ in\ Eq.(21)} + \mathbb{E}[\|\nabla_\theta\mathcal{L}^D(\theta_T, \zeta^*, \xi^*) - \nabla_\theta\mathcal{L}^D(\theta_T, \zeta_T, \phi_{\theta_T}(\zeta_T))\|]$$

$$+ \underbrace{\mathbb{E}[\|\nabla_\theta\mathcal{L}^D(\theta_T, \zeta^*, \xi^*) - \nabla_\theta J(\pi_{\theta_T})\|]}_{Bounded\ in\ Theorem 3.3}$$

$$\leq \frac{\varepsilon}{(\kappa_\xi + 1)(\kappa_\zeta + 1)} + \varepsilon_{func} + \varepsilon_{reg} + \varepsilon_{data} + \mathbb{E}[\|\nabla_\theta\mathcal{L}^D(\theta_T, \zeta^*, \xi^*) - \nabla_\theta\mathcal{L}^D(\theta_T, \zeta_T, \phi_{\theta_T}(\zeta_T))\|]$$

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.(21) and Theorem 3.3.

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, Lemma A.3, and inequality (21), we obtain that

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

Then we can bound:

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

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,

$$\mathbb{E}[\|\nabla_\theta J(\pi_{\theta_T})\|] \leq \frac{\varepsilon}{(\kappa_\xi + 1)(\kappa_\zeta + 1)} + \frac{\varepsilon\kappa_\zeta}{1 + \kappa_\zeta} + \varepsilon_{func} + \varepsilon_{reg} + \varepsilon_{data}$$
$$\leq \varepsilon + \varepsilon_{func} + \varepsilon_{reg} + \varepsilon_{data}$$

$\square$

In the following subsections, we will introduce the Projected-SREDA Algorithm revised from [Luo et al., 2020] and prove that it provide us a stationary points required by Theorem D.1.

We choose $\Theta = \mathbb{R}^{d_\theta}$, $\Xi = \Xi_0$, and $Z = \{\zeta\|\|\zeta\| \leq R'\}$, where Ξ_0 is defined in Property 4.1 and R' will be determined later. For simplicity, we use $\mathcal{L}_-^D = -\mathcal{L}^D$ to denote the minus of original loss function, which should be a non-convex-strongly-concave problem and aligns with the setting of [Luo et al., 2020].

D.2 Verification of the Assumptions in [Luo et al., 2020]

In this section, we verify that Assumptions 1-5 in [Luo et al., 2020] are satisfied under our Assumption A, B and D.

Assumption 1

$$\inf_{\theta\in\mathbb{R}^{d_\theta}, \zeta\in Z} \max_{\xi\in\Xi} \mathcal{L}_-^D(\theta, \zeta, \xi)$$
$$\geq -\max_{\theta\in\mathbb{R}^{d_\theta}, \zeta\in Z, \xi\in\Xi} (1-\gamma)\|(\nu_D^\pi)^\top \mathbf{\Phi}_Q \xi\| + \|\zeta\|\|\mathbf{\Phi}_w^\top \Lambda^D R\| + \|\zeta\|\|\mathbf{M}_\pi\|\|\xi\| + \frac{\lambda_Q}{2}\|\xi\|^2\|\mathbf{K}_Q\| + \frac{\lambda_w}{2}\|\zeta\|\|\mathbf{K}_w\|$$
$$\geq -\max_{\theta\in\mathbb{R}^{d_\theta}, \zeta\in Z, \xi\in\Xi} (1-\gamma) + \|\zeta\| + (1+\gamma)\|\zeta\|\|\xi\| + + \frac{\lambda_Q}{2}\|\xi\|^2 + \frac{\lambda_w}{2}\|\zeta\|$$

Because $\|\zeta\|$ and $\|\xi\|$ are bounded for arbitrary $\zeta, \xi \in Z \times \Xi$, Assumption 1 holds.

Assumption 2 The proof is almost identical to the proof of Property B.1-(b) and we omit here. Assumption 2 holds by choosing L according to (11).

Assumption 3 Under our linear function classes setting, it holds obviously.

Assumption 4 Hold directly by choosing $\mu = \lambda_Q v_Q$.

Assumption 5 Identicial to the Condition B.3. We prove Condition B.3 holds in Appendix B under our Assumptions.

D.3 Useful Lemma

In this subsection, we first prove several useful lemma.

Lemma D.2. *Under Assumption B, for arbitrary $\theta \in \Theta$ and $\xi \in \Xi$, the solution for $\max_{\zeta\in\mathbb{R}^{d_Z}} \mathcal{L}^D(\theta, \zeta, \xi)$ has bounded ℓ_2 norm.*

Proof. Recall the loss function $\mathcal{L}^D$:

$$\mathcal{L}^D(\pi, \zeta, \xi) = (1-\gamma)(\nu_D^\pi)^\top \boldsymbol{\Phi}_Q \xi + \zeta^\top \boldsymbol{\Phi}_w^\top \Lambda^D R - \zeta^\top \mathbf{M}_\pi \xi + \frac{\lambda_Q}{2} \xi^\top \mathbf{K}_Q \xi - \frac{\lambda_w}{2} \zeta^\top \mathbf{K}_w \zeta$$

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

$$\zeta^* = \frac{1}{\lambda_w} \mathbf{K}_w^{-1} (\boldsymbol{\Phi}_w^\top \Lambda^D R - \mathbf{M}_\pi \xi)$$

Given that $\|\xi\| \le R_\xi$ for $\xi \in \Xi$, we have:

$$\|\zeta^*\| \le \frac{1}{\lambda_w} \|\mathbf{K}_w^{-1}\| (\|\boldsymbol{\Phi}^\top \Lambda^D R\| + \|\mathbf{M}_\pi\| \|\xi\|) \le \frac{1}{\lambda_w v_w} (1 + (1+\gamma)R_\xi)$$

$\square$

In the following, we will use $R_0 := \frac{1}{\lambda_w v_w}(1 + (1+\gamma)R_\xi)$ as a shortnote. Next, we are ready to prove the following lemma which is crucial for the analysis of the effect of projection step.

Lemma 4.2. *According to Alg 1, at iteration k, we have $\zeta_k \in Z$, $\zeta_{k+1}^+ = \zeta_k - \eta_k \mathbf{v}_k^\zeta$ and $\zeta_{k+1} = P_Z(\zeta_{k+1}^+)$. Denote $R_0 := \frac{1}{\lambda_w v_w}(1 + (1+\gamma)R_\xi)$ where R_ξ is defined in Property 4.1. If we choose $\Xi = \Xi_0$ and $Z = \{\zeta | \|\zeta\| \le R'\}$ with $R' \ge R_0$, we have:*

$$\langle \nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k), \zeta_{k+1} - \zeta_{k+1}^+ \rangle$$
$$\le \eta_k \frac{R_0 + \eta_k \|\mathbf{v}_k^\zeta\|}{R' + \eta_k \|\mathbf{v}_k^\zeta\|} \|\nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k)\| \|\mathbf{v}_k^\zeta\|.$$

Proof. First of all, if $\zeta_{k+1}^+ \in Z$, then $\zeta_{k+1} - \zeta_{k+1}^+ = 0$, and the Lemma holds. Therefore, in the following, we only consider the case when $\zeta_{k+1}^+ \notin Z$. Because $\zeta_k \in Z$, in the case, we must have $\|\mathbf{v}_k^\zeta\| > 0$.

Because we are considering Z is a high dimensional ball. For $\zeta_{k+1}^+ \notin Z$, we have

$$\zeta_{k+1} = P_Z(\zeta_{k+1}^+) = \zeta_{k+1}^+ \frac{R'}{\|\zeta_{k+1}^+\|}$$

which means,

$$\zeta_{k+1} - \zeta_{k+1}^+ = \left(\frac{R'}{\|\zeta_{k+1}^+\|} - 1\right)\zeta_{k+1}^+$$

Denote $\zeta_k^* = \min_{\zeta \in Z} \mathcal{L}_-^D(\theta_k, \zeta, \xi_k)$. Then we have:

$$\langle \nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k), \zeta_k - \zeta_k^* \rangle \ge 0, \quad \|\zeta_k^*\| \le R_0$$

Then we have:

$$\frac{\langle \nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k), \zeta_{k+1} - \zeta_{k+1}^+ \rangle}{\|\nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k)\| \|\mathbf{v}_k^\zeta\|}$$
$$= \left(\frac{R'}{\|\zeta_{k+1}^+\|} - 1\right) \frac{\langle \nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k), \zeta_{k+1}^+ \rangle}{\|\nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k)\| \|\mathbf{v}_k^\zeta\|}$$
$$= \left(\frac{R'}{\|\zeta_{k+1}^+\|} - 1\right) \frac{\langle \nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k), \zeta_k - \eta_k \mathbf{v}_k^\zeta \pm \zeta_k^* \rangle}{\|\nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k)\| \|\mathbf{v}_k^\zeta\|}$$

28

$$= \left(\frac{R'}{\|\zeta_{k+1}^+\|} - 1\right) \underbrace{\frac{\langle \nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k), \zeta_k - \zeta_k^*\rangle}{\|\nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k)\|\|\mathbf{v}_k^\zeta\|}}_{larger\ than\ 0} + \left(\underbrace{\frac{R'}{\|\zeta_{k+1}^+\|} - 1}_{smaller\ than\ 0}\right) \frac{\langle \nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k), \zeta_k^* - \eta_k \mathbf{v}_k^\zeta\rangle}{\|\nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k)\|\|\mathbf{v}_k^\zeta\|}$$

$$\leq \left(\frac{R'}{\|\zeta_{k+1}^+\|} - 1\right) \frac{\langle \nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k), \zeta_k^* - \eta_k \mathbf{v}_k^\zeta\rangle}{\|\nabla_\zeta \mathcal{L}_-^D(\theta_k, \zeta_k, \xi_k)\|\|\mathbf{v}_k^\zeta\|}$$

$$\leq \left(1 - \frac{R'}{\|\zeta_{k+1}^+\|}\right) \frac{\|\zeta_k^*\| + \eta_k\|\mathbf{v}_k^\zeta\|}{\|\mathbf{v}_k^\zeta\|} \qquad\qquad (\|\zeta_{k+1}^+\| \geq R' \text{ and } \langle a, b\rangle \leq \|a\|\|b\|)$$

$$\leq \left(1 - \frac{R'}{R' + \eta_k\|\mathbf{v}_k^\zeta\|}\right) \frac{R_0 + \eta_k\|\mathbf{v}_k^\zeta\|}{\|\mathbf{v}_k^\zeta\|} \qquad\qquad (\|\zeta_{k+1}^+\| \leq R' + \eta_k\|\mathbf{v}_k^\zeta\|)$$

$$= \eta_k \frac{R_0 + \eta_k\|\mathbf{v}_k^\zeta\|}{R' + \eta_k\|\mathbf{v}_k^\zeta\|}$$

Because $R' = 8\max\{R_0, 1\}$, and $\|\mathbf{v}_k\| \geq \|\mathbf{v}_k^\zeta\|$, we have

$$\frac{R_0 + \eta_k\|\mathbf{v}_k^\zeta\|}{R' + \eta_k\|\mathbf{v}_k^\zeta\|} \leq \frac{R_0 + \eta_k\|\mathbf{v}_k\|}{R' + \eta_k\|\mathbf{v}_k\|} \leq \frac{R_0 + \frac{\varepsilon}{5\kappa_\xi L}}{R' + \frac{\varepsilon}{5\kappa_\xi L}} \leq \frac{R_0 + 1}{R' + 1} \leq \frac{1}{4}$$

where in the third inequality we use the constraint that $\varepsilon < 1$ and the fact that $\kappa_\xi > 1, L > 1$ in our setting. $\qquad\square$

D.4 Main Proofs for Theorem 3.1

The proofs for Theorem D.5 is almost the same as those for the original SREDA algorithm. We will only show those key Lemmas or Theorems in [Luo et al., 2020] which need to be modified as a result of the additional projection step, and omit those untouched. In the following, we will frequently use x_{k+1}^+ to denote $(\theta_{k+1}, \zeta_{k+1}^+)$ before the projection and use x_{k+1} to denote $(\theta_{k+1}, \zeta_{k+1})$ after the projection.

First of all, the following condition still holds

$$\|x_{k+1} - x_k\|^2 \leq \|x_{k+1}^+ - x_k\|^2 \leq \frac{\varepsilon^2}{25\kappa_\xi^2}$$

where the first inequality results from the property of projection and the second one holds because of the choice of learning rate η_k. The condition above corresponds to the $\|x_{k+1} - x_k\|^2 \leq \varepsilon_x^2$ in [Luo et al., 2020]. As a result, all the Lemmas and Theorems in the Appendix B of [Luo et al., 2020] still hold for our Projected-SREDA. Besides, because the PiSARAH will not be effected by our projection step, the results in Appendix C of [Luo et al., 2020] also holds.

Similarly, we consider the following decomposition:

$$\mathcal{L}_-^D(x_{k+1}, \xi_{k+1}) - \mathcal{L}_-^D(x_k, \xi_k) = \underbrace{\mathcal{L}_-^D(x_{k+1}, \xi_k) - \mathcal{L}_-^D(x_k, \xi_k)}_{A_k} + \underbrace{\mathcal{L}_-^D(x_{k+1}, \xi_{k+1}) - \mathcal{L}_-^D(x_{k+1}, \xi_k)}_{B_k}$$

Because the proof of Lemma 14 and Lemma 15 in [Luo et al., 2020] only depends on the previous lemmas, they still hold and we list them here.

Lemma D.3. *Under Assumptions of Theorem D.5, we have* $\mathbb{E}[B_k] \leq \frac{134\varepsilon^2}{\kappa_\xi L}$ *for any* $k \geq 1$

Lemma D.4. *Under Assumptions of Theorem D.5, we have*

$$\mathbb{E}\|\nabla_{\theta,\zeta} \max_{\xi \in \Xi} \mathcal{L}_-^D(\theta, \zeta, \xi)\| \leq \mathbb{E}\|\mathbf{v}_k\| + \frac{15}{7}\varepsilon$$

However, the final proof for Theorem D.5 can not be adapted from [Luo et al., 2020] directly because of the projection. In the following, we show our proof targeted at our Projected-SREDA:

Theorem D.5. *For $\varepsilon < 1$, under Assumption A, B, D, with the following parameter choices:*

$$R' = 8\max\{\frac{1}{\lambda_w v_w}(1 + (1+\gamma)R_\xi), 1\}, R_\xi = \frac{1}{\lambda_w \lambda_Q v_Q + v_\mathbf{M}^2}((1-\gamma)\lambda_w + \frac{1+\gamma}{v_w}),$$

$$\eta_k = \min\left(\frac{\varepsilon}{5\kappa_\xi L\|\mathbf{v}_k\|}, \frac{1}{10\kappa_\xi L}\right), \quad \lambda = \frac{1}{8L}, \quad S_1 = \lceil\frac{2250}{19}\sigma^2\kappa_\xi^2\varepsilon^2\rceil, \quad S_2 = \lceil\frac{3687}{76}\kappa_\xi q\rceil, \quad q = \lceil\varepsilon^{-1}\rceil,$$

$$\Delta\mathcal{L}_-^D = \max_{\theta\in\Theta,\zeta\in Z,\xi\in\Xi}\mathcal{L}_-^D(\theta,\zeta,\xi) - \mathcal{L}_-^D(\theta_0,\zeta_0,\xi_0), \quad K = \lceil\frac{50\kappa_\xi L\Delta\mathcal{L}_-^D}{\varepsilon^2}\rceil \quad and \quad m = \lceil 1024\kappa_\xi\rceil$$

and the same parameter choices for PiSADAH as in [Luo et al., 2020], Algorithm 3 outputs $\widehat{\theta}, \widehat{\zeta}$ such that

$$\mathbb{E}[\|\nabla_{\theta,\zeta}\max_{\xi\in\Xi}\mathcal{L}_-^D(\widehat{\theta},\widehat{\zeta},\xi)\|] \leq O(\varepsilon)$$

with $O(\kappa_\xi^3\varepsilon^{-3})$ stochastic gradient evaluations.

Proof of Theorem D.5. Based on the update rule of θ and ζ in Algorithm 1, we have:

$$A_k \leq \langle\nabla_{\theta,\zeta}\mathcal{L}_-^D(x_k,\xi_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \qquad \text{(Smoothness of } \mathcal{L}_-^D\text{)}$$

$$\leq \langle\nabla_{\theta,\zeta}\mathcal{L}_-^D(x_k,\xi_k), x_{k+1}^+ - x_k\rangle + \langle\nabla_{\theta,\zeta}\mathcal{L}_-^D(x_k,\xi_k), x_{k+1} - x_{k+1}^+\rangle + \frac{L}{2}\|x_{k+1}^+ - x_k\|^2$$

$$\text{(the property of projection)}$$

$$= -\eta_k\langle\nabla_{\theta,\zeta}\mathcal{L}_-^D(x_k,\xi_k), \mathbf{v}_k\rangle + \langle\nabla_\zeta\mathcal{L}_-^D(x_k,\xi_k), \zeta_{k+1} - \zeta_{k+1}^+\rangle + \frac{L\eta_k^2}{2}\|\mathbf{v}_k\|^2$$

$$\leq -\eta_k\langle\nabla_{\theta,\zeta}\mathcal{L}_-^D(x_k,\xi_k), \mathbf{v}_k\rangle + \frac{\eta_k}{4}\|\nabla_\zeta\mathcal{L}_-^D(x_k,\xi_k)\|\|\mathbf{v}_k^\zeta\| + \frac{L\eta_k^2}{2}\|\mathbf{v}_k\|^2 \qquad \text{(Lemma 4.2)}$$

$$\leq -\eta_k\langle\nabla_{\theta,\zeta}\mathcal{L}_-^D(x_k,\xi_k), \mathbf{v}_k\rangle + \frac{\eta_k}{4}\|\mathbf{v}_k\|^2 + \frac{\eta_k}{4}\|\nabla_{\theta,\zeta}\mathcal{L}_-^D(x_k,\xi_k) - \mathbf{v}_k\|\|\mathbf{v}_k\| + \frac{L\eta_k^2}{2}\|\mathbf{v}_k\|^2$$

$$\text{(Triangle Ineq.; } \|\mathbf{v}_k^\zeta\| \leq \|\mathbf{v}_k\|; \|\nabla_\zeta\mathcal{L}_-^D\| \leq \|\nabla_{\theta,\zeta}\mathcal{L}_-^D\|; \|\nabla_\zeta\mathcal{L}_-^D - \mathbf{v}_k^\zeta\| \leq \|\nabla_{\theta,\zeta}\mathcal{L}_-^D - \mathbf{v}_k\|)$$

$$\leq \frac{\eta_k}{2}\|\nabla_{\theta,\zeta}\mathcal{L}_-^D(x_k,\xi_k) - \mathbf{v}_k\|^2 - (\frac{\eta_k}{4} - \frac{L\eta_k^2}{2})\|\mathbf{v}_k\|^2 + \frac{\eta_k}{4}\|\nabla_{\theta,\zeta}\mathcal{L}_-^D(x_k,\xi_k) - \mathbf{v}_k\|\|\mathbf{v}_k\| \quad (\|\nabla_{\theta,\zeta}\mathcal{L}_-^D\| \geq 0)$$

The choice of the step size implies that

$$(\frac{\eta_k}{4} - \frac{L\eta_k^2}{2})\|\mathbf{v}_k\|^2 \geq \min\{\frac{1}{40\kappa_\xi L} - \frac{1}{200\kappa_\xi^2 L}, \frac{\varepsilon}{20\kappa_\xi L\|\mathbf{v}_k\|} - \frac{\varepsilon^2}{50\kappa_\xi^2 L\|\mathbf{v}_k\|^2}\}\|\mathbf{v}_k\|^2$$

$$\geq \min\{\frac{1}{50\kappa_\xi L}, \frac{3\varepsilon}{100\kappa_\xi L\|\mathbf{v}_k\|}\}\|\mathbf{v}_k\|^2$$

$$\geq \frac{\varepsilon^2}{50\kappa_\xi L}\min\left(\frac{\|\mathbf{v}_k\|}{\varepsilon}, \frac{\|\mathbf{v}_k\|^2}{2\varepsilon^2}\right)$$

$$\geq \frac{\varepsilon^2}{50\kappa_\xi L}(\frac{\|\mathbf{v}_k\|}{\varepsilon} - 2)) \qquad (\min(|x|, x^2/2) \geq |x| - 2)$$

$$= \frac{1}{50\kappa_\xi L}(\varepsilon\|\mathbf{v}_k\| - 2\varepsilon^2)$$

Therefore,

$$\mathbb{E}[A_k] \leq \frac{1}{20\kappa_\xi L}\mathbb{E}[\|\nabla_{\theta,\zeta}\mathcal{L}_-^D(x_k,\xi_k) - \mathbf{v}_k\|^2] - \frac{1}{50\kappa_\xi L}(\varepsilon\|\mathbf{v}_k\| - 2\varepsilon^2) + \frac{\varepsilon}{20\kappa_\xi L}\mathbb{E}[\|\nabla_{\theta,\zeta}\mathcal{L}_-^D(x_k,\xi_k) - \mathbf{v}_k\|]$$

$$\leq \frac{1}{20\kappa_\xi L}\mathbb{E}[\|\nabla_{\theta,\zeta}\mathcal{L}_-^D(x_k,\xi_k) - \mathbf{v}_k\|^2] - \frac{1}{50\kappa_\xi L_{30}}(\varepsilon\|\mathbf{v}_k\| - 2\varepsilon^2) + \frac{\varepsilon}{20\kappa_\xi L}\sqrt{\mathbb{E}[\|\nabla_{\theta,\zeta}\mathcal{L}_-^D(x_k,\xi_k) - \mathbf{v}_k\|^2]}$$

$$\leq \frac{1}{20\kappa_\xi L} \cdot \frac{19}{1125}\kappa_\xi^{-2}\varepsilon^2 - \frac{1}{50\kappa_\xi L}(\varepsilon\|\mathbf{v}_k\| - 2\varepsilon^2) + \frac{\varepsilon}{20\kappa_\xi L} \cdot \sqrt{\frac{19}{1125}}\kappa_\xi^{-1}\varepsilon$$

<div align="right">(Corollary 2 in [Luo et al., 2020])</div>

$$\leq \frac{\varepsilon^2}{20\kappa_\xi L} - \frac{1}{50\kappa_\xi L}\varepsilon\|\mathbf{v}_k\|$$

Therefore, combining with Lemma D.3 and taking average over K, we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\mathcal{L}_-^D(x_{k+1}, \xi_{k+1}) - \mathcal{L}_-^D(x_k, \xi_k)] \leq \frac{1}{K}\sum_{k=0}^{K-1}(\frac{\varepsilon^2}{20\kappa_\xi L} - \frac{1}{50\kappa_\xi L}\varepsilon\|\mathbf{v}_k\| + \frac{134\varepsilon^2}{\kappa_\xi L})$$

Consequently, we have:

$$\frac{\varepsilon}{50\kappa_\xi L}\frac{1}{K}\sum_{k=0}^{K-1}\|\mathbf{v}_k\| \leq \frac{135\varepsilon^2}{\kappa_\xi L} + \frac{\Delta\mathcal{L}_-^D}{K}$$

which means

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\mathbf{v}_k\| \leq 6750\varepsilon + \frac{50\kappa_\xi L\Delta\mathcal{L}_-^D}{K\varepsilon}$$

Under Assumptions A and B and Condition that that both Z and Ξ have finite diameter, $\Delta\mathcal{L}_-^D$ is a finite constant. By choosing $K = \lceil\frac{50\kappa_\xi L\Delta\mathcal{L}_-^D}{\varepsilon^2}\rceil$, we have:

$$\mathbb{E}\|\nabla_{\theta,\zeta}\max_{\xi\in\Xi}\mathcal{L}_-^D(\widehat{\theta}, \widehat{\zeta}, \xi)\| = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla_{\theta,\zeta}\max_{\xi\in\Xi}\mathcal{L}_-^D(\theta_k, \zeta_k, \xi)\| \leq \frac{1}{K}\sum_{k=0}^{K-1}(\mathbb{E}\|\mathbf{v}_k\| + \frac{15\varepsilon}{7}) \leq 6754\varepsilon$$

which finishes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

E   Concrete Examples for Saddle-Point Solver Oracle

E.1   Connection with Previous Methods

The inner optimization oracle in our second strategy essentially solves the off-policy policy evaluation problem, i.e., given a policy, compute its value function and marginalized importance weighting function. Among the plethora of works studying off-policy policy evaluation with linear function approximation, [Liu et al., 2020] connected the GTD family and stochastic gradient optimization, and established finite-sample analysis for their off-policy algorithms. Their convergence rate is worse than ours, because their objective is only convex-strongly-concave, whereas our objective is strongly-convex-strongly-concave thanks to the regularization on the parameters of the Q function.

Besides, [Du et al., 2017] adapted variance-reduced stochastic gradient optimization algorithms for policy evaluation, which can be extended to off-policy setting. However, they focused on the finite-sum case and their algorithms will be inefficient for our purpose when the dataset is prohibitively large. Besides, they did not analyze the bias resulting from regularization and approximation error, which we do in this paper.

E.2   The First Example for Saddle-Point Solver Oracle

In the proof of Property 4.1, we show that in our linear setting, it's possible to derive a close form solution for $\zeta^*, \xi^* = \arg\max_{\zeta\in Z}\min_{\xi\in\Xi}\mathcal{L}^D(\theta, \zeta, \xi)$ for arbitrary $\theta\in\Theta$:

$$\zeta_\pi^* = \left(\lambda_w\lambda_Q\mathbf{K}_w + \mathbf{M}_\pi\mathbf{K}_Q^{-1}\mathbf{M}_\pi^\top\right)^{-1}\left(\frac{1}{31}(1-\gamma)\mathbf{M}_\pi\mathbf{K}_Q^{-1}\mathbf{\Phi}_Q^\top\nu_D^\pi + \lambda_Q\mathbf{\Phi}_w^\top\Lambda^D R\right),$$

$$\xi_\pi^* = \left(\lambda_w\lambda_Q\mathbf{K}_Q + \mathbf{M}_\pi^\top\mathbf{K}_w^{-1}\mathbf{M}_\pi\right)^{-1}\left((1-\gamma)\lambda_w\boldsymbol{\Phi}_Q^\top\nu_D^\pi + \mathbf{M}_\pi^\top\mathbf{K}_w^{-1}\boldsymbol{\Phi}_w^\top\Lambda^D R\right).$$

Therefore, we can directly use our sample data to estimate $\mathbf{K}$, $\mathbf{M}$, $\boldsymbol{\Phi}^\top\Lambda^D R$ and $\boldsymbol{\Phi}^\top\nu_D^\pi$ and then estimate (ζ^*, ξ^*). In this section, we provide an algorithm based on this idea in Alg. 3, and prove that such algorithm satisfies the requirement of the ORACLE in Definition 5.1.

E.2.1   Estimation Error of Least-Square Oracle

We first show a useful Lemma:

Lemma E.1 (Matrix Bernstein Theorem (Theorem 6.1.1 in [Tropp, 2015])). *Consider a finite sequence* $\{\mathbf{S}_k\}$ *of independent, random matrices with common dimension* $d_1 \times d_2$. *Assume that*

$$\mathbb{E}\mathbf{S}_k = 0 \quad and \quad \|\mathbf{S}_k\| \leq L \quad for\ each\ index\ k$$

Introduce the random matrix

$$\mathbf{Z} = \sum_k \mathbf{S}_k$$

Let $v(\mathbf{Z})$ *be the matrix variance statistic of the sum:*

$$v(\mathbf{Z}) = \max\{\|\mathbb{E}(\mathbf{Z}\mathbf{Z}^*)\|, \|\mathbb{E}(\mathbf{Z}^*\mathbf{Z})\|\}$$
$$= \max\{\|\sum_k \mathbb{E}(\mathbf{S}_k\mathbf{S}_k^*)\|, \|\sum_k \mathbb{E}(\mathbf{S}_k^*\mathbf{S}_k)\|\}$$

Then,

$$\mathbb{E}\|\mathbf{Z}\| \leq \sqrt{2v(\mathbf{Z})\log(d_1+d_2)} + \frac{1}{3}L\log(d_1+d_2)$$

Furthermore, for all $t \geq 0$

$$\mathbb{P}\{\|\mathbf{Z}\| \geq t\} \leq (d_1+d_2)\exp\left(\frac{-t^2/2}{v(\mathbf{Z}) + Lt/3}\right)$$

For Algorithm 3, we have the following guarantee:

Proposition E.2. *Suppose we have* $N - (s, a, r, s', a', a_0)$ *tuples independently sampling according to* $d^D \times \pi$, *and* $N \geq \max\{2\sigma_\mathbf{K}^2 + \frac{4\sigma_{\min}}{3}, 2\sigma_\mathbf{M}^2 + \frac{8(1+\gamma)\sigma_{\min}}{3}\}\frac{4}{\sigma_{\min}^2}\log\frac{2d}{\delta}$. *For arbitrary* $1/5 > \delta > 0$, *with probability at least* $1 - 5\delta$, $\widehat{\mathbf{K}}_w$, $\widehat{\mathbf{K}}_Q$, $\widehat{\mathbf{M}}$ *are invertible and their smallest eigenvalues (for* $\widehat{\mathbf{M}}$ *we consider the smallest sigular values) are larger than* $v_w/2, v_Q/2$ *and* $v_M/2$, *respectively, while the following conditions hold at the same time:*

$$\|\mathbf{K}_w - \widehat{\mathbf{K}}_w\| \leq \frac{2}{3N_{all}}\log\frac{2d}{\delta} + \sqrt{\frac{2\sigma_\mathbf{K}^2}{N_{all}}\log\frac{2d}{\delta}} = O\left(\frac{\sigma_\mathbf{K}}{\sqrt{N_{all}}}\right)$$

$$\|\mathbf{K}_Q - \widehat{\mathbf{K}}_Q\| \leq \frac{2}{3N_{all}}\log\frac{2d}{\delta} + \sqrt{\frac{2\sigma_\mathbf{K}^2}{N_{all}}\log\frac{2d}{\delta}} = O\left(\frac{\sigma_\mathbf{K}}{\sqrt{N_{all}}}\right)$$

$$\|\mathbf{M}_\pi - \widehat{\mathbf{M}}_\pi\| = \|\mathbf{M}_\pi^\top - \widehat{\mathbf{M}}_\pi^\top\| \leq \frac{4(1+\gamma)}{3N_{all}}\log\frac{2d}{\delta} + \sqrt{\frac{2\sigma_\mathbf{M}^2}{N_{all}}\log\frac{2d}{\delta}} = O\left(\frac{\sigma_\mathbf{M}}{\sqrt{N_{all}}}\right)$$

$$\|\widehat{\mathbf{u}}_\nu^\pi - \mathbf{u}_\nu^\pi\| \leq \frac{4}{3N_{all}}\log\frac{2d}{\delta} + \sqrt{\frac{2\sigma_\nu^2}{N_{all}}\log\frac{2d}{\delta}} = O\left(\frac{\sigma_\nu}{\sqrt{N_{all}}}\right)$$

$$\|\widehat{\mathbf{u}}_R - \mathbf{u}_R\| \leq \frac{4}{3N_{all}}\log\frac{2d}{\delta} + \sqrt{\frac{2\sigma_R^2}{N_{all}}\log\frac{2d}{\delta}} = O\left(\frac{\sigma_R}{\sqrt{N_{all}}}\right)$$

Proof. First, we try to bound $\|\mathbf{K}_w - \widehat{\mathbf{K}}_w\|$. We take a look at a series of random matrices

$$A_i = \mathbf{K}_w - \boldsymbol{\phi}_w(s_i, a_i)\boldsymbol{\phi}_w(s_i, a_i)^\top, \quad i = 1, 2, ...$$

where (s_i, a_i) are sampled from d^D independently. Easy to verify that A_i has the following properties as a result of $\|\boldsymbol{\phi}_w(s, a)\| \leq 1$:

$$\mathbb{E}[A_i] = 0, \quad \|A_i\| = \|\mathbf{K}_w - \boldsymbol{\phi}_w(s, a)^\top \boldsymbol{\phi}_w(s_i, a_i)^\top\| \leq 1$$

Under Assumption D, according to the Matrix Bernstein Theorem and the union bound, we have:

$$P(\frac{\|\sum_{i=1}^N A_i\|}{N_{all}} \geq \varepsilon) \leq P(\sigma_{\max}(\sum_{i=1}^N A_i) \geq N_{all}\varepsilon) + P(\sigma_{\max}(\sum_{i=1}^N -A_i) \geq N_{all}\varepsilon)$$

$$\leq 2d \exp(-\frac{N_{all}\varepsilon^2/2}{\sigma_{\mathbf{K}}^2 + \varepsilon/3})$$

which implies that, with probability $1 - \delta$:

$$\|\mathbf{K}_w - \widehat{\mathbf{K}}_w\| \leq \frac{2}{3N_{all}} \log \frac{2d}{\delta} + \sqrt{\frac{2\sigma_{\mathbf{K}}^2}{N_{all}} \log \frac{2d}{\delta}} = O(\frac{\sigma_{\mathbf{K}}}{\sqrt{N_{all}}})$$

The discussion for $\widehat{\mathbf{K}}_Q$ is similar and we omit here. As for $\|\mathbf{M}_\pi - \widehat{\mathbf{M}}_\pi\|$, notice that,

$$\|\mathbf{M}_\pi - (\boldsymbol{\phi}_w(s, a)\boldsymbol{\phi}_Q(s, a)^\top - \gamma\boldsymbol{\phi}_w(s, a)\boldsymbol{\phi}_Q(s', a'))\| \leq 2(1 + \gamma)$$
$$\|\mathbf{M}_\pi^\top - (\boldsymbol{\phi}_Q(s, a)\boldsymbol{\phi}_w(s, a)^\top - \gamma\boldsymbol{\phi}_Q(s', a')\boldsymbol{\phi}_w(s, a))\| \leq 2(1 + \gamma)$$

Therefore, w.p. $1 - \delta$,

$$\|\mathbf{M}_\pi - \widehat{\mathbf{M}}_\pi\| = \|\mathbf{M}_\pi^\top - \widehat{\mathbf{M}}_\pi^\top\| \leq \frac{4(1 + \gamma)}{3N_{all}} \log \frac{2d}{\delta} + \sqrt{\frac{2\sigma_{\mathbf{M}}^2}{N_{all}} \log \frac{2d}{\delta}} = O(\frac{\sigma_{\mathbf{M}}}{\sqrt{N_{all}}})$$

As for the bounds for the difference of vectors $\|\widehat{\mathbf{u}}_\nu^\pi - \mathbf{u}_\nu^\pi\|$ and $\|\widehat{\mathbf{u}}_R - \mathbf{u}_R\|$, since vectors are special cases of matrices (or we can concatenate the vector with a zero-vector to make a $n \times 2$ matrix and make the proof more rigorous), again, we can apply the same technique again. Notice that,

$$\|\mathbf{u}_R - \boldsymbol{\phi}_w(s, a)r(s, a)\| \leq 2$$
$$\|\mathbf{u}_\nu^\pi - \boldsymbol{\phi}_Q(s_0, a_0)\| \leq 2$$

Besides, for random column vectors x with bounded norm, we should have

$$\|\mathbb{E}_x[xx^\top]\| \leq tr(\mathbb{E}_x[xx^\top]) = \mathbb{E}_x[tr(xx^\top)] = \mathbb{E}_x[x^\top x] = \|\mathbb{E}_x[x^\top x]\|$$

As a result, combining Assumption D, we have

$$\|\widehat{\mathbf{u}}_\nu^\pi - \mathbf{u}_\nu^\pi\| \leq \frac{4}{3N_{all}} \log \frac{2d}{\delta} + \sqrt{\frac{2\sigma_\nu^2}{N_{all}} \log \frac{2d}{\delta}} = O(\frac{\sigma_\nu}{\sqrt{N_{all}}}), \quad w.p. \ 1 - \delta$$

$$\|\widehat{\mathbf{u}}_R - \mathbf{u}_R\| \leq \frac{4}{3N_{all}} \log \frac{2d}{\delta} + \sqrt{\frac{2\sigma_R^2}{N_{all}} \log \frac{2d}{\delta}} = O(\frac{\sigma_R}{\sqrt{N_{all}}}), \quad w.p. \ 1 - \delta$$

Therefore, w.p. $1 - 2\delta$ the concentration results in this Proposition hold. Next, we derive the smallest eigenvalues of $\widehat{\mathbf{K}}_w$, $\widehat{\mathbf{K}}_Q$ and $\widehat{\mathbf{M}}_\pi$ when n is large enough. For arbitrary $x \in \mathbb{R}^{d \times 1}$ with $\|x\| = 1$, we have:

$$x^\top \widehat{\mathbf{K}}_w x = x^\top \mathbf{K}_w x + x^\top (\widehat{\mathbf{K}}_w - \mathbf{K}_w)x \geq v_w - \|\mathbf{K}_w - \widehat{\mathbf{K}}_w\|$$

33

$$x^\top \widehat{\mathbf{K}}_Q x = x^\top \mathbf{K}_Q x + x^\top (\widehat{\mathbf{K}}_Q - \mathbf{K}_Q) x \geq v_Q - \|\mathbf{K}_Q - \widehat{\mathbf{K}}_Q\|$$
$$|x^\top \widehat{\mathbf{M}}_\pi x| \geq |x^\top \mathbf{M} x| - |x^\top (\widehat{\mathbf{M}} - \mathbf{M}) x| \geq v_M - \|\mathbf{M} - \widehat{\mathbf{M}}\|$$

Therefore, easy to verify that, when the concentration results hold, and $N \geq \max\{8\frac{\sigma_{\mathbf{K}}^2}{v_w^2} + \frac{16}{3v_w}, 8\frac{\sigma_{\mathbf{K}}^2}{v_Q^2} + \frac{16}{3v_Q}, 8\frac{\sigma_{\mathbf{K}}^2}{v_M^2} + \frac{32(1+\gamma)}{3v_M}\} \log \frac{2d}{\delta}$, $\widehat{\mathbf{K}}_w, \widehat{\mathbf{K}}_Q, \widehat{\mathbf{M}}$ are invertible and their smallest eigenvalues (for $\widehat{\mathbf{M}}$ we consider the smallest sigular values) are larger than $v_w/2, v_Q/2$ and $v_M/2$ $\qquad\square$

Next, we are ready to show that Algorithm 3 is also an example of the Oracle Algorithm.

Theorem E.3. *Algorithm 3 satisfies the Oracle Condition 5.1 with $\beta = 0$ and arbitrary $c > 0$, with a proper choice of N_{all}.*

Proof. Denote ζ^*, ξ^* as the saddle-point of $\mathcal{L}^D$ given θ, and use $\widehat{\zeta}^*, \widehat{\xi}^*$ to denote the ζ and ξ in Algorithm 3 before the projection. Given the proposition above, we are ready to bound $\|\zeta^* - \widehat{\zeta}^*\|$ and $\|\xi^* - \widehat{\xi}^*\|$. We list two properties below which we will use frequently later. Firstly, for arbitrary invertible $d \times d$ matrices $\mathbf{A}$ and $\mathbf{B}$, we have:

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| = \|\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}\| \leq \|\mathbf{A}^{-1}\|\|\mathbf{B} - \mathbf{A}\|\|\mathbf{B}^{-1}\|$$

Secondly, for arbitrary $d \times d$ matrices $\mathbf{X}_1, \mathbf{Y}_1, \mathbf{X}_2, \mathbf{Y}_2$, we have:

$$\|\mathbf{X}_1 \mathbf{Y}_1 - \mathbf{X}_2 \mathbf{Y}_2\| = \|\mathbf{X}_1(\mathbf{Y}_1 - \mathbf{Y}_2) + \mathbf{Y}_2(\mathbf{X}_1 - \mathbf{X}_2)\| \leq \|\mathbf{X}_1\|\|\mathbf{Y}_1 - \mathbf{Y}_2\| + \|\mathbf{Y}_2\|\|\mathbf{X}_1 - \mathbf{X}_2\|$$

Then, we have

$$\begin{aligned}
&\|\zeta^* - \widehat{\zeta}^*\| \\
\leq &(1-\gamma)\|\Big(\big(\lambda_w \lambda_Q \mathbf{K}_w + \mathbf{M}_\pi \mathbf{K}_Q^{-1} \mathbf{M}_\pi^\top\big)^{-1} \mathbf{M}_\pi \mathbf{K}_Q^{-1} - \big(\lambda_w \lambda_Q \widehat{\mathbf{K}}_w + \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1} \widehat{\mathbf{M}}_\pi^\top\big)^{-1} \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1}\Big)\mathbf{u}_\nu^\pi\| \\
&+ (1-\gamma)\|\big(\lambda_w \lambda_Q \widehat{\mathbf{K}}_w + \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1} \widehat{\mathbf{M}}_\pi^\top\big)^{-1} \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1}(\mathbf{u}_\nu^\pi - \widehat{\mathbf{u}}_\nu^\pi)\| \\
&+ \|\big(\lambda_w \lambda_Q \widehat{\mathbf{K}}_w + \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1} \widehat{\mathbf{M}}_\pi^\top\big)^{-1} \lambda_Q(\mathbf{u}_R - \widehat{\mathbf{u}}_R)\| \\
&+ \|\Big(\big(\lambda_w \lambda_Q \mathbf{K}_w + \mathbf{M}_\pi \mathbf{K}_Q^{-1} \mathbf{M}_\pi^\top\big)^{-1} - \big(\lambda_w \lambda_Q \widehat{\mathbf{K}}_w + \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1} \widehat{\mathbf{M}}_\pi^\top\big)^{-1}\Big) \lambda_Q \mathbf{u}_R\| \\
\leq &(1-\gamma)\|\big(\lambda_w \lambda_Q \mathbf{K}_w + \mathbf{M}_\pi \mathbf{K}_Q^{-1} \mathbf{M}_\pi\big)^{-1}\|\|\lambda_w \lambda_Q \mathbf{K}_w + \mathbf{M}_\pi \mathbf{K}_Q^{-1} \mathbf{M}_\pi - \lambda_w \lambda_Q \widehat{\mathbf{K}}_w - \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1} \widehat{\mathbf{M}}_\pi^\top\| \\
&\cdot \|\big(\lambda_w \lambda_Q \widehat{\mathbf{K}}_w + \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1} \widehat{\mathbf{M}}_\pi^\top\big)^{-1}\|\|\mathbf{M}_\pi \mathbf{K}_Q^{-1}\| \\
&+ (1-\gamma)\|\big(\lambda_w \lambda_Q \widehat{\mathbf{K}}_w + \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1} \widehat{\mathbf{M}}_\pi^\top\big)^{-1} \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1}\|\|\mathbf{M}_\pi \mathbf{K}_Q^{-1} - \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1}\| \\
&+ (1-\gamma)\frac{\sigma_\nu}{\sqrt{N}}\|\big(\lambda_w \lambda_Q \widehat{\mathbf{K}}_w + \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1} \widehat{\mathbf{M}}_\pi^\top\big)^{-1} \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1}\| + \lambda_Q \frac{\sigma_R}{\sqrt{N_{all}}}\|\big(\lambda_w \lambda_Q \widehat{\mathbf{K}}_w + \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1} \widehat{\mathbf{M}}_\pi^\top\big)^{-1}\| \\
&+ \lambda_Q\|\big(\lambda_w \lambda_Q \mathbf{K}_w + \mathbf{M}_\pi \mathbf{K}_Q^{-1} \mathbf{M}_\pi\big)^{-1}\|\|\lambda_w \lambda_Q \mathbf{K}_w + \mathbf{M}_\pi \mathbf{K}_Q^{-1} \mathbf{M}_\pi - \lambda_w \lambda_Q \widehat{\mathbf{K}}_w - \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1} \widehat{\mathbf{M}}_\pi^\top\| \\
&\cdot \|\big(\lambda_w \lambda_Q \widehat{\mathbf{K}}_w + \widehat{\mathbf{M}}_\pi \widehat{\mathbf{K}}_Q^{-1} \widehat{\mathbf{M}}_\pi^\top\big)^{-1}\| \\
= &O\Big(\frac{1}{v_Q(\lambda_w \lambda_Q v_w + v_{\mathbf{M}}^2)^2 \sqrt{N_{all}}}(\lambda_w \lambda_Q \sigma_{\mathbf{K}} + \frac{\sigma_{\mathbf{M}}}{v_Q} + \frac{\sigma_{\mathbf{K}}}{v_Q^2}) + \frac{\sigma_\nu}{v_Q(\lambda_w \lambda_Q v_w + v_{\mathbf{M}}^2)\sqrt{N_{all}}} + \frac{\lambda_Q \sigma_R}{(\lambda_w \lambda_Q v_w + v_{\mathbf{M}}^2)\sqrt{N_{all}}} \\
&+ \frac{\lambda_Q}{(\lambda_w \lambda_Q v_w + v_{\mathbf{M}}^2)^2 \sqrt{N_{all}}}(\lambda_w \lambda_Q \sigma_{\mathbf{K}} + \frac{\sigma_{\mathbf{M}}}{v_Q} + \frac{\sigma_{\mathbf{K}}}{v_Q^2})\Big) \log \frac{2d}{\delta}
\end{aligned}$$

$$=O\left(\frac{1+\lambda_Q v_Q}{v_Q(\lambda_w\lambda_Q v_w+v_{\mathbf{M}}^2)^2}(\lambda_w\lambda_Q\sigma_{\mathbf{K}}+\frac{\sigma_{\mathbf{M}}}{v_Q}+\frac{\sigma_{\mathbf{K}}}{v_Q^2})+\frac{\sigma_\nu}{v_Q(\lambda_w\lambda_Q v_w+v_{\mathbf{M}}^2)}+\frac{\lambda_Q\sigma_R}{(\lambda_w\lambda_Q v_w+v_{\mathbf{M}}^2)}\right)\sqrt{N_{all}}\log\frac{2d}{\delta}$$

where we omit constant number in $O(\cdot)$.

Similarly, we have

$$\|\xi^*-\widehat{\xi}^*\|$$
$$\le\|(1-\gamma)\lambda_w\left(\left(\lambda_w\lambda_Q\mathbf{K}_Q+\mathbf{M}_\pi^\top\mathbf{K}_w^{-1}\mathbf{M}_\pi\right)^{-1}-\left(\lambda_w\lambda_Q\widehat{\mathbf{K}}_Q+\widehat{\mathbf{M}}_\pi^\top\widehat{\mathbf{K}}_w^{-1}\widehat{\mathbf{M}}_\pi\right)^{-1}\right)\mathbf{u}_\nu^\pi\|$$
$$\quad+\|(1-\gamma)\lambda_w\left(\lambda_w\lambda_Q\widehat{\mathbf{K}}_Q+\widehat{\mathbf{M}}_\pi^\top\widehat{\mathbf{K}}_w^{-1}\widehat{\mathbf{M}}_\pi\right)^{-1}(\mathbf{u}_\nu^\pi-\widehat{\mathbf{u}}_\nu^\pi)\|$$
$$\quad+\|\left(\left(\lambda_w\lambda_Q\mathbf{K}_Q+\mathbf{M}_\pi^\top\mathbf{K}_w^{-1}\mathbf{M}_\pi\right)^{-1}\mathbf{M}_\pi^\top\mathbf{K}_w^{-1}-\left(\lambda_w\lambda_Q\widehat{\mathbf{K}}_Q+\widehat{\mathbf{M}}_\pi^\top\widehat{\mathbf{K}}_w^{-1}\widehat{\mathbf{M}}_\pi\right)^{-1}\widehat{\mathbf{M}}_\pi^\top\widehat{\mathbf{K}}_w^{-1}\right)\mathbf{u}_R\|$$
$$\quad+\|\left(\lambda_w\lambda_Q\widehat{\mathbf{K}}_Q+\widehat{\mathbf{M}}_\pi^\top\widehat{\mathbf{K}}_w^{-1}\widehat{\mathbf{M}}_\pi\right)^{-1}\widehat{\mathbf{M}}_\pi^\top\widehat{\mathbf{K}}_w^{-1}(\mathbf{u}_R-\widehat{\mathbf{u}}_R)\|$$
$$=O\left(\frac{\lambda_w v_w+1}{v_w(\lambda_w\lambda_Q v_Q+v_{\mathbf{M}}^2)^2}(\lambda_w\lambda_Q\sigma_{\mathbf{K}}+\frac{\sigma_{\mathbf{M}}}{v_w}+\frac{\sigma_{\mathbf{K}}}{v_w^2})+\frac{\sigma_R}{v_w(\lambda_w\lambda_Q v_Q+v_{\mathbf{M}}^2)}+\frac{\lambda_w\sigma_\nu}{(\lambda_w\lambda_Q v_Q+v_{\mathbf{M}}^2)}\right)\sqrt{N_{all}}\log\frac{2d}{\delta}$$

As a result, for arbitrary $1/5>\delta>0$, w.p. at least $1-5\delta$,

$$\sqrt{\|\zeta^*-\widehat{\zeta}^*\|^2+\|\xi^*-\widehat{\xi}^*\|^2}\le\|\zeta^*-\widehat{\zeta}^*\|+\|\xi^*-\widehat{\xi}^*\|$$
$$=O\left(\frac{1+\lambda_Q v_Q}{v_Q(\lambda_w\lambda_Q v_w+v_{\mathbf{M}}^2)^2\sqrt{N_{all}}}(\lambda_w\lambda_Q\sigma_{\mathbf{K}}+\frac{\sigma_{\mathbf{M}}}{v_Q}+\frac{\sigma_{\mathbf{K}}}{v_Q^2})+\frac{\sigma_\nu}{v_Q(\lambda_w\lambda_Q v_w+v_{\mathbf{M}}^2)\sqrt{N_{all}}}+\frac{\lambda_Q\sigma_R}{(\lambda_w\lambda_Q v_w+v_{\mathbf{M}}^2)\sqrt{N_{all}}}\right.$$
$$\left.+\frac{\lambda_w v_w+1}{v_w(\lambda_w\lambda_Q v_Q+v_{\mathbf{M}}^2)^2}(\lambda_w\lambda_Q\sigma_{\mathbf{K}}+\frac{\sigma_{\mathbf{M}}}{v_w}+\frac{\sigma_{\mathbf{K}}}{v_w^2})+\frac{\sigma_R}{v_w(\lambda_w\lambda_Q v_Q+v_{\mathbf{M}}^2)}+\frac{\lambda_w\sigma_\nu}{(\lambda_w\lambda_Q v_Q+v_{\mathbf{M}}^2)}\right)\sqrt{N_{all}}\log\frac{2d}{\delta}$$

$$(22)$$

For simplicity, we use C_{LS} to as a shorthand of

$$O\left(\frac{1+\lambda_Q v_Q}{v_Q(\lambda_w\lambda_Q v_w+v_{\mathbf{M}}^2)^2}(\lambda_w\lambda_Q\sigma_{\mathbf{K}}+\frac{\sigma_{\mathbf{M}}}{v_Q}+\frac{\sigma_{\mathbf{K}}}{v_Q^2})+\frac{\sigma_\nu}{v_Q(\lambda_w\lambda_Q v_w+v_{\mathbf{M}}^2)}+\frac{\lambda_Q\sigma_R}{(\lambda_w\lambda_Q v_w+v_{\mathbf{M}}^2)}\right.$$
$$\left.+\frac{\lambda_w v_w+1}{v_w(\lambda_w\lambda_Q v_Q+v_{\mathbf{M}}^2)^2}(\lambda_w\lambda_Q\sigma_{\mathbf{K}}+\frac{\sigma_{\mathbf{M}}}{v_w}+\frac{\sigma_{\mathbf{K}}}{v_w^2})+\frac{\sigma_R}{v_w(\lambda_w\lambda_Q v_Q+v_{\mathbf{M}}^2)}+\frac{\lambda_w\sigma_\nu}{(\lambda_w\lambda_Q v_Q+v_{\mathbf{M}}^2)}\right)$$

Next, we try to convert the above high probability bound to an upper bound for expectation. We use X to denote $\|\zeta^*-\widehat{\zeta}^*\|^2+\|\xi^*-\widehat{\xi}^*\|^2$ and treat it as a r.v.. Then, we have:

$$P(X\le\frac{C_{LS}^2}{N_{all}}\log^2\frac{2d}{\delta})\ge1-5\delta$$

or equivalently,

$$P(X\le x)\ge1-10d\exp(-\frac{\sqrt{xN_{all}}}{C_{LS}})$$

Therefore, computing the expectation of the distribution described by the following C.D.F. can provide us an upper bound for $\mathbb{E}[X]$:

$$F_X(x)=P(X\le x)=1-10d\exp(-\frac{\sqrt{xN_{all}}}{C_{LS}})$$

As a result,

$$\mathbb{E}[X]\le\int_{x=0}^\infty(1-F_X(x))\mathrm{d}x=\int_{x=0}^\infty10d\exp(-\frac{\sqrt{xN_{all}}}{C_{LS}})\mathrm{d}x$$

$$=10d \int_{x=0}^{\infty} \exp(-x)\mathrm{d}\frac{C_{LS}^2}{N_{all}}x^2 = 20d\frac{C_{LS}^2}{N_{all}}$$

which means that,

$$\mathbb{E}[\|\zeta^* - \widehat{\zeta}^*\|^2 + \|\xi^* - \widehat{\xi}^*\|^2] = 20d\frac{C_{LS}^2}{N_{all}} \tag{23}$$

Because $\zeta \in Z$ and $\xi \in \Xi$, the projection can only shrink the distance:

$$\mathbb{E}[\|\zeta^* - P_Z(\widehat{\zeta}^*)\|^2 + \|\xi^* - P_\Xi(\widehat{\xi}^*)\|^2] \leq \mathbb{E}[\|\zeta^* - \widehat{\zeta}^*\|^2 + \|\xi^* - \widehat{\xi}^*\|^2]$$

As a result, for arbitrary c, we can choose $N_{all} = 20d\frac{C_{LS}^2}{c}$ in Algorithm 3, and Algorithm 3 satisfies the oracle condition with $\beta = 0$:

$$\mathbb{E}[\|P_Z(\zeta) - \zeta^*\|^2 + \|P_\Xi(\xi) - \xi^*\|^2] \leq c \tag{24}$$

$\square$

E.3 The Second Example for Saddle-Point Solver Oracle

In this section, we provide another example for the oracle in Definition 5.1 based on first-order optimization, which is inspired by SVRE[Chavdarova et al., 2019].

E.3.1 Proofs for Algorithm 4

For simplification, we will use $\omega = [\zeta, \xi] \in Z \times \Xi := \Omega$ to denote the vector concatenated by ζ and ξ. Similarly, $g_t = [-g_t^\zeta, g_t^\xi]$, and $F_N(\omega) = \mathbb{E}_{(s,a,r,s',a_0,a') \sim N}\{[-\nabla_\zeta \mathcal{L}^{(s,a,r,s',a_0,a')}(\theta, \zeta, \xi), \nabla_\xi \mathcal{L}^{(s,a,r,s',a_0,a')}(\theta, \zeta, \xi)]\}$, where N is the mini batch data sampled according to d^D, and $\nabla \mathcal{L}^{(s,a,r,s',a_0,a')}(\theta, \zeta, \xi)$ is the gradient computed with one sample (s, a, r, s', a_0, a'). We use $F(\omega) := \mathbb{E}_{N \sim D}[F_N(\omega)]$ to denote the gradient expected over entire dataset distribution. Besides,

$$\eta g_t = [-\eta_\zeta g_t^\zeta, \eta_\xi g_t^\xi]; \qquad \eta^2 \|\omega\|^2 = \eta_\theta^2 \|\zeta\|^2 + \eta_\xi^2 \|\xi\|^2; \qquad \mu\|\omega\|^2 = \mu_\zeta\|\zeta\|^2 + \mu_\xi\|\xi\|^2$$

$$\bar{L}^2 \|w\|^2 = \bar{L}_\zeta^2 \|\zeta\|^2 + \bar{L}_\xi^2 \|\xi\|^2; \qquad \eta^2 \bar{L}^2 = \eta_\zeta^2 \bar{L}_\zeta^2 + \eta_\xi^2 \bar{L}_\xi^2; \qquad \eta\mu = \eta_\zeta \mu_\zeta + \eta_\xi \mu_\xi$$

The update rule for Algorithm 4 can be summarized as

$$\text{Extrapolation}: \quad \omega_{t+1/2} = P_\Omega(\omega_t - \eta g_t)$$
$$\text{Update}: \quad \omega_{t+1} = P_\Omega(\omega_t - \eta g_{t+1/2})$$

Besides, in this section, the expectation $\mathbb{E}$ concerns all the randomness starting from the beginning of the algorithm.

Lemma E.4 (Lemma 1 in [Chavdarova et al., 2019]). *Let $\omega \in \Omega$ and $\omega^+ := P_\Omega(w + u)$, then for all $w' \in \Omega$, we have*

$$\|\omega^+ - \omega'\|^2 \leq \|\omega - \omega'\|^2 + 2u^\top(\omega^+ - \omega') - \|\omega^+ - \omega\|^2$$

Lemma E.5 (Adapted from Lemma 3 in [Chavdarova et al., 2019]). *For any $w \in \Omega$, when $t > 0$, we have*

$$\|\omega_{t+1} - \omega\|^2 \leq \|\omega_t - \omega\|^2 - 2\eta g_{t+1/2}^\top(\omega_{t+1/2} - \omega) + \eta^2 \|g_t - g_{t+1/2}\| - \|\omega_{t+1/2} - \omega_t\|^2$$

and when $t = 0$, we have

$$\|\omega_1 - \omega\|^2 \leq \|\omega_0 - \omega\|^2 - 2\eta g_0^\top(\omega_1 - \omega)$$

36

Proof. For $t = 0$, by simply applying Lemma E.4 for $(\omega, u, \omega^+, \omega') = (\omega_0, -\eta g_0^\top, \omega_1, \omega)$, we have:

$$\|\omega_1 - \omega\|^2 \leq \|\omega_0 - \omega\|^2 - 2\eta g_0^\top (\omega_1 - \omega) - \|\omega_1 - \omega_0\|^2 \leq \|\omega_0 - \omega\|^2 - 2\eta g_0^\top (\omega_1 - \omega)$$

For $t > 0$, the proof is exactly the same as Lemma 3 in [Chavdarova et al., 2019]

$\square$

Lemma E.6 (Bound $\|g_t - g_{t+1/2}\|^2$). *For $t > 0$, we have:*

$$\mathbb{E}[\|g_t - g_{t+1/2}\|^2] \leq 10\mathbb{E}[\|F_N(w_t) - F_N(\omega^*)\|^2] + 10\mathbb{E}[\|F_N(\omega^*) - F_N(w_{t-1})\|^2] + 5\bar{L}^2\mathbb{E}[\|w_t - w_{t+1/2}\|]$$

Proof. For $t > 0$:

$$
\begin{aligned}
&\mathbb{E}[\|g_t - g_{t+1/2}\|^2] \\
=&\mathbb{E}[\|F_N(w_t) - F_N(w_{t-1}) + m_t - F_{N'}(w_{t+1/2}) + F_{N'}(w_{t-1}) - m_t\|^2] \\
=&\mathbb{E}[\|F_N(w_t) \pm F_N(w^*) - F_N(w_{t-1}) - F_{N'}(w_{t+1/2}) \pm F_{N'}(w_t) \pm F_{N'}(w^*) + F_{N'}(w_{t-1})\|^2] \\
\leq&5\mathbb{E}[\|F_N(w_t) - F_N(\omega^*)\|^2] + 5\mathbb{E}[\|F_N(\omega^*) - F_N(w_{t-1})\|^2] \\
&+ 5\mathbb{E}[\|F_{N'}(w_{t+1/2}) - F_{N'}(w_t)\|^2]] + 5\mathbb{E}[\|F_{N'}(w_t) - F_{N'}(\omega^*)\|^2] + 5\mathbb{E}[\|F_{N'}(\omega^*) - F_{N'}(w_{t-1})\|^2] \\
=&10\mathbb{E}[\|F_N(w_t) - F_N(\omega^*)\|^2] + 10\mathbb{E}[\|F_N(\omega^*) - F_N(w_{t-1})\|^2] + 5\mathbb{E}[\|F_{N'}(w_{t+1/2}) - F_{N'}(w_t)\|^2]]
\end{aligned}
$$

where in the inequality we use the extended Young's inequality; in the last equation we use the fact that

$$\mathbb{E}_{N \sim D}[\|F_N(w_t) - F_N(w)\|^2] = \mathbb{E}_{N' \sim D}[\|F_{N'}(w_t) - F_{N'}(w)\|^2], \qquad \forall w \in \Omega$$

Besides, according to Assumption B.4

$$\mathbb{E}[\|F_{N'}(w_{t+1/2}) - F_{N'}(w_t)\|^2]] \leq \bar{L}^2\mathbb{E}[\|w_t - w_{t+1/2}\|]$$

As a result,

$$\mathbb{E}[\|g_t - g_{t+1/2}\|^2] \leq 10\mathbb{E}[\|F_N(w_t) - F_N(\omega^*)\|^2] + 10\mathbb{E}[\|F_N(\omega^*) - F_N(w_{t-1})\|^2] + 5\bar{L}^2\mathbb{E}[\|w_t - w_{t+1/2}\|]$$

$\square$

Proposition E.7. *With Property B.1, for arbitrary θ, the operator $F(\omega)$ satisfying:*

$$\left(F(\omega_1) - F(\omega_2)\right)^\top \left(\omega_1 - \omega_2\right) \geq \mu\|\omega_1 - \omega_2\|^2$$

Proof. Based on Property B.1, we have:

$$
\begin{aligned}
-\mathcal{L}^D(\theta, \zeta_1, \xi_2) &\geq -\mathcal{L}^D(\theta, \zeta_2, \xi_2) - \nabla_\zeta \mathcal{L}^D(\theta, \zeta_1, \xi_1)^\top (\zeta_2 - \zeta_1) + \frac{\mu_\zeta}{2}\|\zeta_2 - \zeta_1\|^2 \\
-\mathcal{L}^D(\theta, \zeta_2, \xi_1) &\geq -\mathcal{L}^D(\theta, \zeta_2, \xi_2) - \nabla_\zeta \mathcal{L}^D(\theta, \zeta_2, \xi_2)^\top (\zeta_1 - \zeta_2) + \frac{\mu_\zeta}{2}\|\zeta_2 - \zeta_1\|^2 \\
\mathcal{L}^D(\theta, \zeta_1, \xi_2) &\geq \mathcal{L}^D(\theta, \zeta_1, \xi_1) + \nabla_\xi \mathcal{L}^D(\theta, \zeta_1, \xi_1)^\top (\xi_2 - \xi_1) + \frac{\mu_\xi}{2}\|\xi_2 - \xi_1\|^2 \\
\mathcal{L}^D(\theta, \zeta_2, \xi_1) &\geq \mathcal{L}^D(\theta, \zeta_2, \xi_2) + \nabla_\xi \mathcal{L}^D(\theta, \zeta_2, \xi_2)^\top (\xi_1 - \xi_2) + \frac{\mu_\xi}{2}\|\xi_2 - \xi_1\|^2
\end{aligned}
$$

Sum up and we can obtain

$$
\begin{aligned}
&\left(F(\omega_1) - F(\omega_2)\right)^\top \left(\omega_1 - \omega_2\right) := \left(F(\zeta_1, \xi_1) - F(\zeta_2, \xi_2)\right)^\top \left([\zeta_1, \xi_1] - [\zeta_2, \xi_2]\right) \\
=&-\left(\nabla_\zeta \mathcal{L}^D(\theta, \zeta_1, \xi_1) - \nabla_\zeta \mathcal{L}^D(\theta, \zeta_2, \xi_2)\right)^\top (\zeta_1 - \zeta_2) + \left(\nabla_\xi \mathcal{L}^D(\theta, \zeta_1, \xi_1) - \nabla_\xi \mathcal{L}^D(\theta, \zeta_2, \xi_2)\right)^\top (\xi_1 - \xi_2) \\
\geq&\mu_\zeta\|\zeta_2 - \zeta_1\|^2 + \mu_\xi\|\xi_2 - \xi_1\|^2 := \mu\|\omega_1 - \omega_2\|^2
\end{aligned}
$$

$\square$

Theorem 5.4. *Under Assumptions in Section 2.2, in Algorithm 4, if step sizes satisfy*

$$\eta_\zeta \le \frac{1}{50\max\{\bar{L}_\zeta, \mu_\zeta\}}, \qquad \eta_\xi \le \frac{1}{50\max\{\bar{L}_\xi, \mu_\xi\}}$$

after K iterations, the algorithm will output (ζ_K, ξ_K):

$$\mathbb{E}[\|\zeta_K - \zeta^*\|^2 + \|\xi_K - \xi^*\|^2]$$
$$\le \frac{201}{100}\left(1 - \frac{\mu\eta}{4}\right)^K \mathbb{E}[\|\zeta_0 - \zeta^*\|^2 + \|\xi_0 - \xi^*\|^2]$$
$$+ \frac{8\sigma^2}{\min\{\frac{\mu_\zeta\eta_\zeta}{4}, \frac{\mu_\xi\eta_\xi}{4}\}|N|}\left(\frac{\eta_\zeta}{\mu_\zeta} + \frac{\eta_\xi}{\mu_\xi}\right)$$

where (ζ^*, ξ^*) is the saddle point of $\mathcal{L}^D(\theta, \zeta, \xi)$ given input θ.

Proof. When $t > 0$, from Lemma E.5, we have

$$\|\omega_{t+1} - \omega^*\|^2 \le \|\omega_t - \omega^*\|^2 - 2\eta g_{t+1/2}^\top(\omega_{t+1/2} - \omega^*) - \|\omega_{t+1/2} - \omega_t\|^2 + \eta^2\|g_t - g_{t+1/2}\|^2$$

Next, we use P_{t+1} to denote $\mathbb{E}[\|\omega_{t+1} - \omega^*\|^2] + \tau\mathbb{E}[\|F_N(\omega^*) - F_N(w_t)\|^2]$, where τ will be determined later, then we have

$$
\begin{aligned}
P_{t+1} =& \mathbb{E}[\|\omega_{t+1} - \omega^*\|^2] + \tau\mathbb{E}[\|F_N(\omega^*) - F_N(w_t)\|^2]\\
\le& \mathbb{E}[\|\omega_t - \omega^*\|^2] - 2\eta\mathbb{E}[F(\omega_{t+1/2})^\top(\omega_{t+1/2} - \omega^*)] - \mathbb{E}[\|\omega_{t+1/2} - \omega_t\|^2]\\
&+ \eta^2\mathbb{E}[\|g_t - g_{t+1/2}\|^2] + \tau\mathbb{E}[\|F_N(\omega^*) - F_N(w_t)\|^2]\\
&- 2\eta\mathbb{E}[(m_t - F(\omega_{t-1}))^\top(\omega_{t+1/2} - \omega^*)]\\
&\qquad\qquad (\mathbb{E}[g_{t+1/2}^\top(\omega_{t+1/2} - \omega^*)] = \mathbb{E}[(F(\omega_{t+1/2}) - F(\omega_{t-1}) + m_t)^\top(\omega_{t+1/2} - \omega^*)])\\
\le& \mathbb{E}[\|\omega_t - \omega^*\|^2] - 2\eta\mathbb{E}[F(\omega_{t+1/2})^\top(\omega_{t+1/2} - \omega^*)] - (1 - 5\eta^2\bar{L}^2)\mathbb{E}[\|\omega_{t+1/2} - \omega_t\|^2]\\
&+ (\tau + 10\eta^2)\mathbb{E}[\|F_N(w_t) - F_N(\omega^*)\|^2] + 10\eta^2\mathbb{E}[\|F_N(\omega^*) - F_N(w_{t-1})\|^2]\\
&+ 2\eta\sqrt{\mathbb{E}[\|m_t - F(\omega_{t-1})\|^2]\mathbb{E}[\|\omega_{t+1/2} - \omega^*\|^2]}\\
&\qquad\qquad (\text{Lemma E.6 and Cauthy Inequality: } \mathbb{E}[a^\top b|c] \le \sqrt{\mathbb{E}[\|a\|^2|c]\mathbb{E}[\|b\|^2|c]})\\
\le& \mathbb{E}[\|\omega_t - \omega^*\|^2] - 2\eta\mathbb{E}[F(\omega_{t+1/2})^\top(\omega_{t+1/2} - \omega^*)] - (1 - 5\eta^2\bar{L}^2)\mathbb{E}[\|\omega_{t+1/2} - \omega_t\|^2]\\
&+ (\tau + 10\eta^2)\mathbb{E}[\|F_N(w_t) - F_N(\omega^*)\|^2] + 10\eta^2\mathbb{E}[\|F_N(\omega^*) - F_N(w_{t-1})\|^2]\\
&+ \frac{8\eta}{\mu}\mathbb{E}[\|m_t - F(\omega_{t-1})\|^2] + \frac{\mu\eta}{8}\mathbb{E}[\|\omega_{t+1/2} - \omega^*\|^2] \qquad\qquad (2\sqrt{|a^\top b|} \le \|a\|^2 + \|b\|^2)\\
\le& \mathbb{E}[\|\omega_t - \omega^*\|^2] - 2\eta\mathbb{E}[F(\omega_{t+1/2})^\top(\omega_{t+1/2} - \omega^*)] - (1 - 25\eta^2\bar{L}^2 - 2\tau\bar{L}^2)\mathbb{E}[\|\omega_{t+1/2} - \omega_t\|^2]\\
&+ (2\tau + 20\eta^2)\mathbb{E}[\|F_N(w_{t+1/2}) - F_N(\omega^*)\|^2] + 10\eta^2\mathbb{E}[\|F_N(\omega^*) - F_N(w_{t-1})\|^2]\\
&+ \frac{8\sigma^2}{|N|}\left(\frac{\eta_\zeta}{\mu_\zeta} + \frac{\eta_\xi}{\mu_\xi}\right) + \frac{\mu\eta}{4}(\mathbb{E}[\|\omega_{t+1/2} - \omega_t\|^2] + \mathbb{E}[\|\omega_t - \omega^*\|^2])\\
&\qquad (\text{Condition B.3; Young's Inequality; } \mathbb{E}[\|F_N(\omega_{t+1/2}) - F_N(\omega_t)\|^2] \le \bar{L}^2\mathbb{E}[\|\omega_{t+1/2} - \omega_t\|^2])\\
\le& \mathbb{E}[\|\omega_t - \omega^*\|^2] - 2\eta\mathbb{E}[F(\omega_{t+1/2})^\top(\omega_{t+1/2} - \omega^*)] - (1 - 25\eta^2\bar{L}^2 - 2\tau\bar{L}^2)\mathbb{E}[\|\omega_{t+1/2} - \omega_t\|^2]\\
&+ (2\tau\bar{L} + 20\eta^2\bar{L})\mathbb{E}[(F_N(\omega^*) - F_N(w_{t+1/2}))^\top(\omega^* - w_{t+1/2})] + 10\eta^2\mathbb{E}[\|F_N(\omega^*) - F_N(w_{t-1})\|^2]\\
&+ \frac{8\sigma^2}{|N|}\left(\frac{\eta_\zeta}{\mu_\zeta} + \frac{\eta_\xi}{\mu_\xi}\right) + \frac{\mu\eta}{4}(\mathbb{E}[\|\omega_{t+1/2} - \omega_t\|^2] + \mathbb{E}[\|\omega_t - \omega^*\|^2]) \qquad (\text{Assumption B.4})\\
=& \mathbb{E}[\|\omega_t - \omega^*\|^2] - (2\eta - 20\bar{L}\eta^2 - 2\tau\bar{L})\mathbb{E}[(F(\omega_{t+1/2}) - F(w^*))^\top(\omega_{t+1/2} - \omega^*)]\\
&- (1 - 25\eta^2\bar{L}^2 - 2\tau\bar{L}^2)\mathbb{E}[\|\omega_{t+1/2} - \omega_t\|^2] + 10\eta^2\mathbb{E}[\|F_N(\omega^*) - F_N(w_{t-1})\|^2]
\end{aligned}
$$

38

$$+ \frac{8\sigma^2}{|N|}\left(\frac{\eta_\varsigma}{\mu_\varsigma} + \frac{\eta_\xi}{\mu_\xi}\right) + \frac{\mu\eta}{4}\left(\mathbb{E}[\|\omega_{t+1/2} - \omega_t\|^2 + \mathbb{E}[\|\omega_t - \omega^*\|^2])\right)$$

By Prop. E.7, we have:

$$(F(w^*) - F(\omega_{t+1/2}))^\top(\omega^* - \omega_{t+1/2}) \geq \mu\|\omega^* - \omega_{t+1/2}\|^2 \geq \frac{\mu}{2}\|w_t - \omega^*\|^2 - \mu\|w_{t+1/2} - w_t\|^2 \tag{25}$$

By choosing $0 < \eta_\varsigma \leq \frac{1}{50\max\{\bar{L}_\varsigma, \mu_\varsigma\}}$, $0 < \eta_\xi \leq \frac{1}{50\max\{\bar{L}_\xi, \mu_\xi\}}$, and $\tau = 15\eta^2$, we know

$$2\eta - 20\bar{L}\eta^2 - 2\tau\bar{L} = 2\eta - 50\bar{L}\eta^2 \geq 0$$

As a result, we can use (25) to get:

$$
\begin{aligned}
P_{t+1} \leq & \left(1 - \mu\eta + 10\mu\eta^2\bar{L} + \tau\mu\bar{L} + \frac{\mu\eta}{4}\right)\mathbb{E}[\|\omega_t - \omega^*\|^2] \\
& - \left(1 - 25\eta^2\bar{L}^2 - 2\tau\bar{L}^2 - 2\mu\eta + 20\mu\bar{L}\eta^2 + 2\mu\tau\bar{L} - \frac{\mu\eta}{4}\right)\mathbb{E}[\|\omega_{t+1/2} - \omega_t\|^2] \\
& + \frac{10\eta^2}{\tau}\tau\mathbb{E}[\|F_N(\omega^*) - F_N(w_{t-1})\|^2] + \frac{8\sigma^2}{|N|}\left(\frac{\eta_\varsigma}{\mu_\varsigma} + \frac{\eta_\xi}{\mu_\xi}\right) \\
\leq & \underbrace{\left(1 - \frac{3}{4}\mu\eta + 25\mu\eta^2\bar{L}\right)}_{p_1}\mathbb{E}[\|\omega_t - \omega^*\|^2] + \underbrace{\left(55\eta^2\bar{L}^2 + \frac{9}{4}\mu\eta - 50\mu\bar{L}\eta^2 - 1\right)}_{p_2}\mathbb{E}[\|\omega_{t+1/2} - \omega_t\|^2] \\
& + \frac{2}{3}\tau\mathbb{E}[\|F_N(\omega^*) - F_N(w_{t-1})\|^2] + \frac{8\sigma^2}{|N|}\left(\frac{\eta_\varsigma}{\mu_\varsigma} + \frac{\eta_\xi}{\mu_\xi}\right)
\end{aligned}
$$

since $0 < \eta\mu \leq 1/50$ and $0 < \eta\bar{L} \leq 1/50$

$$
\begin{aligned}
p_1 \leq & 1 - \frac{3}{4}\mu\eta + \frac{25\mu\eta}{50} = 1 - \frac{\mu\eta}{4} \\
p_2 \leq & \frac{11}{500} + \frac{9}{200} - 1 \leq \frac{11}{500} + \frac{9}{200} - 1 \leq 0
\end{aligned}
$$

As a result

$$
\begin{aligned}
P_{t+1} \leq & \left(1 - \frac{\mu\eta}{4}\right)\mathbb{E}[\|w_t - \omega^*\|^2] + \frac{2}{3}\tau\mathbb{E}[\|F_N(\omega^*) - F_N(w_{t-1})\|^2] + \frac{8\sigma^2}{|N|}\left(\frac{\eta_\varsigma}{\mu_\varsigma} + \frac{\eta_\xi}{\mu_\xi}\right) \\
\leq & \left(1 - \min\{\frac{\mu\eta}{4}, \frac{1}{3}\}\right)P_t + \frac{8\sigma^2}{|N|}\left(\frac{\eta_\varsigma}{\mu_\varsigma} + \frac{\eta_\xi}{\mu_\xi}\right) \\
= & \left(1 - \frac{\mu\eta}{4}\right)P_t + \frac{8\sigma^2}{|N|}\left(\frac{\eta_\varsigma}{\mu_\varsigma} + \frac{\eta_\xi}{\mu_\xi}\right) & (\mu\eta \leq 1/50) \\
\leq & \left(1 - \frac{\mu\eta}{4}\right)^t P_1 + \frac{8\sigma^2}{\min\{\frac{\mu_\varsigma\eta_\varsigma}{4}, \frac{\mu_\xi\eta_\xi}{4}\}|N|}\left(\frac{\eta_\varsigma}{\mu_\varsigma} + \frac{\eta_\xi}{\mu_\xi}\right) \\
= & \left(1 - \frac{\mu\eta}{4}\right)^t\left(\mathbb{E}[\|w_1 - \omega^*\|] + \tau\mathbb{E}[\|F_N(\omega^*) - F_N(\omega_0)\|^2]\right) + \frac{8\sigma^2}{\min\{\frac{\mu_\varsigma\eta_\varsigma}{4}, \frac{\mu_\xi\eta_\xi}{4}\}|N|}\left(\frac{\eta_\varsigma}{\mu_\varsigma} + \frac{\eta_\xi}{\mu_\xi}\right)
\end{aligned}
$$

Next, we take a look at $\mathbb{E}[\|w_1 - \omega^*\|]$, from Lemma E.5, we have:

$$
\begin{aligned}
& \mathbb{E}[\|\omega_1 - \omega^*\|^2] \\
\leq & \mathbb{E}[\|\omega_0 - \omega^*\|^2 - 2\eta g_0^\top(\omega_1 - \omega^*)] \\
= & \mathbb{E}[\|\omega_0 - \omega^*\|^2] + 2\eta\mathbb{E}[(F(\omega_0) - g_0)^\top(\omega_1 - \omega^*)] + 2\eta\mathbb{E}[(F(\omega^*) - F(\omega_0))^\top(\omega_1 - \omega^*)]
\end{aligned}
$$

$$\leq \mathbb{E}[\|\omega_0 - \omega^*\|^2] + 2\eta\mathbb{E}[\|F(\omega_0) - g_0\|^2]\mathbb{E}[\|\omega_1 - \omega^*\|^2] + 2\eta\mathbb{E}[(F(\omega^*) - F(\omega_0))^\top(\omega_1 - \omega^*)]$$

$$\leq \mathbb{E}[\|\omega_0 - \omega^*\|^2] + \frac{2\eta\sigma^2}{|N|}\mathbb{E}[\|\omega_1 - \omega^*\|] + 3\eta^2\mathbb{E}[\|F(\omega^*) - F(\omega_0)\|^2] + \frac{1}{3}\mathbb{E}[\|\omega_1 - \omega^*\|^2]$$

$$\leq \mathbb{E}[\|\omega_0 - \omega^*\|^2] + \frac{\mu\eta}{4}\mathbb{E}[\|\omega_1 - \omega^*\|] + 3\eta^2\mathbb{E}[\|F(\omega^*) - F(\omega_0)\|^2] + \frac{1}{3}\mathbb{E}[\|\omega_1 - \omega^*\|^2]$$

$$\leq \mathbb{E}[\|\omega_0 - \omega^*\|^2] + \frac{1}{2}\mathbb{E}[\|\omega_1 - \omega^*\|] + 3\eta^2\mathbb{E}[\|F(\omega^*) - F(\omega_0)\|^2] \qquad (\eta\mu < 1/50)$$

Therefore,

$$\mathbb{E}[\|\omega_1 - \omega\|^2] \leq 2\mathbb{E}[\|\omega_0 - \omega^*\|^2] + 6\eta^2\mathbb{E}[\|F(\omega^*) - F(\omega_0)\|^2]$$

Finally, using the fact that $\mathbb{E}[\|F_N(\omega^*) - F_N(\omega_0)\|^2] \leq \bar{L}^2\mathbb{E}[\|\omega^* - \omega_0\|^2]$, we have

$$\mathbb{E}[\|\omega_{t+1} - \omega^*\|^2] - \frac{8\sigma^2}{\min\{\frac{\mu_\zeta\eta_\zeta}{4}, \frac{\mu_\xi\eta_\xi}{4}\}|N|}\Big(\frac{\eta_\zeta}{\mu_\zeta} + \frac{\eta_\xi}{\mu_\xi}\Big) \leq P_{t+1} - \frac{8\sigma^2}{\min\{\frac{\mu_\zeta\eta_\zeta}{4}, \frac{\mu_\xi\eta_\xi}{4}\}|N|}\Big(\frac{\eta_\zeta}{\mu_\zeta} + \frac{\eta_\xi}{\mu_\xi}\Big)$$

$$\leq \Big(1 - \frac{\mu\eta}{4}\Big)^t \Big(2\mathbb{E}[\|\omega_0 - \omega^*\|^2] + 6\eta^2\mathbb{E}[\|F(\omega^*) - F(\omega_0)\|^2] + \tau\mathbb{E}[\|F_N(\omega^*) - F_N(\omega_0)\|^2]\Big)$$

$$= \Big(1 - \frac{\mu\eta}{4}\Big)^t \Big(2\mathbb{E}[\|\omega_0 - \omega^*\|^2] + 6\eta^2\bar{L}^2\mathbb{E}[\|\omega^* - \omega_0\|^2] + 15\eta^2\bar{L}^2\mathbb{E}[\|\omega^* - \omega_0\|^2]\Big)$$

$$\leq \Big(1 - \frac{\mu\eta}{4}\Big)^t \Big(2 + \frac{6}{2500} + \frac{15}{2500}\Big)\mathbb{E}[\|\omega^* - \omega_0\|^2]$$

$$\leq \frac{201}{100}\Big(1 - \frac{\mu\eta}{4}\Big)^t\mathbb{E}[\|\omega^* - \omega_0\|^2]$$

which finishes the proof. □

F Missing details for Algorithm 2

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 F.1. *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. Under Assumptions in Section 2.2 and Condition 4.1, we have*

$$\|\zeta_1^* - \zeta_2^*\| \leq \kappa_\zeta(\kappa_\xi + 1)\|\theta_1 - \theta_2\|, \quad \|\xi_1^* - \xi_2^*\| \leq \kappa_\xi(\kappa_\zeta + 1)\|\theta_1 - \theta_2\|$$

Proof. With Condition 4.1, we have

$$\|\nabla_\zeta\mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| = \|\nabla_\zeta\mathcal{L}^D(\theta_1, \zeta_1^*, \xi_1^*) - \nabla_\zeta\mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| \leq L\|\theta_1 - \theta_2\| \qquad (26)$$

$$\|\nabla_\xi\mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| = \|\nabla_\xi\mathcal{L}^D(\theta_1, \zeta_1^*, \xi_1^*) - \nabla_\xi\mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| \leq L\|\theta_1 - \theta_2\| \qquad (27)$$

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

$$\|\zeta_1^* - \zeta_2^*\| \leq \frac{1}{\mu_\zeta}\|\nabla_\zeta\Phi_{\theta_2}(\zeta_1^*)\| = \frac{1}{\mu_\zeta}\|\nabla_\zeta\mathcal{L}^D(\theta_2, \zeta_1^*, \phi_{\theta_2}(\zeta_1^*))\|$$

$$\leq \frac{1}{\mu_\zeta}\|\nabla_\zeta\mathcal{L}^D(\theta_2, \zeta_1^*, \phi_{\theta_2}(\zeta_1^*)) - \nabla_\zeta\mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| + \frac{1}{\mu_\zeta}\|\nabla_\zeta\mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*))\|$$

$$\leq \frac{1}{\mu_\zeta}\|\nabla_\zeta\mathcal{L}^D(\theta_2, \zeta_1^*, \phi_{\theta_2}(\zeta_1^*)) - \nabla_\zeta\mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| + \frac{L}{\mu_\zeta}\|\theta_1 - \theta_2\|$$

$$\leq \frac{L}{\mu_\zeta}\|\phi_{\theta_2}(\zeta_1^*) - \xi_1^*\| + \frac{L}{\mu_\zeta}\|\theta_1 - \theta_2\|$$

$$\leq \frac{L}{\mu_\zeta\mu_\xi}\|\nabla_\xi\mathcal{L}^D(\theta_2, \zeta_1^*, \xi_1^*)\| + \frac{L}{\mu_\zeta}\|\theta_1 - \theta_2\|$$

$$\leq \kappa_\zeta(\kappa_\xi + 1)\|\theta_1 - \theta_2\|$$

where in the first step, we use Lemma A.3; in the fourth and fifth inequalities, we use the Property 2.2; in the last inequality, we use Eq.(26) again.

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

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

$\square$

Lemma 5.2. *[Relate the shift of ζ_t and ξ_t with θ_t] Denote $(\theta_t, \zeta_t, \xi_t)$ as the parameter value at the beginning at the step t in Algorithm 2, and denote $(\zeta_t^*, \xi_t^*) = \arg\max_{\zeta \in Z} \min_{\xi \in \Xi} \mathcal{L}^D(\theta_t, \zeta, \xi)$ as the only saddle point given θ_t. Under Assumptions in Section 2.2, we have:*

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

where $d := \max\{C_{\mathcal{W}}, C_{\mathcal{Q}}\}$ is the maximum of diameters of Z and Ξ, $C_{\zeta,\xi}$ is a short note of $\kappa_\mu^2(\kappa_\xi + 1)^2 + \kappa_\xi^2(\kappa_\mu + 1)^2$, $g_\theta^\tau := \frac{1}{\eta_\theta}(\theta^{\tau+1} - \theta^\tau)$, and η_θ is the step size of θ.

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 $t \geq 1$

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

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

Property 2 For $t \geq 0$,

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

41

$$\leq \beta^{t+1}d^2 + \eta_\theta^2 C_{\zeta,\xi} \sum_{\tau=0}^{t-1} \beta^{t-\tau}\mathbb{E}[\|g_\theta^\tau\|^2] + \frac{c}{1-\beta}$$

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

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

With the two properties above, we can bound:

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

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 $0 < \beta \leq 1$. $\qquad \square$

Lemma F.2. *Define $L^{D*}(\theta) := \max_{\zeta \in Z} \min_{\xi \in \Xi} \mathcal{L}^D(\theta, \zeta, \xi)$. Under the same condition of Lemma 5.2 above, with an additional constraint $\beta \leq (1-\alpha)^2/2$, for $t \geq 0$, we have:*

$$\mathbb{E}[\|g_\theta^{t+1} - \nabla_\theta \mathcal{L}^{D*}(\theta_{t+1})\|^2] \leq 2\lambda^{t+1}\mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2] + 2\frac{\alpha^2\sigma^2}{(1-\lambda)|B|}$$
$$+ 4L^2\left(\beta^{t+2}d^2 + \frac{c}{1-\beta} + \frac{3\lambda c}{(1-\beta)(1-\lambda)} + \frac{3d^2\beta\lambda^{t+2}}{\lambda-\beta}\right)$$
$$+ 2\eta_\theta^2 L^2 \mathbb{E}\left[\sum_{i=0}^{t}\left(\lambda^{t-i+1}\left(6C_{\zeta,\xi}\frac{\lambda}{\lambda-\beta} + 1\right) + 2C_{\zeta,\xi}\beta^{t-i+1}\right)\|g_\theta^i\|^2\right]$$

where $\varepsilon_{data}, \varepsilon_{func}, \varepsilon_{reg}$ are the same as those in Theorem 3.3, and λ as a shortnote of $2(1-\alpha)^2$.

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 $\nabla_\theta \mathcal{L}^B(\theta_t, \zeta_t, \xi_t)$, $\nabla_\theta \mathcal{L}^D(\theta_t, \zeta_t, \xi_t)$, $\nabla_\theta \mathcal{L}^{D*}(\theta_t)$ respectively, and in Property B.3, we have already shown the variance of batch gradient can be bounded. First we can use the Young's inequality to obtain

$$\mathbb{E}[\|g_\theta^{t+1} - \nabla_\theta \mathcal{L}_{t+1}^{D*}\|^2] \leq 2\underbrace{\mathbb{E}[\|g_\theta^{t+1} - \nabla_\theta \mathcal{L}_{t+1}^D\|^2]}_{p_1} + 2\underbrace{\mathbb{E}[\|\nabla_\theta \mathcal{L}_{t+1}^D - \nabla_\theta \mathcal{L}_{t+1}^{D*}\|^2]}_{p_2}$$

Since the first term has already been bounded in Theorem 3.3. Next, we bound p_1 and p_2:

Upper bound p_1 We again use $C_{\zeta,\xi}$ as a short note of $\kappa_\mu^2(\kappa_\xi+1)^2 + \kappa_\xi^2(\kappa_\zeta+1)^2$. From Lemma 5.2, we know that,

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

In the following, we will use λ as a shortnote of $2(1-\alpha)^2$.

$$p_1 = \mathbb{E}[\|g_\theta^{t+1} - \nabla_\theta\mathcal{L}_{t+1}^D\|^2]$$

$$=\mathbb{E}\left[\left\|(1-\alpha)g_\theta^t + \alpha\nabla_\theta\mathcal{L}_{t+1}^B - \nabla_\theta\mathcal{L}_{t+1}^D \pm (1-\alpha)\nabla_\theta\mathcal{L}_t^D\right\|^2\right]$$

$$=\mathbb{E}\left[\left\|(1-\alpha)(g_\theta^t - \nabla_\theta\mathcal{L}_t^D) + \alpha(\nabla_\theta\mathcal{L}_{t+1}^B - \nabla_\theta\mathcal{L}_{t+1}^D) + (1-\alpha)(\nabla_\theta\mathcal{L}_t^D - \nabla_\theta\mathcal{L}_{t+1}^D)\right\|^2\right]$$

$$=\mathbb{E}\left[\left\|(1-\alpha)(g_\theta^t - \nabla_\theta\mathcal{L}_t^D) + (1-\alpha)(\nabla_\theta\mathcal{L}_t^D - \nabla_\theta\mathcal{L}_{t+1}^D)\right\|^2\right] + \mathbb{E}\left[\left\|\alpha(\nabla_\theta\mathcal{L}_{t+1}^B - \nabla_\theta\mathcal{L}_{t+1}^D)\right\|^2\right]$$
$$\text{(Drop 0 expectation)}$$

$$\leq\lambda\mathbb{E}\left[\left\|(g_\theta^t - \nabla_\theta\mathcal{L}_t^D)\right\|^2\right] + \lambda\mathbb{E}\left[\left\|(\nabla_\theta\mathcal{L}_t^D - \nabla_\theta\mathcal{L}_{t+1}^D)\right\|^2\right] + \frac{\alpha^2\sigma^2}{|B|} \qquad\text{(Young's Ineq.)}$$

$$\leq\lambda\mathbb{E}\left[\left\|(g_\theta^t - \nabla_\theta\mathcal{L}_t^D)\right\|^2\right] + \lambda L^2\mathbb{E}[\|\zeta_{t+1}-\zeta_t\|^2 + \|\xi_{t+1}-\xi_t\|^2 + \|\theta_{t+1}-\theta_t\|^2] + \frac{\alpha^2\sigma^2}{|B|} \quad\text{(Smoothness of } \mathcal{L}^D)$$

$$\leq\lambda^{t+1}\mathbb{E}[\|g_\theta^0 - \nabla_\theta\mathcal{L}_0^D\|^2] + \frac{\alpha^2\sigma^2}{|B|}\frac{1-\lambda^{t+1}}{1-\lambda} + \mathbb{E}\left[\sum_{i=0}^{t}\lambda^{t-i+1}L^2\left(\|\zeta_{i+1}-\zeta_i\|^2 + \|\xi_{i+1}-\xi_i\|^2 + \|\theta_{i+1}-\theta_i\|^2\right)\right]$$

$$\leq\lambda^{t+1}\mathbb{E}[\|g_\theta^0 - \nabla_\theta\mathcal{L}_0^D\|^2] + \frac{\alpha^2\sigma^2}{(1-\lambda)|B|}$$
$$+ \mathbb{E}\left[\sum_{i=0}^{t}\lambda^{t-i+1}L^2\left(6\beta^{i+1}d^2 + 6\eta_\theta^2 C_{\zeta,\xi}\sum_{\tau=0}^{i}\beta^{i-\tau}\mathbb{E}[\|g_\theta^\tau\|^2] + \frac{6c}{1-\beta} + \eta_\theta^2\|g_\theta^i\|^2\right)\right]$$

$$\leq\lambda^{t+1}\mathbb{E}[\|g_\theta^0 - \nabla_\theta\mathcal{L}_0^D\|^2] + \mathbb{E}\left[\sum_{i=0}^{t}\eta_\theta^2 L^2\left(6C_{\zeta,\xi}\sum_{\tau=i}^{t}\lambda^{t-\tau+1}\beta^{\tau-i} + \lambda^{t-i+1}\right)\|g_\theta^i\|^2\right]$$
$$+ \frac{\alpha^2\sigma^2}{(1-\lambda)|B|} + \frac{6\lambda c L^2}{(1-\beta)(1-\lambda)} + \frac{6L^2 d^2\beta\lambda^{t+2}}{\lambda-\beta}$$

$$\leq\lambda^{t+1}\mathbb{E}[\|g_\theta^0 - \nabla_\theta\mathcal{L}_0^D\|^2] + \mathbb{E}\left[\sum_{i=0}^{t}\eta_\theta^2 L^2\lambda^{t-i+1}\left(6C_{\zeta,\xi}\frac{\lambda}{\lambda-\beta} + 1\right)\|g_\theta^i\|^2\right] + \frac{\alpha^2\sigma^2}{(1-\lambda)|B|}$$
$$+ \frac{6\lambda c L^2}{(1-\beta)(1-\lambda)} + \frac{6L^2 d^2\beta\lambda^{t+2}}{\lambda-\beta}$$

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 Assumption A; in the last but two inequality, we apply the summation formula of equal ratio sequence and use the fact that $0 < \alpha \leq 1, \beta \leq 1$; in the last step, we use our condition $\beta \leq (1-\alpha)^2/2$

Upper bound p_2 Next, we give an upper bound for p_2. From the Property 2 in Lemma 5.2, we know that

$$\Delta_{t+1}(\zeta_{t+1},\xi_{t+1}) = \mathbb{E}[\|\zeta_{t+1}-\zeta_{t+1}^*\|^2] + \mathbb{E}[\|\xi_{t+1}-\xi_{t+1}^*\|^2]$$

$$\leq\beta^{t+2}d^2 + \eta_\theta^2 C_{\zeta,\xi}\sum_{\tau=0}^{t}\beta^{t-\tau+1}\mathbb{E}[\|g_\theta^\tau\|^2] + \frac{c}{1-\beta}$$

As a result

$$p_2 = \mathbb{E}[\|\nabla_\theta\mathcal{L}_{t+1}^D - \nabla_\theta\mathcal{L}_{t+1}^{D*}\|^2] \leq 2L^2\mathbb{E}[\|\zeta_{t+1}-\zeta_{t+1}^*\|^2 + \|\xi_{t+1}-\xi_{t+1}^*\|^2]$$

43

$$\leq 2L^2 \Big(\beta^{t+2} d^2 + \eta_\theta^2 C_{\zeta,\xi} \sum_{\tau=0}^{t} \beta^{t-\tau+1} \mathbb{E}[\|g_\theta^\tau\|^2] + \frac{c}{1-\beta} \Big)$$

Combine these two results we can finish the proof:

$$\mathbb{E}[\|g_\theta^{t+1} - \nabla_\theta \mathcal{L}^{D*}(\theta_{t+1})\|^2] \leq 2p_1 + 2p_2$$

$$\leq 2\lambda^{t+1} \mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2] + 2\mathbb{E}\Big[\sum_{i=0}^{t} \eta_\theta^2 L^2 \lambda^{t-i+1} \Big(6C_{\zeta,\xi} \frac{\lambda}{\lambda - \beta} + 1 \Big) \|g_\theta^i\|^2 \Big]$$

$$+ 2\frac{\alpha^2 \sigma^2}{(1-\lambda)|B|} + \frac{12\lambda c L^2}{(1-\beta)(1-\lambda)} + \frac{12L^2 d^2 \beta \lambda^{t+2}}{\lambda - \beta} + 4L^2 \Big(\beta^{t+2} d^2 + \eta_\theta^2 C_{\zeta,\xi} \sum_{\tau=0}^{t} \beta^{t-\tau+1} \mathbb{E}[\|g_\theta^\tau\|^2] + \frac{c}{1-\beta} \Big)$$

$$\leq 2\lambda^{t+1} \mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2] + 2\frac{\alpha^2 \sigma^2}{(1-\lambda)|B|} + 4L^2 \Big(\beta^{t+2} d^2 + \frac{c}{1-\beta} + \frac{3\lambda c}{(1-\beta)(1-\lambda)} + \frac{3d^2 \beta \lambda^{t+2}}{\lambda - \beta} \Big)$$

$$+ 2\eta_\theta^2 L^2 \mathbb{E}\Big[\sum_{i=0}^{t} \Big(\lambda^{t-i+1} \Big(6C_{\zeta,\xi} \frac{\lambda}{\lambda - \beta} + 1 \Big) + 2C_{\zeta,\xi} \beta^{t-i+1} \Big) \|g_\theta^i\|^2 \Big]$$

$\square$

Proposition F.3. *Under Assumption A and B, define $L^{D*}(\theta) := \max_{\zeta \in Z} \min_{\xi \in \Xi} \mathcal{L}^D(\theta, \zeta, \xi)$ is $L_{\zeta,\xi}$ smooth with $L_{\zeta,\xi} := L(1 + \kappa_\zeta(\kappa_\xi + 1) + \kappa_\zeta(\kappa_\zeta + 1))$.*

Proof. For arbitrary $\theta_1, \theta_2 \in \Theta$, denote (ζ_1^*, ξ_1^*) and (ζ_2^*, ξ_2^*) as the saddle-point of $\arg\max_{\zeta \in Z} \min_{\xi \in \Xi} \mathcal{L}^D(\theta, \zeta, \xi)$ when $\theta = \theta_1$ and $\theta = \theta_2$, respectively, according to Lemma F.1 we have:

$$\|\nabla_\theta L^{D*}(\theta_1) - \nabla_\theta L^{D*}(\theta_2)\| \leq L(\|\theta_1 - \theta_2\| + \|\zeta_1 - \zeta_2\| + \|\xi_1 - \xi_2\|)$$
$$\leq L(1 + \kappa_\zeta(\kappa_\xi + 1) + \kappa_\zeta(\kappa_\zeta + 1))\|\theta_1 - \theta_2\|$$

$\square$

Theorem F.4. *Under Assumption A, B and C, given arbitrary ε, Algorithm 2 will return us a policy $\pi_{\widehat{\theta}}$, satisfying*

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

if the hyper-parameters in Alg. 2 satisfy the following constraints:

$$T = \lceil \max\{96, \frac{16\Delta L^{D*} L_{\zeta,\xi}}{\varepsilon^2}, \frac{16\Delta L^{D*} L \sqrt{28 C_{\zeta,\xi} + 2}}{\varepsilon^2}\} \rceil = O(\varepsilon^{-2}), \quad |B| = \lceil \frac{36\sigma^2}{\varepsilon^2} \rceil = O(\varepsilon^{-2})],$$

$$\alpha = 0.5, \quad \lambda = 2(1-\alpha)^2 = 0.5, \quad \beta = \min\{\frac{\varepsilon^2}{L^2 d^2}, (1-\lambda)^2, 1/2\}, \quad \eta_\theta = \min\{\frac{1}{2L_{\zeta,\xi}}, \sqrt{\frac{1-\lambda}{6L^2(14C_{\zeta,\xi} + 1)}}\}.$$

where $\Delta L^{D*} := \max_{\theta \in \Theta} \mathcal{L}^{D*}(\theta) - \mathcal{L}^{D*}(\theta_0)$, $C_{\zeta,\xi} = \kappa_\mu^2(\kappa_\xi + 1)^2 + \kappa_\xi^2(\kappa_\zeta + 1)^2$, and c_{oracle} is an independent constant.

Besides, the total gradient computation for O-SRM to obtain $\widehat{\theta}$ should be $O(\varepsilon^{-4})$, if:

(1) either we choose Algorithm 3 in Appendix E.2 as the ORACLE *and set the hyper-parameter $|N_{all}|$ in Algorithm 3 to be*

$$|N_{all}| = 480 \frac{d(2L^2 + 1)C_{LS}^2}{\varepsilon^2} = O(\varepsilon^{-2})$$

where C_{LS} is a constant defined in Appendix E.2,

(2) or we choose Algorithm 4 in Appendix E.3 as the ORACLE and set the hyper-parameters in Algorithm 4 to be:

$$|N| = \lceil \frac{240(L^2+1)\sigma^2}{\min\{\frac{\mu_\zeta \eta_\zeta}{4}, \frac{\mu_\xi \eta_\xi}{4}\}\varepsilon^2}(\frac{\eta_\zeta}{\mu_\zeta} + \frac{\eta_\xi}{\mu_\xi})\rceil, \quad K = c_{oracle}\log(\frac{1}{\beta}),$$

$$\eta_\zeta \leq \frac{1}{50\max\{\bar{L}_\zeta, \mu_\zeta\}}, \quad \eta_\xi \leq \frac{1}{50\max\{\bar{L}_\xi, \mu_\xi\}}.$$

where $\bar{L}_\zeta$ and $\bar{L}_\xi$ are defined in Condition B.4.

Proof. Recall the definition of $\mathcal{L}^{D*}$ in Prop F.3 and $\mathcal{L}^{D*}$ is $L + C_{\zeta,\xi}$-smooth.

$$L^{D*}(\theta_{T+1}) = L^{D*}(\theta_T + \eta_\theta g_\theta^T)$$

$$\geq L^{D*}(\theta_T) + \eta_\theta (g_\theta^T)^\top \nabla_\theta L^{D*}(\theta_T) - \frac{\eta_\theta^2 L_{\zeta,\xi}}{2}\|g_\theta^T\|^2$$

$$= L^{D*}(\theta_T) + \frac{\eta_\theta}{2}\|\nabla_\theta L^{D*}(\theta_T)\|^2 - \frac{\eta_\theta}{2}\|g_\theta^T - \nabla_\theta L^{D*}(\theta_T)\|^2 + (\frac{\eta_\theta}{2} - \frac{\eta_\theta^2 L_{\zeta,\xi}}{2})\|g_\theta^T\|^2$$

$$\geq L^{D*}(\theta_T) + \frac{\eta_\theta}{2}\|\nabla_\theta L^{D*}(\theta_T)\|^2 - \frac{\eta_\theta}{2}\|g_\theta^T - \nabla_\theta L^{D*}(\theta_T)\|^2 + \frac{\eta_\theta}{4}\|g_\theta^T\|^2$$

$$\geq L^{D*}(\theta_0) + \frac{\eta_\theta}{2}\sum_{t=0}^{T}\|\nabla_\theta L^{D*}(\theta_T)\|^2 - \frac{\eta_\theta}{2}\underbrace{\Big(\sum_{t=0}^{T}\|g_\theta^t - \nabla_\theta L^{D*}(\theta_T)\|^2 - \frac{1}{2}\|g_\theta^t\|^2\Big)}_{p}$$

where in the second equation, we use the fact that $(g_\theta^T)^\top \nabla_\theta L^{D*}(\theta_T) = \frac{1}{2}\|\nabla_\theta L^{D*}(\theta_T)\|^2 + \frac{1}{2}\|g_\theta^T\|^2 - \frac{1}{2}\|g_\theta^T - \nabla_\theta L^{D*}(\theta_T)\|^2$; in the second inequality, we add a constraint for η_θ that $\eta_\theta \leq \frac{1}{2L_{\zeta,\xi}}$;

Next, we give a upper bound for p with Lemma F.2:

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

$$\leq \sum_{t=0}^{T}\Big\{2\lambda^{t+1}\mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2] + 2\frac{\alpha^2\sigma^2}{(1-\lambda)|B|} + 4L^2\Big(\beta^{t+2}d^2 + \frac{c}{1-\beta} + \frac{3\lambda c}{(1-\beta)(1-\lambda)} + \frac{3d^2\beta\lambda^{t+2}}{\lambda-\beta}\Big)$$

$$+ 2\eta_\theta^2 L^2 \mathbb{E}\Big[\sum_{i=0}^{t}\Big(\lambda^{t-i+1}\Big(6C_{\zeta,\xi}\frac{\lambda}{\lambda-\beta}+1\Big) + 2C_{\zeta,\xi}\beta^{t-i+1}\Big)\|g_\theta^i\|^2\Big] - \frac{1}{2}\mathbb{E}[\|g_\theta^t\|^2]\Big\}$$

$$\leq 2(T+1)\frac{\alpha^2\sigma^2}{(1-\lambda)|B|} + 4(T+1)cL^2(\frac{1}{1-\beta} + \frac{3\lambda}{(1-\beta)(1-\lambda)}) + \frac{2(T+1)}{1-\lambda}\mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2]$$

$$+ 4L^2 d^2(\frac{\beta^2}{1-\beta} + \frac{3\beta\lambda^2}{(1-\lambda)(\lambda-\beta)}) + \sum_{t=0}^{T}\mathbb{E}[\|g_\theta^t\|^2]\Big\{-\frac{1}{2} + 2\eta_\theta^2 L^2 \sum_{i=0}^{T-t+1}\mathbb{E}\Big[\lambda^i\Big(6C_{\zeta,\xi}\frac{\lambda}{\lambda-\beta}+1\Big) + 2C_{\zeta,\xi}\beta^i\Big]\Big\}$$

$$\leq 2(T+1)\frac{\alpha^2\sigma^2}{(1-\lambda)|B|} + 4(T+1)cL^2(\frac{1}{1-\beta} + \frac{3\lambda}{(1-\beta)(1-\lambda)}) + \frac{2(T+1)}{1-\lambda}\mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2]$$

$$+ 4L^2 d^2(\frac{\beta^2}{1-\beta} + \frac{3\beta\lambda^2}{(1-\lambda)(\lambda-\beta)}) + \sum_{t=0}^{T}\mathbb{E}[\|g_\theta^t\|^2]\Big\{-\frac{1}{2} + 2\eta_\theta^2 L^2 \Big(\frac{1}{1-\lambda}\Big(6C_{\zeta,\xi}\frac{\lambda}{\lambda-\beta}+1\Big) + 2C_{\zeta,\xi}\frac{1}{1-\beta}\Big)\Big\}$$

$$\leq 2(T+1)\frac{\alpha^2\sigma^2}{(1-\lambda)|B|} + 4(T+1)cL^2(\frac{1}{1-\beta} + \frac{3\lambda}{(1-\beta)(1-\lambda)}) + \frac{2(T+1)}{1-\lambda}\mathbb{E}[\|g_\theta^0 - \nabla_\theta \mathcal{L}_0^D\|^2]$$

$$+ 4L^2 d^2(\frac{\beta^2}{1-\beta} + \frac{6\beta\lambda}{1-\lambda}) + \sum_{t=0}^{T}\mathbb{E}[\|g_\theta^t\|^2]\Big\{-\frac{1}{2} + \frac{2\eta_\theta^2 L^2}{1-\lambda}\Big(14C_{\zeta,\xi}+1\Big)\Big\} \qquad (\beta \leq (1-\alpha)^2 = \frac{\lambda}{2} \leq \lambda)$$

$$\leq 2(T+1)\frac{\alpha^2\sigma^2}{(1-\lambda)|B|} + 4(T+1)cL^2(\frac{1}{1-\beta} + \frac{3\lambda}{(1-\beta)(1-\lambda)})) + \frac{2(T+1)}{1-\lambda}\mathbb{E}[\|g_\theta^0 - \nabla_\theta\mathcal{L}_0^D\|^2]$$
$$+ 4L^2d^2(\frac{\beta^2}{1-\beta} + \frac{6\beta\lambda}{1-\lambda}).$$

In the fourth inequality, we add the following constraint to drop the terms containing $\|g_\theta\|$:

$$\eta_\theta \leq \sqrt{\frac{1-\lambda}{4L^2(14C_{\zeta,\xi}+1)}}. \tag{28}$$

Therefore,

$$\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla_\theta\mathcal{L}^{D*}(\theta_\tau)\|^2 \leq \frac{2}{(T+1)\eta_\theta}(\mathcal{L}^{D*}(\theta_T) - \mathcal{L}^{D*}(\theta_0)) + \frac{1}{T+1}\sum_{\tau=0}^{T}\left(\|g_\theta^\tau - \nabla_\theta\mathcal{L}^{D*}(\theta_\tau)\|^2 - \frac{1}{2}\|g_\theta^\tau\|^2\right)$$

$$\leq 4cL^2(\frac{1}{1-\beta} + \frac{3\lambda}{(1-\beta)(1-\lambda)})) + \frac{2\Delta\mathcal{L}^{D*}}{(T+1)\eta_\theta}$$

$$+ \frac{2}{1-\lambda}\mathbb{E}[\|g_\theta^0 - \nabla_\theta\mathcal{L}_0^D\|^2] + 2\frac{\alpha^2\sigma^2}{(1-\lambda)|B|} + 4L^2d^2(\frac{\beta^2}{1-\beta} + \frac{6\beta\lambda}{1-\lambda})$$

$$\leq \underbrace{4cL^2\frac{(6\lambda+1)}{1-\lambda}}_{p_0} + \underbrace{\frac{2\Delta\mathcal{L}^{D*}}{(T+1)\eta_\theta}}_{p_1} + \underbrace{\frac{2(1+\alpha^2)}{1-\lambda}\frac{\sigma^2}{|B|}}_{p_2} + \underbrace{24L^2d^2\frac{\beta\lambda}{(T+1)(1-\lambda)}}_{p_3}.$$

$$(\beta \leq 1/2, \beta \leq \lambda, \mathbb{E}[g_\theta^0] = \mathcal{L}_0^D)$$

The choice of α In order to guarantee that $\lambda = 2(1-\alpha)^2 < 1$ and $\alpha < 1$, we first need $\alpha \in (1 - \frac{\sqrt{2}}{2}, 1)$. Easy to observe that, the coefficient of p_2:

$$\frac{1+\alpha^2}{1-\lambda} = \frac{1+\alpha^2}{1-2(1-\alpha)^2} = \frac{1/2+2\alpha}{-1+4\alpha-2\alpha^2} - 1/2.$$

By taking the derivative w.r.t. α, we have:

$$(\frac{1+\alpha^2}{1-2(1-\alpha)^2})' = \frac{2(-1+4\alpha-2\alpha^2)-(1/2+2\alpha)(4-4\alpha)}{(-1+4\alpha-2\alpha^2)^2} = \frac{4\alpha^2+2\alpha-4}{(-1+4\alpha-2\alpha^2)^2}.$$

We can observe easily that the zero point of the first-order derivative should located in interval $(\frac{1}{2}, 1)$. Besides, p_0, p_1, p_3 are decreasing as α approaching 1 (i.e. λ approaching 0). Therefore, the optimal choice of α should locates in $(\frac{1}{2}, 1)$.

As we can see, the introduction of α is to balancing the bias of momentum term and the variance in the current step to provide a better estimation of the gradient in current step. In the following, for simplicity, we directly choose $\alpha = 1/2$.

Next, we want to carefully choose other hyper-parameters to make sure $p_0, p_1, p_2, p_3 \leq \varepsilon^2/4$. We consider $\beta \leq \min\{\frac{\varepsilon^2}{L^2d^2}, (1-\lambda)^2, 1/2\}$.

Control p_0 Since $\beta < 1/2$, we know

$$p_0 \leq 32cL^2.$$

To make sure $p_0 \leq \varepsilon^2/4$, we need

$$c \leq \frac{\varepsilon^2}{128L^2}.$$

If we choose Algorithm 3 as ORACLE, the hypyer-parameter $|N|$ (total sample size) in Algorithm 3 should be set to:

$$|N| \geq 20d\frac{C_{LS}^2}{c} = O(\varepsilon^{-2})$$

If we choose Algorithm 2, the hyper-parameter $|N|$ (mini batch size) and K (total iteration) in Algorithm 3 should be set to:

$$\frac{8\sigma^2}{\min\{\frac{\mu_\zeta \eta_\zeta}{4}, \frac{\mu_\xi \eta_\xi}{4}\}|N|}(\frac{\eta_\zeta}{\mu_\zeta} + \frac{\eta_\xi}{\mu_\xi}) \leq \frac{\varepsilon^2}{128L^2}, \quad K = c_{oracle}\log(1/\beta) = O(\log\frac{1}{\varepsilon}).$$

where c_{oracle} is an independent constant. Therefore,

$$|N| \geq \frac{240(L^2+1)\sigma^2}{\min\{\frac{\mu_\zeta \eta_\zeta}{4}, \frac{\mu_\xi \eta_\xi}{4}\}\varepsilon^2}(\frac{\eta_\zeta}{\mu_\zeta} + \frac{\eta_\xi}{\mu_\xi}) = O(\varepsilon^{-2}).$$

Control p_1 Since we have two constrains on η_θ: $\eta_\theta \leq \frac{1}{2L_{\zeta,\xi}}$ and $\eta_\theta \leq \sqrt{\frac{1-\lambda}{4L^2(14C_{\zeta,\xi}+1)}}$. To make sure $p_1 \leq \frac{\varepsilon^2}{4}$, we need,

$$\frac{2\Delta L^{D*}}{(T+1)}\max\{\sqrt{\frac{4L^2(14C_{\zeta,\xi}+1)}{1-\lambda}}, 2L_{\zeta,\xi}\} \leq \frac{\varepsilon^2}{4}.$$

Therefore, we requirement

$$T \geq \max\{\frac{16\Delta L^{D*}L_{\zeta,\xi}}{\varepsilon^2}, \frac{16\Delta L^{D*}L\sqrt{28C_{\zeta,\xi}+2}}{\varepsilon^2}\}. \tag{29}$$

Control p_2 Recall that we choose $\alpha = 1/2$, we need:

$$p_2 = \frac{2(1+\alpha)^2}{1-\lambda}\frac{\sigma^2}{|B|} = 9\frac{\sigma^2}{|B|} \leq \frac{\varepsilon^2}{4}.$$

which requires:

$$|B| \geq \frac{36\sigma^2}{\varepsilon^2} \tag{30}$$

Control p_3 Because K (iteration number of *Oracle* Algorithm) only depends on $\log\frac{1}{\beta}$, we choose $\beta \leq \frac{\varepsilon^2}{L^2d^2}$. Then, as long as $T \geq 168$, we have:

$$p_3 = 24L^2d^2\frac{\beta\lambda}{(T+1)(1-\lambda)} = \frac{24\beta L^2d^2}{T+1} \leq \frac{\varepsilon^2}{4} \tag{31}$$

Combine (29)-(31), we need

$$T \geq \max\{96, \frac{16\Delta L^{D*}L_{\zeta,\xi}}{\varepsilon^2}, \frac{16\Delta L^{D*}L\sqrt{28C_{\zeta,\xi}+2}}{\varepsilon^2}\}, \quad |B| \geq \frac{36\sigma^2}{\varepsilon^2};$$

As a result, if we use Algorithm 3 as ORACLE, the total computation before obtaining θ_{out} should be

$$|B| \cdot (T+1) + |N| \cdot T = O(\varepsilon^{-2}) + O(\varepsilon^{-4}) + O(\varepsilon^{-2}) \cdot O(\varepsilon^{-2}) = O(\varepsilon^{-4})$$

and if we use Algorithm 4 as ORACLE, it should be

$$|B| \cdot (T+1) + |N| \cdot K \cdot T = O(\varepsilon^{-2}) + O(\varepsilon^{-4}) + O(\varepsilon^{-2}) \cdot O(\log\frac{1}{\varepsilon}) \cdot O(\varepsilon^{-2}) = O(\varepsilon^{-4})$$

where we omit log terms in $O(\cdot)$.

Guarantee of $\mathbb{E}[\|\nabla_\theta J(\pi_{\widehat{\theta}})\|]$ By applying the hyper-parameters discussed above, we have:

$$
\begin{aligned}
\mathbb{E}[\|\nabla_\theta J(\pi_{\widehat{\theta}})\|] \leq & \mathbb{E}[\|\nabla_\theta \mathcal{L}^{D*}(\widehat{\theta})\|] + \varepsilon_{reg} + \varepsilon_{func} + \varepsilon_{data} \\
\leq & \sqrt{\mathbb{E}[\|\nabla_\theta \mathcal{L}^{D*}(\widehat{\theta})\|^2]} + \varepsilon_{reg} + \varepsilon_{func} + \varepsilon_{data} \\
= & \sqrt{\frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}[\|\nabla_\theta \mathcal{L}^{D*}(\theta_t)\|^2]} + \varepsilon_{reg} + \varepsilon_{func} + \varepsilon_{data} \\
\leq & \varepsilon + \varepsilon_{reg} + \varepsilon_{func} + \varepsilon_{data}
\end{aligned}
$$

$\square$