
A Safe Exploration Approach to Constrained Markov Decision Processes

Tingting Ni
Sycamore, EPFL Lausanne

Maryam Kamgarpour
Sycamore, EPFL Lausanne

Abstract

We consider discounted infinite-horizon constrained Markov decision processes (CMDPs), where the goal is to find an optimal policy that maximizes the expected cumulative reward while satisfying expected cumulative constraints. Motivated by the application of CMDPs in online learning for safety-critical systems, we focus on developing a model-free and *simulator-free* algorithm that ensures *constraint satisfaction during learning*. To this end, we employ the LB-SGD algorithm proposed in (Usmanova et al., 2024), which utilizes an interior-point approach based on the log-barrier function of the CMDP. Under the commonly assumed conditions of relaxed Fisher non-degeneracy and bounded transfer error in policy parameterization, we establish the theoretical properties of the LB-SGD algorithm. In particular, unlike existing CMDP approaches that ensure policy feasibility only upon convergence, the LB-SGD algorithm guarantees feasibility throughout the learning process and converges to the ε -optimal policy with a sample complexity of $\mathcal{O}(\varepsilon^{-6})$. Compared to the state-of-the-art policy gradient-based algorithm, C-NPG-PDA (Bai et al., 2023), the LB-SGD algorithm requires an additional $\mathcal{O}(\varepsilon^{-2})$ samples to ensure policy feasibility during learning with the same Fisher non-degenerate parameterization.

1 INTRODUCTION

Reinforcement learning (RL) involves studying sequential decision-making problems, where an agent

aims to maximize an expected cumulative reward by interacting with an unknown environment (Sutton and Barto, 2018). While RL has achieved impressive success in domains like video games and board games (Berner et al., 2019; Silver et al., 2016, 2017), safety concerns arise when applying RL to real-world problems, such as autonomous driving (Fazel et al., 2018), robotics (Koppejan and Whiteson, 2011; Ono et al., 2015), and cyber-security (Zhang et al., 2019). Incorporating safety into RL algorithms can be done in various ways (Garcia and Fernández, 2015). From a problem formulation perspective, one natural approach to incorporate safety constraints is through the framework of discounted infinite horizon constrained Markov decision processes (CMDPs).

In a CMDP, the agent aims to maximize an expected cumulative reward subject to expected cumulative constraints. The CMDP formulation has a long history (Altman, 1999; Puterman, 2014) and has been applied in several realistic scenarios (Kalweit et al., 2020; Mirchevska et al., 2018; Zang et al., 2020). Due to its applicability, there has been a growing body of literature in recent years that develops learning-based algorithms for CMDPs, employing both model-free (Bai et al., 2023, 2022; Ding et al., 2020, 2022a; Liu et al., 2022; Xu et al., 2021; Zeng et al., 2022) and model-based approaches (Agarwal et al., 2022; Hasan-zadeZonuzy et al., 2021; Jayant and Bhatnagar, 2022).

Existing learning-based approaches to the CMDP problem offer various theoretical guarantees regarding constraint violations. Some of these approaches only ensure constraint satisfaction upon algorithm convergence (Ding et al., 2020, 2022a; Liu et al., 2022; Xu et al., 2021; Zeng et al., 2022), bounding the average constraint violation by ε . Others enhance these guarantees by aiming for averaged zero constraint violation (Bai et al., 2023; Kalagarla et al., 2023; Wei et al., 2022a,b).

For the practical deployment of RL algorithms in real-world scenarios, particularly those requiring online tuning, it is important to satisfy the constraints during the learning process (Abe et al., 2010). This

property is referred to as *safe exploration* (Koller et al., 2019). Ensuring constraint satisfaction during learning is challenging, as it limits exploration and also requires a more accurate estimation of model parameters or gradients (Vaswani et al., 2022).

To address safe exploration, model-based methods employ either Gaussian processes to learn system dynamics (Koller et al., 2019; Cheng et al., 2019; Wachi et al., 2018; Berkenkamp et al., 2017; Fisac et al., 2018) or leverage Lyapunov-based analysis (Chow et al., 2018, 2019) to ensure safe exploration with high probability. However, these approaches lack guarantees on the performance of the learned policy. An alternative model-based approach, known as the constrained upper confidence RL, offers convergence guarantees and ensures safe exploration with high probability. This algorithm has been applied in both infinite horizon average reward scenarios with known transition dynamics (Zheng and Ratliff, 2020) and finite horizon reward scenarios with unknown transition dynamics (Liu et al., 2021a; Bura et al., 2022). However, in complex environments, accurately modeling system dynamics can be computationally challenging (Sutton and Barto, 2018).

Policy gradient (PG) algorithms demonstrate their advantage in handling complex environments in a model-free manner (Agarwal et al., 2021). They have shown empirical success in solving CMDPs (Liang et al., 2018; Achiam et al., 2017; Tessler et al., 2019; Liu et al., 2020a). Initial guarantees for safe exploration in CMDPs were provided by (Achiam et al., 2017), relying on exact policy gradient information. However, with unknown transition dynamics, we can only estimate the gradient. To address this, several works (Ding et al., 2020, 2022a; Xu et al., 2021; Ding et al., 2024) assume access to a simulator (Koenig and Simmons, 1993) (also known as a generative model (Azar et al., 2012)), which provides information about the reward and constraint values associated with any state and action of the CMDP. However, practical RL requires learning in real-world scenarios, where access to a simulator may not be feasible. Theoretically, the analysis becomes significantly more challenging without a simulator (Jin et al., 2018).

Among the above works addressing CMDPs with simulator access, (Ding et al., 2020, 2022a; Xu et al., 2021) provided theoretical guarantees on bounding the average constraint violation over the iterates by ε . However, ensuring an averaged ε constraint violation is problematic in safety-critical CMDPs, as constraint values may overshoot in each iteration (Stooke et al., 2020; Calvo-Fullana et al., 2023), thereby failing to provide safety guarantees for each policy iterate. To partially mitigate this issue, (Ding et al., 2024) pro-

posed an approach that ensures constraint satisfaction and optimality only for the last iterate policy. Without simulator access, (Zeng et al., 2022) provided theoretical guarantees on averaged ε constraint violation, whereas (Bai et al., 2023) strengthened this guarantee to an averaged zero constraint violation, albeit at the cost of requiring an additional $\mathcal{O}(\varepsilon^{-2})$ samples compared to the state-of-the-art policy gradient-based algorithm (Ding et al., 2020). However, like all the aforementioned works, these approaches do not ensure constraint satisfaction during learning. A summary of these works, their constraint satisfaction and convergence guarantees are provided in Table 1.

In the field of constrained optimization, safe exploration has been extensively studied using a Bayesian approach based on Gaussian processes (Berkenkamp et al., 2017; Sui et al., 2015; Berkenkamp et al., 2021; Amani et al., 2019). However, Bayesian optimization algorithms suffer from the curse of dimensionality (Frazier, 2018; Moriconi et al., 2020; Eriksson and Jankowiak, 2021), and this challenges their applicability to model-free RL settings, where large state and action spaces are often encountered. To address this limitation, (Usmanova et al., 2020, 2024) proposed a first-order interior point approach, inspired by (Hinder and Ye, 2019), that tackles the issue by incorporating constraints into the objective using a log barrier function. While their vanilla non-convex analysis could directly apply to policy gradient-based algorithms, the convergence result would be limited to ε -approximate stationary points rather than optimal points using $\mathcal{O}(\varepsilon^{-7})$ samples in total. While both (Usmanova et al., 2024) and (Liu et al., 2020a) demonstrated the success of the log barrier approach on benchmark continuous control problems, to our knowledge, safe exploration and tight convergence guarantees for the log barrier policy gradient method in a CMDP had not been proven.

Our paper is dedicated to providing provable non-asymptotic convergence guarantees for solving CMDPs while ensuring safe exploration under a simulator-free setting. Our contributions are as follows.

1.1 Contributions

- We employ the LB-SGD algorithm proposed in (Usmanova et al., 2024) and develop an interior-point stochastic policy gradient approach for the CMDP problem. Further, we prove that the last iterate policy is $\mathcal{O}(\sqrt{\varepsilon_{bias}}) + \tilde{\mathcal{O}}(\varepsilon)$ optimal,¹ while ensuring safe exploration with high probability, utilizing $\tilde{\mathcal{O}}(\varepsilon^{-6})$ samples. The term ε_{bias} represents the function approximation error resulting from the policy class (see Theorem 4.12).

¹The notation $\tilde{\mathcal{O}}(\cdot)$ hides the $\log(\frac{1}{\varepsilon})$ term.

Table 1: Sample complexity for achieving ε -optimal objectives with guarantees on constraint violations in stochastic policy gradient-based algorithms, considering various parameterizations for discounted infinite horizon CMDPs. For constraint violation, Averaged $\mathcal{O}(\varepsilon)$ refers to the average constraint violation being bounded by ε . Averaged zero strengthens the above bound to 0. RPG-PD (Ding et al., 2024) ensures the last iterate’s safety. Here, we refer to “w.h.p” as with high probability.

Stochastic policy gradient-based algorithms					
Parameterization	Algorithm	Sample complexity	Constraint violation	Optimality	Generative model
Softmax	NPG-PD Ding et al. (2020)	$\mathcal{O}(\varepsilon^{-2})$	Averaged $\mathcal{O}(\varepsilon)$	Average	✓
Softmax	PD-NAC (Zeng et al., 2022)	$\mathcal{O}(\varepsilon^{-6})$	Averaged $\mathcal{O}(\varepsilon)$	Average	×
Neural softmax(ReLU)	CRPO (Xu et al., 2021)	$\mathcal{O}(\varepsilon^{-6})$	Averaged $\mathcal{O}(\varepsilon)$	Average	✓
Log-linear	RPG-PD (Ding et al., 2024)	$\tilde{\mathcal{O}}(\varepsilon^{-14})$	0 at last iterate	Last iterate	✓
Fisher non-degenerate	PD-ANPG (Mondal and Aggarwal, 2024)	$\tilde{\mathcal{O}}(\varepsilon^{-3})$	Averaged $\mathcal{O}(\varepsilon)$	Average	✓
Fisher non-degenerate	C-NPG-PDA (Bai et al., 2023)	$\tilde{\mathcal{O}}(\varepsilon^{-4})$	Averaged zero	Average	×
Relaxed Fisher non-degenerate	LB-SGD (Usmanova et al., 2024)	$\tilde{\mathcal{O}}(\varepsilon^{-6})$	Safe exploration w.h.p	Last iterate	×

- On the technical aspect, we construct an accurate gradient estimator for the log barrier and establish its local smoothness properties. Furthermore, by incorporating common assumptions on the policy class, including relaxed Fisher non-degeneracy and bounded transfer error (Liu et al., 2020b; Yuan et al., 2022; Fatkhullin et al., 2023; Ding et al., 2022b), we establish the gradient dominance property for the *log barrier function* (see Lemma 4.10). This in turn enables us to derive convergence guarantees for the last iterate.
- We contribute to the understanding of the (relaxed) Fisher non-degeneracy and the bounded transfer error assumptions by narrowing down the classes of policies that satisfy these assumptions (see Facts 4.6, 4.9).

1.2 Notations

For a set \mathcal{X} , $\Delta(\mathcal{X})$ denotes the probability simplex over \mathcal{X} , while $|\mathcal{X}|$ and $\text{Cl}\{\mathcal{X}\}$ denote its cardinality and closure, respectively. For any integer m , we set $[m] := \{1, \dots, m\}$. Here, $\|\cdot\|$ denotes the Euclidean ℓ_2 -norm for vectors and the operator norm for matrices, respectively. The notation $A \succeq B$ indicates that the matrix $A - B$ is positive semi-definite. We denote the image space and kernel space of the matrix A as $\text{Im}(A)$ and $\text{Ker}(A)$, respectively. The function $f(x)$ is said to be M -smooth on \mathcal{X} if the inequality $f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{M}{2} \|x - y\|^2$ holds $\forall x, y \in \mathcal{X}$, and L -Lipschitz continuous on \mathcal{X} if $|f(x) - f(y)| \leq L \|x - y\|$ holds $\forall x, y \in \mathcal{X}$.

2 PROBLEM FORMULATION

We consider an infinite-horizon discounted constrained Markov decision process (CMDP) defined by the tuple $\{\mathcal{S}, \mathcal{A}, P, \rho, \{r_i\}_{i=0}^m, \gamma\}$. Here, \mathcal{S} and \mathcal{A} are the state and action spaces, respectively, $\rho \in \Delta(\mathcal{S})$ denotes the initial state distribution, and $P(s'|s, a)$ is the probability of

transitioning from state s to state s' under action a . Additionally, $r_0 : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ is the reward function, and $r_i : \mathcal{S} \times \mathcal{A} \rightarrow [-1, 1]$ is the utility function for $i \in [m]$. We denote the discount factor as $\gamma \in (0, 1)$.

We consider a stationary stochastic policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$, which maps states to probability distributions over actions, and we denote Π as the set containing all stationary stochastic policies. We introduce the performance measure $V_i^\pi(\rho) := \mathbb{E}_{\tau \sim \pi} [\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t)]$, which is the infinite horizon discounted total return concerning the function r_i . Here, τ denotes a trajectory $\{(s_0, a_0, s_1, a_1, \dots) : s_h \in \mathcal{S}, a_h \in \mathcal{A}, h \in \mathbb{N}\}$ induced by the initial distribution $s_0 \sim \rho$, the policy $a_t \sim \pi(\cdot|s_t)$, and the transition dynamics $s_{t+1} \sim P(\cdot|s_t, a_t)$.

In a CMDP, the objective is to find a policy that maximizes the objective function $V_0^\pi(\rho)$ subject to the constraints $V_i^\pi(\rho)$ for $i \in [m]$:

$$\max_{\pi} V_0^\pi(\rho) \quad \text{s.t.} \quad V_i^\pi(\rho) \geq 0, \quad i \in [m]. \quad (\text{RL-O})$$

The choice of optimizing only over stationary policies is justified: it has been shown that the set of all optimal policies for a CMDP includes stationary policies (Altman, 1999). We assume the existence of a stationary optimal policy π^* for problem (RL-O), ensured by Slater’s condition for the finite action space (Altman, 1999) and by suitable measurability and compactness assumptions for the continuous action space, see (Hernández-Lerma and González-Hernández, 2000, Theorem 3.2).

For large or continuous CMDPs, solving (RL-O) is intractable due to the curse of dimensionality (Sutton et al., 1999). The policy gradient method allows us to search for the optimal policy π^* within a parameterized policy set $\{\pi_\theta, \theta \in \mathbb{R}^d\}$. For example, we can apply neural softmax parametrization for discrete action space, or Gaussian parameterization for continuous action space. For simplicity, we denote $V_i^{\pi_\theta}(\rho)$ as $V_i^\theta(\rho)$, as it is a function of θ . Due to the policy parameterization, we can reformulate problem (RL-O)

into a constrained optimization problem over the finite-dimensional parameter space, as follows:

Problem 2.1. Consider $\theta \in \mathbb{R}^d$, and we are solving the following optimization problem:

$$\max_{\theta} V_0^{\theta}(\rho) \quad \text{s.t.} \quad V_i^{\theta}(\rho) \geq 0, \quad i \in [m]. \quad (\text{RL-P})$$

Here, the feasible set is denoted as $\Theta := \{\theta \mid V_i^{\theta}(\rho) \geq 0, i \in [m]\}$, and the corresponding feasible parameterized policy set is $\Pi_{\Theta} := \{\pi_{\theta} \mid \theta \in \Theta\}$.

Due to the parameterization, our parameterized policy set may not cover the entire stochastic policy set. Our goal is to find a policy π_{θ} that closely approximates the optimal policy π^* while ensuring safe exploration, as defined below.

Definition 2.2. Given an algorithm providing a sequence $\{\theta_t\}_{t=0}^T$, we say it ensures safe exploration if $\forall t \in \{0, \dots, T\}$, we have $\min_{i \in [m]} V_i^{\theta_t}(\rho) \geq 0$.

3 LOG BARRIER POLICY GRADIENT APPROACH

Our approach to safe exploration is based on maximizing the unconstrained log barrier surrogate of Problem (RL-P), defined as

$$B_{\eta}^{\theta}(\rho) := V_0^{\theta}(\rho) + \eta \sum_{i=1}^m \log V_i^{\theta}(\rho)$$

where $\eta > 0$. The log barrier algorithm (Liu et al., 2020a; Usmanova et al., 2024) can be summarized as

$$\theta_{t+1} = \theta_t + \gamma_t \hat{\nabla}_{\theta} B_{\eta}^{\theta}(\rho), \quad (1)$$

where γ_t is the stepsize, and $\hat{\nabla}_{\theta} B_{\eta}^{\theta}(\rho)$ is an estimate of the true gradient $\nabla_{\theta} B_{\eta}^{\theta}(\rho)$, computed as

$$\nabla_{\theta} B_{\eta}^{\theta}(\rho) = \nabla V_0^{\theta}(\rho) + \eta \sum_{i=1}^m \frac{\nabla V_i^{\theta}(\rho)}{V_i^{\theta}(\rho)}. \quad (2)$$

The intuition is that the iterates approach the stationary point of the log barrier function from the interior, thereby ensuring safe exploration. Furthermore, we establish that the stationary points of the log barrier function correspond to approximately optimal points of the CMDP objective, ensuring optimality.

The iteration above is arguably simple; In contrast to the approach in (Bai et al., 2023; Ding et al., 2022a; Liu et al., 2022; Xu et al., 2021; Zeng et al., 2022), our method eliminates the need for projection, adjustment of the learning rate for a dual variable, and the requirement of a simulator capable of simulating the MDP from any state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$. However, the challenge lies in fine-tuning the stepsize γ_t to ensure safe exploration while maintaining convergence. Next, we address these aspects and formalize the algorithm.

3.1 Estimating the Log Barrier Gradient

Given that we do not have access to the generative model, we need to estimate the log barrier gradient to implement (1), which by Equation (2) it implies that we need to estimate both $V_i^{\theta}(\rho)$ and its gradient $\nabla_{\theta} V_i^{\theta}(\rho)$. Gradient estimators using Monte Carlo approaches have been addressed in past work (Sutton et al., 1999; Williams, 1992; Baxter and Bartlett, 2001), and we will apply them to estimate the log barrier gradient as follows.

We sample n truncated trajectories with a fixed horizon H , where each trajectory is denoted as $\tau_j := (s_t^j, a_t^j, \{r_i(s_t^j, a_t^j)\}_{i=0}^m)_{t=0}^{H-1}$, including the corresponding reward and utility functions. The estimator of the value function, denoted as $\hat{V}_i^{\theta}(\rho)$, is computed as the average value over the sampled trajectories:

$$\hat{V}_i^{\theta}(\rho) := \frac{1}{n} \sum_{j=1}^n \sum_{t=0}^{H-1} \gamma^t r_i(s_t^j, a_t^j).$$

To estimate the gradient $\nabla_{\theta} V_i^{\theta}(\rho)$, we consider the so-called GPOMDP gradient estimator, computed as:

$$\hat{\nabla}_{\theta} V_i^{\theta}(\rho) := \frac{1}{n} \sum_{j=1}^n \sum_{t=0}^{H-1} \sum_{t'=0}^t \gamma^{t'} r_i(s_{t'}^j, a_{t'}^j) \nabla_{\theta} \log \pi_{\theta}(a_{t'}^j | s_{t'}^j).$$

Note that our work extends to other gradient estimators such as REINFORCE for which error bounds similar to those in Proposition 3.3 below can be derived.

Bounds on the error in the above estimations can be derived given the following assumption.

Assumption 3.1. The gradient and Hessian of the function $\log \pi_{\theta}(a|s)$ are bounded, i.e., there exist constants $M_g, M_h > 0$ such that $\|\nabla_{\theta} \log \pi_{\theta}(a|s)\| \leq M_g$ and $\|\nabla_{\theta}^2 \log \pi_{\theta}(a|s)\| \leq M_h$ for all $\theta \in \Theta$.

Remark 3.2. Assumption 3.1 has been widely utilized in the analysis of policy gradient methods (Liu et al., 2020b; Yuan et al., 2022; Ding et al., 2022b; Xu et al., 2020). It is satisfied for softmax policy, the log-linear policy with bounded feature vectors (Agarwal et al., 2021, Section 6.1.1), as well as Cauchy policy (Fatkhullin et al., 2023, Appendix B) and Gaussian policy with action clipping (Xu et al., 2020, Appendix D).

As will be shown in Proposition 3.3 below, Assumption 3.1 ensures smoothness and Lipschitz continuity properties for the value functions $V_i^{\theta}(\rho)$. It also ensures that the estimators of $V_i^{\theta}(\rho)$ and its gradient $\nabla_{\theta} V_i^{\theta}(\rho)$ have sub-Gaussian tail bounds. These tail bounds provide probabilistic guarantees that the estimators do not deviate significantly from their expected values. Such a result is needed for ensuring safe exploration and convergence of the algorithm.

Proposition 3.3. *Let Assumption 3.1 hold. The following properties hold $\forall i \in \{0, \dots, m\}$ and $\forall \theta \in \Theta$.*

1. $V_i^\theta(\rho)$ are L -Lipschitz continuous and M -smooth, where $L := \frac{M_g}{(1-\gamma)^2}$ and $M := \frac{M_g^2 + M_h}{(1-\gamma)^2}$.

2. Let $b^0(H) := \frac{\gamma^H}{1-\gamma}$ and $b^1(H) := \frac{M_g \gamma^H}{1-\gamma} \sqrt{\frac{1}{1-\gamma}} + H$, we have

$$\begin{aligned} \left| V_i^\theta(\rho) - \mathbb{E} [\hat{V}_i^\theta(\rho)] \right| &\leq b^0(H), \\ \left\| \nabla_\theta V_i^\theta(\rho) - \mathbb{E} [\hat{\nabla}_\theta V_i^\theta(\rho)] \right\| &\leq b^1(H). \end{aligned}$$

3. Let $\sigma^0(n) := \frac{\sqrt{2}}{\sqrt{n}(1-\gamma)}$ and $\sigma^1(n) := \frac{2\sqrt{2}M_g}{\sqrt{n}(1-\gamma)^{\frac{3}{2}}}$. For any $\delta \in (0, 1)$, we have

$$\mathbb{P} \left(\left| \hat{V}_i^\theta(\rho) - \mathbb{E} [\hat{V}_i^\theta(\rho)] \right| \leq \sigma^0(n) \sqrt{\ln \frac{2}{\delta}} \right) \geq 1 - \delta.$$

Furthermore, if $n \geq 8 \ln \frac{1}{\delta}$, we have

$$\mathbb{P} \left(\left\| \hat{\nabla}_\theta V_i^\theta(\rho) - \mathbb{E} [\hat{\nabla}_\theta V_i^\theta(\rho)] \right\| \leq \sigma^1(n) \sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}} \right) \geq 1 - \delta.$$

The first property has been proven in (Bai et al., 2023, Lemma 2) and (Yuan et al., 2022, Lemma 4.4). The bias bound $b^1(H)$ has been established in (Yuan et al., 2022, Lemma 4.5). We prove the remaining properties of the concentration bounds for the value estimators using Hoeffding's inequality and for the gradient estimators using the vector Bernstein inequality (see Appendix D.1). We note that (Papini et al., 2022) also proposed tail bounds for gradient estimators; however, we emphasize that our bound is dimension-free due to the use of the vector Bernstein inequality.

Based on the estimators of the value function $V_i^\theta(\rho)$ and its gradient $\nabla_\theta V_i^\theta(\rho)$, we construct the estimator for the log barrier gradient required in iteration (1) as follows:

$$\hat{\nabla}_\theta B_\eta^\theta(\rho) := \hat{\nabla}_\theta V_0^\theta(\rho) + \eta \sum_{i=1}^m \frac{\hat{\nabla}_\theta V_i^\theta(\rho)}{\hat{V}_i^\theta(\rho)}. \quad (3)$$

Based on Proposition 3.3, we proceed to establish the sub-Gaussian tail bound for the estimator (3).

Lemma 3.4. *Let Assumption 3.1 hold. Fix a confidence level $\delta \in (0, 1]$ and an accuracy $\varepsilon > 0$. By setting $n = \mathcal{O}(\varepsilon^{-2}(1-\gamma)^{-6}(\min_{i \in [m]} V_i^\theta(\rho))^{-4} \eta^2 \ln \delta^{-1})$ and $H = \mathcal{O}(\ln(\varepsilon \min_{i \in [m]} V_i^\theta(\rho))^{-1})$, we have*

$$\mathbb{P} \left(\left\| \hat{\nabla}_\theta B_\eta^\theta(\rho) - \nabla_\theta B_\eta^\theta(\rho) \right\| \leq \varepsilon \right) \geq 1 - \delta.$$

The proof of Lemma 3.4, along with the explicit forms for n and H , can be found in Appendix D.2.

Lemma 3.4 indicates that the sample complexity for obtaining an accurate estimate of the log barrier gradient depends on the distance of the iterates to the boundary, which is of order $\mathcal{O}((\min_{i \in [m]} V_i^\theta(\rho))^{-4})$. This observation will be crucial in deriving the convergence rate and sample complexity of our algorithm.

3.2 Tuning the Stepsize

The log barrier function is not smooth globally because its gradient becomes unbounded as the iterate approaches the boundary of the feasible domain. However, given a local region, the gradient of the log barrier function has bounded growth since $V_i^\theta(\rho)$ is bounded and smooth in that area. Based on this observation, the LB-SGD algorithm (Usmanova et al., 2024) developed a local smoothness constant, M_t by assuming access to unbiased estimators of $\{V_i^{\theta_t}(\rho)\}_{i=0}^m$. In our RL setting, we incorporate the biases of the estimators (see Proposition 3.3), and accordingly tune the stepsize for staying in the local region as follows.

The local smoothness constant \hat{M}_t , accounting for the biases and variances of the objective values and gradients of $\{V_i^{\theta_t}(\rho)\}_{i=0}^m$, are:

$$\hat{M}_t := M + \sum_{i=1}^m \frac{10M\eta}{\underline{\alpha}_i(t)} + 8\eta \sum_{i=1}^m \frac{(\bar{\beta}_i(t))^2}{(\underline{\alpha}_i(t))^2}.$$

Here, $\underline{\alpha}_i(t)$ represents the lower confidence bound of the constraint function $V_i^{\theta_t}(\rho)$, and $\bar{\beta}_i(t)$ denotes the upper confidence bound of $|\langle \nabla_\theta V_i^{\theta_t}(\rho), \nabla_\theta B_\eta^{\theta_t}(\rho) / \|\nabla_\theta B_\eta^{\theta_t}(\rho)\| \rangle|$. These confidence bounds are derived from Proposition 3.3 and detailed in Appendix D.3.1. To prevent overshooting and ensure that iterations remain within the local region where the estimator is valid, we set the stepsize γ_t as follows:

$$\gamma_t = \min \left\{ \frac{1}{\hat{M}_t}, \min_{i \in [m]} \left\{ \frac{\underline{\alpha}_i(t)}{\sqrt{M\underline{\alpha}_i(t)} + 2|\bar{\beta}_i(t)|} \right\} \frac{1}{\|\hat{\nabla}_x B_\eta^{\theta_t}(\rho)\|} \right\}. \quad (4)$$

Above, the second term inside the minimization corresponds to the region around the current iterate θ_t where the estimator is valid (see Appendix D.3.1).

With the gradient estimators and stepsize defined, we can now provide the LB-SGD approach in Algorithm 1. In summary, the LB-SGD algorithm implements stochastic gradient ascent on the log barrier function $B_\eta^{\theta_t}(\rho)$ using the sampling scheme provided in Section 3.1, where η controls the optimality of the algorithm's output. If the norm of the estimated gradient is smaller than $\frac{\eta}{2}$, the algorithm terminates. However, if the norm exceeds this threshold, the algorithm proceeds with stochastic gradient ascent, based on the stepsize specified in line 8.

Algorithm 1 LB-SGD

```

1: Input: Smoothness parameter  $M = \frac{M_g^2 + M_h}{(1-\gamma)^2}$ ,
   batch size  $n$ , truncated horizon  $H$ , number of it-
   erations  $T$ , confidence bound  $\delta \in (0, 1)$ ,  $\eta > 0$ .
2: for  $t = 0, 1, \dots, T - 1$  do
3:   Compute  $\hat{V}_i^{\theta_t}(\rho)$  and  $\hat{\nabla}_{\theta} V_i^{\theta_t}(\rho)$  using sampling
   scheme 3.1.
4:   Compute  $\hat{\nabla}_{\theta} B_{\eta}^{\theta_t}(\rho)$  using Eq.(3).
5:   if  $\|\hat{\nabla}_{\theta} B_{\eta}^{\theta_t}(\rho)\| \leq \frac{\eta}{2}$  then
6:     Break and return  $\theta_{\text{break}}$ .
7:   end if
8:    $\theta_{t+1} = \theta_t + \gamma_t \hat{\nabla}_{\theta} B_{\eta}^{\theta_t}(\rho)$ , where  $\gamma_t$  is defined in
   Eq.(4).
9: end for
10: Return  $\theta_{\text{out}}$ , which can be either  $\theta_{\text{break}}$  or  $\theta_T$ .

```

Under appropriate assumptions on the policy parameterization, we prove that Algorithm 1 can find a policy that is $\mathcal{O}(\sqrt{\varepsilon_{\text{bias}}}) + \tilde{\mathcal{O}}(\varepsilon)$ -optimal with a sample complexity of $\tilde{\mathcal{O}}(\varepsilon^{-6})$. Here, $\varepsilon_{\text{bias}}$ represents the transfer error from Assumption 4.8, as detailed in Section 4.2. We provide an informal statement of the main result of our paper here, and in the next section, we elaborate on the assumptions to formalize and prove this statement.

Theorem 3.5. *Under suitable assumptions (see Theorem 4.12 for the precise statement), Algorithm 1 has the following properties using $\tilde{\mathcal{O}}(\varepsilon^{-6})$ samples:*

1. *Safe exploration (see Definition 2.2) is satisfied with high probability.*
2. *The output policy $\pi_{\theta_{\text{out}}}$ achieves $\mathcal{O}(\sqrt{\varepsilon_{\text{bias}}}) + \tilde{\mathcal{O}}(\varepsilon)$ -optimality with high probability.*

This result extends the findings of (Usmanova et al., 2024) in the optimization setting to the RL setting. Their work focuses on non-convex objective and constraint functions $V_i^{\theta}(\rho)$, and consequently, only convergence to a stationary point can be guaranteed. In this paper, we further demonstrate that LB-SGD ensures that the last iterate policy converges towards a globally optimal policy while guaranteeing safe exploration. Despite the non-convex nature of the objective and constraint functions in the RL setting (Agarwal et al., 2021), our sample complexity aligns with that of zeroth-order feasible iterate method for solving convex constrained optimization problems, as demonstrated in (Usmanova et al., 2024, Theorem 10).

4 TECHNICAL ANALYSIS OF LOG BARRIER FOR CMDPS

The proof of Theorem 4.12 is divided into safe exploration analysis (Section 4.1) and convergence analysis (Section 4.2). In the safe exploration analysis (Proposition 4.3), we utilize Slater’s condition and an Extended Mangasarian-Fromovitz constraint qualification (MFCQ) assumption (see Assumptions 4.1 and 4.2, respectively) to establish lower bounds on the distance of the iterates from the boundary. For the convergence analysis, we rely on the properties of the policy parametrization. Specifically, under the relaxed Fisher non-degeneracy assumption and the bounded transfer error assumption (see Assumptions 4.4 and 4.8, respectively), we establish the gradient dominance property of the log barrier function in Lemma 4.10. This allows us to bound the gap between $V_0^{\theta}(\rho)$ and $V_0^{\pi^*}(\rho)$ by the norm of the gradient $\nabla_{\theta} B_{\eta}^{\theta}(\rho)$.

4.1 Safe Exploration of the Algorithm

Assumption 4.1 (Slater’s condition). *There exist a known starting point $\theta_0 \in \Theta$ and $\nu_s > 0$ such that $V_i^{\theta_0}(\rho) \geq \nu_s$, $\forall i \in [m]$.*

Assumption 4.1 has been commonly used in the analysis of CMDPs (Bai et al., 2023; Ding et al., 2020; Liu et al., 2021a,b). It is natural to consider this assumption since without a safe initial policy π_{θ_0} , the safe exploration (see Definition 2.2) cannot be satisfied.

Our next assumption requires the existence of a direction that ensures the iterates too close to the boundary (within a distance of ν_{emf}), can move away from it (by ℓ). To formalize it, for $p > 0$, let $\mathbf{B}_p(\theta) := \{i \in [m] \mid 0 < V_i^{\theta}(\rho) \leq p\}$ be the set of constraints indicating that θ is approximately p -close to the boundary.

Assumption 4.2 (Extended MFCQ). *There exist constants $0 < \nu_{\text{emf}} \leq \nu_s$ and $\ell > 0$ such that for any $\theta \in \Theta$, there is a direction $s_{\theta} \in \mathbb{R}^d$ with $\|s_{\theta}\| = 1$ satisfying $\langle s_{\theta}, \nabla V_i^{\theta}(\rho) \rangle > \ell$ for all $i \in \mathbf{B}_{\nu_{\text{emf}}}(\theta)$.*

Note that the MFCQ assumption (Mangasarian and Fromovitz, 1967), commonly considered in non-convex constrained optimization (Muehlebach and Jordan, 2022; Boob et al., 2023), corresponds to the above with $l = 0$ and $\nu_{\text{emf}} = 0$.² Without Assumption 4.2, the iterates generated by the log barrier approach can be $\mathcal{O}(\exp -\eta^{-1})$ close to the boundary, as illustrated in an example in Appendix B.2. This, in turn, would require very accurate gradient estimators to ensure safe exploration, resulting in extremely high sample complexity,

²The MFCQ assumption ensures that the Karush-Kuhn-Tucker conditions are necessary optimality conditions (Muehlebach and Jordan, 2022; Boob et al., 2023).

as inferred from Lemma 3.4. For further insights, we show the cases in which this assumption is implied by the MFCQ assumption in Appendix B.1.

We are now ready to establish the safe exploration property of Algorithm 1.

Proposition 4.3. *Let Assumptions 3.1, 4.1, and 4.2 hold. Fix a confidence level $\delta \in (0, 1]$. By setting $\eta \leq \nu_{emf}$, $n = \mathcal{O}(\eta^{-4} \ell^{-4m} (1 - \gamma)^{-6-8m} \ln \delta^{-1})$ and $H = \mathcal{O}(\ln(\ell\eta)^{-1})$, we have*

$$\mathbb{P} \left\{ \forall t \in \{0, \dots, T\}, \min_{i \in [m]} V_i^{\theta_t}(\rho) \geq c\eta \right\} \geq 1 - mT\delta$$

$$\text{with } c := \left(\frac{\ell(1-\gamma)^2}{4M_g(1+\frac{4m}{3})} \right)^m.$$

The proof of Proposition 4.3 can be found in Appendix D.3. As the transition dynamics are unknown, Algorithm 1 only relies on estimates of the true gradients. Accordingly, Proposition 4.3 demonstrates that the LB-SGD algorithm ensures safe exploration with high probability. Importantly, it also shows that the iterates remain within a distance of $\Omega(\eta)$ from the boundary. This ensures an upper bound on the sample complexity for an accurate estimate of the log barrier gradient, as inferred from Lemma 3.4, and consequently helps derive the sample complexity of our algorithm.

4.2 Convergence and Sample Complexity

To establish algorithm convergence, we rely on two commonly employed assumptions. The first assumption, relaxed Fisher non-degeneracy (see Assumption 4.4), ensures that the policy can adequately explore the state-action space. The second assumption, bounded transfer error (see Assumption 4.8), ensures that our parameterized policy set sufficiently covers the entire stochastic policy set.

To introduce the relaxed Fisher non-degeneracy assumption, we first define the discounted state-action visitation distribution as $d_\rho^\theta(s, a) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t P(s_t = s, a_t = a)$. The Fisher information matrix induced by policy π_θ is defined as

$$F^\theta(\rho) := \mathbb{E}_{(s,a) \sim d_\rho^\theta} [\nabla \log \pi_\theta(a|s) (\nabla \log \pi_\theta(a|s))^T].$$

Assumption 4.4 (Relaxed Fisher non-degeneracy). *For any $\theta \in \Theta$, there exists a positive constant μ_F such that the smallest non-zero eigenvalue of $F^\theta(\rho)$ is lower bounded by μ_F .*

Assumption 4.4 weakens the Fisher non-degeneracy assumption commonly used to prove convergence of policy gradient methods (Bai et al., 2023; Liu et al., 2020b; Yuan et al., 2022; Masiha et al., 2022;

Fatkhullin et al., 2023; Ding et al., 2022b), which further requires $F^\theta(\rho)$ to be strictly positive definite, as formalized in the following definition.

Definition 4.5 (Fisher non-degeneracy). *For any $\theta \in \Theta$, there exists a positive constant μ_F such that all eigenvalues of $F^\theta(\rho)$ are lower bounded by μ_F .*

Despite the importance of the above, there has been only a partial understanding of which policy classes comply with Assumption 4.4 and Definition 4.5. It has been claimed that in the tabular setting, the softmax parameterization fails to satisfy Fisher non-degeneracy, particularly when the policy approaches a deterministic policy (Ding et al., 2022b). Our result below complements this understanding.

Fact 4.6. *1) The softmax parameterization does not satisfy **Fisher non-degeneracy**. 2) Softmax, log-linear, and neural softmax parameterizations fail to satisfy **relaxed Fisher non-degeneracy** as the policy approaches determinism.*

The proof of Fact 4.6 can be found in Appendix C.1. Building upon this fact, to ensure softmax, log-linear, and neural softmax parameterizations satisfy relaxed Fisher non-degeneracy, we can constrain the parameter space Θ to a compact set, preventing the policy from approaching determinism. Note that for continuous action spaces, Gaussian and Cauchy policies satisfy Fisher non-degeneracy (see (Fatkhullin et al., 2023, Appendix B)). While the above assumption ensures sufficient exploration of the policy, the bounded transfer error assumption concerns the richness of the policy class. We define the notion of “richness” in policy parameterization as follows:

Definition 4.7 (Richness of Policy Parameterization). *Define Π as the closure of all stochastically parameterized policies, denoted by $\mathcal{C}\{\pi_\theta \mid \theta \in \mathbb{R}^d\}$. If we have another policy parameterization Π' , we say Π' is a richer parameterization compared to Π if $\Pi' \subsetneq \Pi$.*

To formalize the bounded transfer error assumption, we start by defining the state value function as $V_i^\theta(s) := \mathbb{E}_{\tau \sim \pi_\theta} [\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) \mid s_0 = s]$, the state-action value function as $Q_i^\theta(s, a) := \mathbb{E}_{\tau \sim \pi_\theta} [\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) \mid s_0 = s, a_0 = a]$, and the advantage function as $A_i^\theta(s, a) = Q_i^\theta(s, a) - V_i^\theta(s)$. With these definitions in place, the transfer error is defined as $L(\mu_i^*, \theta, d_\rho^{\pi^*}) := \mathbb{E}_{(s,a) \sim d_\rho^{\pi^*}} \left[\left(A_i^\theta(s, a) - (1 - \gamma) \mu_i^{*T} \nabla \log \pi_\theta(a|s) \right)^2 \right]$, where $\mu_i^* = (F^\theta(\rho))^{-1} \nabla_\theta V_i^\theta(\rho)$. This formulation is termed as the transfer error because it shows the error in approximating the advantage function A_i^θ , which depends on d_ρ^θ , while the expectation of the error is taken with respect to a fixed measure $d_\rho^{\pi^*}$.

Assumption 4.8 (Bounded transfer error). *For any*

$\theta \in \Theta$, there exists a non-negative constant ε_{bias} such that for $i \in \{0, \dots, m\}$, $L(\mu_i^*, \theta, d_\rho^{\pi^*}) \leq \varepsilon_{bias}$.

Assumption 4.8 has been utilized in several works (Liu et al., 2020b; Yuan et al., 2022; Fatkhullin et al., 2023; Ding et al., 2022b). The general understanding is that softmax parameterization results in $\varepsilon_{bias} = 0$. This result is extended by either assuming a very specific class of MDPs, such as a linear MDP model with low-rank transition dynamics (Jiang et al., 2017; Jin et al., 2020; Yang and Wang, 2019), or a very specific policy class, such as a “rich” two-layer neural network (Wang et al., 2020). Building upon these findings, we present a more general result connecting the richness of policy classes to the transfer error.

Fact 4.9. *For log-linear and neural softmax policy parameterizations, a richer policy parameterization leads to a decrease in the transfer error ε_{bias} .*

We have provided the exact formalization of “richness” of a parametrization and the proof of Fact 4.9 in Appendix C.2.

With these assumptions in place, we can establish bounds on the optimality of the policy π_θ given $\nabla_\theta B_\eta^\theta(\rho)$ in the following lemma.

Lemma 4.10. *Let Assumptions 3.1, 4.4, and 4.8 hold. For any $\theta \in \Theta$, we have*

$$V_0^{\pi^*}(\rho) - V_0^\theta(\rho) \leq m\eta + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} \left(1 + \sum_{i \in [m]} \frac{\eta}{V_i^\theta(\rho)} \right) + \frac{M_h}{\mu_F} \|\nabla_\theta B_\eta^\theta(\rho)\|.$$

The proof of Lemma 4.10 is provided in Appendix D.4. Previous works such as (Ding et al., 2022b; Masiha et al., 2022) established gradient dominance of the value function in the MDP setting, bounding the optimality gap for $V_0^\theta(\rho)$ by the norm of its gradient $\nabla_\theta V_0^\theta(\rho)$. Here, we establish this property in the CMDP setting by bounding the optimality gap using the norm of the log barrier gradient, $\nabla_\theta B_\eta^\theta(\rho)$, along with an additional term $\mathcal{O}\left(\sum_{i \in [m]} \eta \sqrt{\varepsilon_{bias}} / V_i^\theta(\rho)\right)$. Therefore, to bound the sub-optimality of the stationary point of the log barrier function, we need to provide a lower bound on the distance of the stationary point from the boundary, which is shown in the following lemma.

Lemma 4.11. *Let Assumptions 3.1 and 4.2 hold. For any stationary point θ_{st} of the log barrier function, we have*

$$\min_{i \in [m]} \left\{ V_i^{\theta_{st}}(\rho) \right\} \geq \min \left\{ \nu_{emf}, \frac{\min\{\eta, \nu_{emf}\} \ell}{mL} \right\}.$$

The proof of Lemma 4.11 can be found in Appendix D.5. From the above lemma, we prove that the stationary points of the log barrier function are at most $\Omega(\nu_{emf} + \eta)$ close to the boundary. Combining this with Lemma 4.10, we conclude that the stationary points are $\mathcal{O}(\eta + \sqrt{\varepsilon_{bias}} \max\{1, \eta/\nu_{emf}\})$ -optimal. Meanwhile, the gradient ascent method ensures convergence to the stationary point of $B_\eta^\theta(\rho)$. Leveraging Lemmas 4.3 and 4.10, we complete the proof of Theorem 3.5. Below, we provide the precise statement of Theorem 3.5.

Theorem 4.12. *Let Assumptions 3.1, 4.1, 4.2, 4.4, and 4.8 hold. Fix a confidence level $\delta \in (0, 1]$, by setting n and H as in Proposition 4.3, after T iterations of the Algorithm 1, the following holds:*

1. $\mathbb{P}\left(\forall t \in \{0, \dots, T\}, \min_{i \in [m]} V_i^{\theta_t}(\rho) \geq 0\right) \geq 1 - mT\delta$.
2. *With a probability of at least $1 - mT\delta$, the output policy $\pi_{\theta_{out}}$ satisfies*

$$V_0^{\pi^*}(\rho) - V_0^{\theta_{out}}(\rho) \leq \mathcal{O}\left(\frac{\sqrt{\varepsilon_{bias}}}{\ell^m(1-\gamma)^{2m+1}}\right) + \mathcal{O}\left(\frac{\eta}{\mu_F}\right) + \mathcal{O}\left(\eta \ln \frac{1}{\eta}\right) + \mathcal{O}\left(\exp(-C\mu_F T \eta^2)\right)(V_0^{\pi^*}(\rho) - V_0^{\theta_0}(\rho)),$$

$$\text{where } C := \frac{c}{2L^2(1+\frac{m}{c}) \max\left\{4 + \frac{5Mc}{L^2}, 1 + \sqrt{\frac{Mc}{4L^2}}\right\}}.$$

The proof of Theorem 4.12 can be found in Appendix D.6. Inequality (5) yields the following key insights: the last iterate of the algorithm converges to the neighborhood of the optimal point of $V_0^{\pi^*}(\rho)$ at a rate of $\mathcal{O}(\exp(-C\mu_F T \eta^2))$. Hence, larger values of μ_F lead to faster convergence. On the other hand, the neighborhood’s radius is influenced by two factors: μ_F and the transfer error ε_{bias} . A smaller μ_F corresponds to a less random policy, reducing exploration. A larger ε_{bias} indicates inadequate policy parameterization. Consequently, smaller μ_F values and larger ε_{bias} values prevent the algorithm from reaching the optimal policy. Therefore, μ_F controls optimality and convergence speed.

Based on Theorem 4.12, we can determine the sample complexity of the algorithm required to ensure safe exploration and achieve ε -optimality, as stated in the following corollary, whose proof is provided in Appendix D.7.

Corollary 4.13. *Given $\varepsilon > 0$, the sample complexity of Algorithm 1 to return an $\mathcal{O}(\varepsilon \ln \varepsilon^{-1} + \varepsilon \mu_F^{-1}) + \mathcal{O}(\varepsilon_{bias}^{1/2} \ell^{-m} (1-\gamma)^{-2m-1})$ -optimal policy while ensuring safe exploration with high probability is $\mathcal{O}(\mu_F^{-1} \varepsilon^{-6} \ell^{-6m} (1-\gamma)^{-10-12m} \ln \ell^{-1} \ln^3 \varepsilon^{-1})$.*

When compared to the state-of-the-art policy gradient-based algorithm, C-NPG-PDA (Bai et al.,

2023), which only provides guarantees for averaged zero constraint violation assuming Fisher non-degeneracy, our algorithm demands an additional $\mathcal{O}(\varepsilon^{-2})$ samples only assuming relaxed Fisher non-degeneracy. This increase in sampling requirement serves as the price for ensuring safe exploration.

5 COMPUTATIONAL EXPERIMENT

While our work primarily focuses on establishing theoretical guarantees for safe exploration in CMDPs, we validate the performance of our algorithm against benchmark algorithms in Appendix E. Since our paper emphasizes the feasibility of iterates rather than bounding average constraint violations, we compare with two typical CMDP learning algorithms: (1) the IPO algorithm (Liu et al., 2020a), which combines the log barrier method with a policy gradient approach and fixed stepsize, and (2) the RPG-PD algorithm (Ding et al., 2024), a primal-dual method that guarantees the feasibility of the last iterate, using an entropy-regularized policy gradient and quadratic-regularized gradient ascent for the dual variable.

As RPG-PD is designed for discrete action spaces, we conducted experiments in a standard gridworld environment with softmax parameterization. Our experiments show that both IPO and RPG-PD ensure safe exploration through manual stepsize tuning, LB-SGD achieves safe exploration using adaptive stepsizes estimated from samples, eliminating the need for tuning. Moreover, LB-SGD is more sample-efficient in finding the optimal policy compared to RPG-PD, which aligns with theoretical results, as RPG-PD requires $\mathcal{O}(\varepsilon^{-14})$ samples to find an $\mathcal{O}(\varepsilon)$ -optimal policy, while LB-SGD requires only $\tilde{\mathcal{O}}(\varepsilon^{-6})$ samples.

Our experiment confirmed that LB-SGD achieves safe exploration while converges to the optimal policy efficiently. However, as expected, ensuring these guarantees necessitates a higher number of samples per iteration near the boundary for accurate estimates. It would be interesting to determine whether this sample complexity is inherent to our algorithm and its analysis or to the safe exploration requirement.

6 CONCLUSION

We employed a log barrier policy gradient approach for ensuring safe exploration in CMDPs. Our work establishes the convergence of the algorithm to an optimal point and characterizes its sample complexity. A potential direction for future research is to explore methods that can further reduce the sample complexity of safe exploration. This could involve incorpo-

rating variance reduction techniques, leveraging MDP structural characteristics (e.g., natural policy gradient method), and extending the relaxed Fisher non-degenerate parameterization to general policy representations. Another potential research avenue is to establish lower bounds for the safe exploration problem.

7 ACKNOWLEDGMENTS

This research is gratefully supported by the Swiss National Science Foundation (SNSF).

References

- Abe, N., Melville, P., Pendus, C., Reddy, C. K., Jensen, D. L., Thomas, V. P., Bennett, J. J., Anderson, G. F., Cooley, B. R., Kowalczyk, M., et al. (2010). Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84.
- Achiam, J., Held, D., Tamar, A., and Abbeel, P. (2017). Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR.
- Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(98):1–76.
- Agarwal, M., Bai, Q., and Aggarwal, V. (2022). Regret guarantees for model-based reinforcement learning with long-term average constraints. In *Uncertainty in Artificial Intelligence*, pages 22–31. PMLR.
- Altman, E. (1999). *Constrained Markov decision processes: stochastic modeling*. Routledge.
- Amani, S., Alizadeh, M., and Thrampoulidis, C. (2019). Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32.
- Azar, M. G., Munos, R., and Kappen, B. (2012). On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*.
- Bai, Q., Bedi, A. S., Agarwal, M., Koppel, A., and Aggarwal, V. (2022). Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3682–3689.
- Bai, Q., Singh Bedi, A., and Aggarwal, V. (2023). Achieving zero constraint violation for constrained

- p reinforcement learning via conservative natural policy gradient primal-dual algorithm.
- Proceedings of the AAAI Conference on Artificial Intelligence*
- , 37(6):6737–6744.
- Baxter, J. and Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350.
- Berkenkamp, F., Krause, A., and Schoellig, A. P. (2021). Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *Machine Learning*, pages 1–35.
- Berkenkamp, F., Turchetta, M., Schoellig, A., and Krause, A. (2017). Safe model-based reinforcement learning with stability guarantees. *Advances in neural information processing systems*, 30.
- Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. (2019). Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*.
- Boob, D., Deng, Q., and Lan, G. (2023). Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, 197(1):215–279.
- Bura, A., HasanzadeZonuzi, A., Kalathil, D., Shakkottai, S., and Chamberland, J.-F. (2022). Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. *Advances in Neural Information Processing Systems*, 35:1047–1059.
- Calvo-Fullana, M., Paternain, S., Chamon, L. F., and Ribeiro, A. (2023). State augmented constrained reinforcement learning: Overcoming the limitations of learning with rewards. *IEEE Transactions on Automatic Control*.
- Cheng, R., Orosz, G., Murray, R. M., and Burdick, J. W. (2019). End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3387–3395.
- Chow, Y., Nachum, O., Duenez-Guzman, E., and Ghavamzadeh, M. (2018). A lyapunov-based approach to safe reinforcement learning. *Advances in neural information processing systems*, 31.
- Chow, Y., Nachum, O., Faust, A., Duenez-Guzman, E., and Ghavamzadeh, M. (2019). Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*.
- Ding, D., Wei, C.-Y., Zhang, K., and Ribeiro, A. (2024). Last-iterate convergent policy gradient primal-dual methods for constrained mdps. *Advances in Neural Information Processing Systems*, 36.
- Ding, D., Zhang, K., Basar, T., and Jovanovic, M. (2020). Natural policy gradient primal-dual method for constrained Markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390.
- Ding, D., Zhang, K., Duan, J., Başar, T., and Jovanović, M. R. (2022a). Convergence and sample complexity of natural policy gradient primal-dual methods for constrained mdps. *arXiv preprint arXiv:2206.02346*.
- Ding, Y., Zhang, J., and Lavaei, J. (2022b). On the global optimum convergence of momentum-based policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 1910–1934. PMLR.
- Eriksson, D. and Jankowiak, M. (2021). High-dimensional bayesian optimization with sparse axis-aligned subspaces. In *Uncertainty in Artificial Intelligence*, pages 493–503. PMLR.
- Fatkhullin, I., Barakat, A., Kireeva, A., and He, N. (2023). Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR.
- Fisac, J. F., Akametalu, A. K., Zeilinger, M. N., Kaynama, S., Gillula, J., and Tomlin, C. J. (2018). A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752.
- Frazier, P. I. (2018). Bayesian optimization. In *Recent advances in optimization and modeling of contemporary problems*, pages 255–278. Informa.
- Garcia, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480.
- HasanzadeZonuzi, A., Kalathil, D., and Shakkottai, S. (2021). Model-based reinforcement learning for infinite-horizon discounted constrained markov decision processes. In *International Joint Conference on Artificial Intelligence (IJCAI)*.
- Hernández-Lerma, O. and González-Hernández, J. (2000). Constrained markov control processes in borel spaces: the discounted case. *Mathematical Methods of Operations Research*, 52:271–285.

- Hinder, O. and Ye, Y. (2019). A polynomial time log barrier method for problems with nonconvex constraints. *arXiv preprint: <https://arxiv.org/pdf/1807.00404.pdf>*.
- Jayant, A. K. and Bhatnagar, S. (2022). Model-based safe deep reinforcement learning via a constrained proximal policy optimization algorithm. *Advances in Neural Information Processing Systems*, 35:24432–24445.
- Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR.
- Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is q-learning provably efficient? *Advances in neural information processing systems*, 31.
- Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.
- Kalagarla, K. C., Jain, R., and Nuzzo, P. (2023). Safe posterior sampling for constrained mdps with bounded constraint violation. *arXiv preprint [arXiv:2301.11547](https://arxiv.org/abs/2301.11547)*.
- Kalweit, G., Huegle, M., Werling, M., and Boedecker, J. (2020). Deep constrained q-learning. *arXiv preprint [arXiv:2003.09398](https://arxiv.org/abs/2003.09398)*.
- Koenig, S. and Simmons, R. G. (1993). Complexity analysis of real-time reinforcement learning. In *AAAI*, volume 93, pages 99–105.
- Kohler, J. M. and Lucchi, A. (2017). Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, pages 1895–1904. PMLR.
- Koller, T., Berkenkamp, F., Turchetta, M., Boedecker, J., and Krause, A. (2019). Learning-based model predictive control for safe exploration and reinforcement learning. *arXiv preprint [arXiv:1906.12189](https://arxiv.org/abs/1906.12189)*.
- Koppejan, R. and Whiteson, S. (2011). Neuroevolutionary reinforcement learning for generalized control of simulated helicopters. *Evolutionary intelligence*, 4:219–241.
- Liang, Q., Que, F., and Modiano, E. (2018). Accelerated primal-dual policy optimization for safe reinforcement learning. *arXiv preprint [arXiv:1802.06480](https://arxiv.org/abs/1802.06480)*.
- Liu, T., Zhou, R., Kalathil, D., Kumar, P., and Tian, C. (2021a). Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193.
- Liu, T., Zhou, R., Kalathil, D., Kumar, P. R., and Tian, C. (2022). Policy optimization for constrained mdps with provable fast global convergence.
- Liu, Y., Ding, J., and Liu, X. (2020a). Ipo: Interior-point policy optimization under constraints. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4940–4947.
- Liu, Y., Halev, A., and Liu, X. (2021b). Policy learning with constraints in model-free reinforcement learning: A survey. In *The 30th International Joint Conference on Artificial Intelligence (IJCAI)*.
- Liu, Y., Zhang, K., Basar, T., and Yin, W. (2020b). An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636.
- Mangasarian, O. L. and Fromovitz, S. (1967). The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. *Journal of Mathematical Analysis and applications*, 17(1):37–47.
- Masiha, S., Salehkaleybar, S., He, N., Kiyavash, N., and Thiran, P. (2022). Stochastic second-order methods improve best-known sample complexity of sgd for gradient-dominated functions. *Advances in Neural Information Processing Systems*, 35:10862–10875.
- Metelli, A. M., Manneschi, G., and Restelli, M. (2022). Policy space identification in configurable environments. *Machine Learning*, pages 1–53.
- Mirchevska, B., Pek, C., Werling, M., Althoff, M., and Boedecker, J. (2018). High-level decision making for safe and reasonable autonomous lane changing using reinforcement learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2156–2162. IEEE.
- Mondal, W. U. and Aggarwal, V. (2024). Sample-efficient constrained reinforcement learning with general parameterization. *arXiv preprint [arXiv:2405.10624](https://arxiv.org/abs/2405.10624)*.
- Moriconi, R., Deisenroth, M. P., and Sesh Kumar, K. (2020). High-dimensional bayesian optimization using low-dimensional feature spaces. *Machine Learning*, 109:1925–1943.
- Muehlebach, M. and Jordan, M. I. (2022). On constraints in first-order optimization: A view from non-smooth dynamical systems. *The Journal of Machine Learning Research*, 23(1):11681–11727.

- Ono, M., Pavone, M., Kuwata, Y., and Balaram, J. (2015). Chance-constrained dynamic programming with application to risk-aware robotic space exploration. *Autonomous Robots*, 39:555–571.
- Papini, M., Pirotta, M., and Restelli, M. (2022). Smoothing policies and safe policy gradients. *Machine Learning*, 111(11):4081–4137.
- Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.
- Schlaginhaufen, A. and Kamgarpour, M. (2023). Identifiability and generalizability in constrained inverse reinforcement learning. In *International Conference on Machine Learning*, pages 30224–30251. PMLR.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *nature*, 550(7676):354–359.
- Stooke, A., Achiam, J., and Abbeel, P. (2020). Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, pages 9133–9143. PMLR.
- Sui, Y., Gotovos, A., Burdick, J., and Krause, A. (2015). Safe exploration for optimization with gaussian processes. In *International conference on machine learning*, pages 997–1005. PMLR.
- Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.
- Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.
- Tessler, C., Mankowitz, D. J., and Mannor, S. (2019). Reward constrained policy optimization. In *International Conference on Learning Representations*.
- Usmanova, I., As, Y., Kamgarpour, M., and Krause, A. (2024). Log barriers for safe black-box optimization with application to safe reinforcement learning. *Journal of Machine Learning Research*, 25(171):1–54.
- Usmanova, I., Krause, A., and Kamgarpour, M. (2020). Safe non-smooth black-box optimization with application to policy search. In *Learning for Dynamics and Control*, pages 980–989. PMLR.
- Vaswani, S., Yang, L., and Szepesvári, C. (2022). Near-optimal sample complexity bounds for constrained mdps. *Advances in Neural Information Processing Systems*, 35:3110–3122.
- Wachi, A., Sui, Y., Yue, Y., and Ono, M. (2018). Safe exploration and optimization of constrained mdps using gaussian processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Wang, L., Cai, Q., Yang, Z., and Wang, Z. (2020). Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*.
- Wei, H., Liu, X., and Ying, L. (2022a). A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3868–3876.
- Wei, H., Liu, X., and Ying, L. (2022b). Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *International Conference on Artificial Intelligence and Statistics*, pages 3274–3307. PMLR.
- Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning*, pages 5–32.
- Xu, P., Gao, F., and Gu, Q. (2020). Sample efficient policy gradient methods with recursive variance reduction. In *International Conference on Learning Representations*.
- Xu, T., Liang, Y., and Lan, G. (2021). CRPO: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR.
- Yang, L. and Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR.
- Yuan, R., Gower, R. M., and Lazaric, A. (2022). A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 3332–3380. PMLR.
- Zang, W., Miao, F., Gravina, R., Sun, F., Fortino, G., and Li, Y. (2020). CMDP-based intelligent transmission for wireless body area network in remote health monitoring. *Neural computing and applications*, 32:829–837.
- Zeng, S., Doan, T. T., and Romberg, J. (2022). Finite-time complexity of online primal-dual natural actor-critic algorithm for constrained markov decision processes. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 4028–4033. IEEE.

Zhang, X., Zhang, K., Miehling, E., and Basar, T. (2019). Non-cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 32.

Zheng, L. and Ratliff, L. (2020). Constrained upper confidence reinforcement learning. In *Learning for Dynamics and Control*, pages 620–629. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:

- (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

We discuss the mathematical setting in Section 2 and illustrate the algorithm in Section 3.2. The assumptions are described in detail in Section 4.

- (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

The results are presented in Theorem 4.12 and further elaborated in Corollary 4.13.

- (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

The code is included in the supplementary file.

2. For any theoretical claim, check if you include:

- (a) Statements of the full set of assumptions of all theoretical results. [Yes]

We state the full set of assumptions needed for the theoretical results.

- (b) Complete proofs of all theoretical results. [Yes]

The detail of the proofs are provided in Appendix D.

- (c) Clear explanations of any assumptions. [Yes]

We explain the Extended Mangasarian-Fromovitz constraint qualification assumption (See Assumption 4.2), relaxed Fisher non-degeneracy (See Assumption 4.4), and the bounded transfer error assumptions (See Assumption 4.8). We discuss the limitations of these assumptions immediately after introducing each of them, and further discussion is provided in Appendices B and C.

3. For all figures and tables that present empirical results, check if you include:

- (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

The code is included in the supplementary file.

- (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
- We provide the training details in Appendix E

- (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

We plotted the error bar with respect to the random seed in Appendix E.

- (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

All the experiments were carried out on a MacBook Pro with an Apple M1 Pro chip and 32 GB of RAM.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

- (a) Citations of the creator If your work uses existing assets. [Yes]

Our code is developed based on (Schlaginhausen and Kamgarpour, 2023, <https://github.com/andrschl/cirl>). We acknowledged it in Appendix E.

- (b) The license information of the assets, if applicable. [Not Applicable]

- (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

- (d) Information about consent from data providers/curators. [Not Applicable]

- (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

- (a) The full text of instructions given to participants and screenshots. [Not Applicable]

- (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

- (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Table 2: Sample complexity for achieving ε -optimal objectives with guarantees on constraint violations in stochastic policy gradient-based algorithms, considering various parameterizations for discounted infinite horizon CMDPs.

Stochastic policy gradient-based algorithms						
Parameterization	Algorithm	Sample complexity	Constraint violation	Optimality	Generative model	Slater's condition
Softmax	NPG-PD (Ding et al., 2020)	$\mathcal{O}(\varepsilon^{-2})$	Averaged $\mathcal{O}(\varepsilon)$	Average	✓	✓
Softmax	PD-NAC (Zeng et al., 2022)	$\mathcal{O}(\varepsilon^{-6})$	Averaged $\mathcal{O}(\varepsilon)$	Average	×	✓
Neural softmax(ReLU)	CRPO (Xu et al., 2021)	$\mathcal{O}(\varepsilon^{-6})$	Averaged $\mathcal{O}(\varepsilon)$	Average	✓	×
Log-linear	RPG-PD (Ding et al., 2024)	$\tilde{\mathcal{O}}(\varepsilon^{-14})$	0 at last iterate	Last iterate	✓	✓
Fisher non-degenerate	PD-ANPG (Mondal and Aggarwal, 2024)	$\tilde{\mathcal{O}}(\varepsilon^{-3})$	Averaged $\mathcal{O}(\varepsilon)$	Average	✓	✓
Fisher non-degenerate	C-NPG-PDA (Bai et al., 2023)	$\tilde{\mathcal{O}}(\varepsilon^{-4})$	Averaged zero	Average	×	✓
Relaxed Fisher non-degenerate	LB-SGD (Usmanova et al., 2024)	$\tilde{\mathcal{O}}(\varepsilon^{-6})$	Safe exploration w.h.p	Last iterate	×	✓

A COMPARISON OF MODEL-FREE SAFE RL ALGORITHMS

Regarding the past work on policy gradient in infinite horizon discounted CMDPs, we further provide details on the notion of constraint satisfaction. To this end, we provide an extended version of Table 1 to include the assumptions.

1. (Slater's condition) Compared to Table 1, the above table includes an additional column detailing the assumptions required for convergence analysis. LB-SGD, unlike all the other methods in the table, requires a feasible initial policy since our work is the only one that focuses on safe exploration.
2. (Constraint violation) In our work, we define safe exploration as ensuring constraint satisfaction throughout the learning process, as defined in the property 2.2. Our LB-SGD algorithm achieves safe exploration with high probability. However, in (Bai et al., 2023), the authors claim to achieve zero constraint violation but employ a different definition, specified as:

$$\frac{1}{T} \sum_{t=0}^{T-1} V_i^{\theta_t}(\rho) \geq 0, \forall i \in [m].$$

It is important to note that while their algorithm aims for zero constraint violation, individual iterates during the learning process may still violate the constraints. Hence, we refer to it as an averaged zero constraint violation, since safe exploration represents a stronger notion of constraint violation guarantees. Additionally, in (Ding et al., 2024), the regularized policy gradient primal-dual (RPG-PD) algorithm returns the last iterate policy satisfying the constraints, but it does not provide guarantees for safe exploration.

3. (Sample complexity) In the constraint-rectified policy optimization (CRPO) algorithm (Xu et al., 2021), the authors provide a general result for measuring the algorithm's performance in (Xu et al., 2021, Theorem 2). We conclude that $\mathcal{O}(\varepsilon^{-6})$ is the optimal sample complexity for achieving an $\mathcal{O}(\varepsilon)$ optimality gap for the CRPO algorithm.

B DISCUSSION ON ASSUMPTION 4.2

B.1 Sufficient Conditions for Assumption 4.2

In this section, we first study the relationship between the extended MFCQ assumption and the MFCQ assumption. Let us state the MFCQ assumption (Mangasarian and Fromovitz, 1967) below.

Assumption B.1 (MFCQ (Mangasarian and Fromovitz, 1967)). *For every $\theta \in \Theta'$, where $\Theta' = \{\theta \in \Theta \mid \exists i \in [m], V_i^\theta(\rho) = 0\}$, there exists a direction s_θ such that $\langle s_\theta, \nabla V_i^\theta(\rho) \rangle > 0$ for all $i \in \mathbf{B}_0(\theta) := \{i \in [m] \mid V_i^\theta(\rho) = 0\}$.*

Let us define

$$\ell_\theta := \min_{i \in \mathbf{B}_0(\theta)} \left\{ \left\langle \frac{s_\theta}{\|s_\theta\|}, \nabla V_i^\theta(\rho) \right\rangle \right\},$$

$$\nu_\theta := \begin{cases} \min_{i \in \{[m] \setminus \mathbf{B}_0(\theta)\}} \{V_i^\theta(\rho)\}, & \{[m] \setminus \mathbf{B}_0(\theta)\} \neq \emptyset, \\ \frac{1}{1-\gamma}, & \text{otherwise.} \end{cases}$$

Under the MFCQ assumption, for $\theta \in \Theta'$, both ℓ_θ and ν_θ are strictly positive. Now, we argue that under Assumption 3.1, we can establish a relationship between the MFCQ assumption and the extended MFCQ assumption.

Proposition B.2. *Let Assumptions 3.1 and B.1 hold. Set $\ell := \inf_{\theta \in \Theta'} \{\frac{\ell_\theta}{2}\}$ and $\nu_1 := \inf_{\theta \in \Theta'} \{\frac{\nu_\theta}{3}\}$. If $\ell, \nu_1 > 0$, then Assumption 4.2, namely, the extended MFCQ Assumption, holds.*

Proof of Proposition B.2. Under Assumption 3.1, we know that $V_i^\theta(\rho)$ is L -Lipschitz continuous and M -smooth, as shown in Proposition 3.3. For each $\theta \in \Theta'$, consider $\theta_1 \in \mathcal{R}(\theta) := \{\theta_1 \mid \theta_1 \in \Theta, \|\theta_1 - \theta\| \leq \min\{\frac{\ell_\theta}{2M}, \frac{\nu_\theta}{3L}\}\}$. For $i \in \mathbf{B}_0(\theta)$, we have

$$V_i^{\theta_1}(\rho) \leq V_i^\theta(\rho) + L\|\theta - \theta_1\| \leq \frac{\nu_\theta}{3}.$$

For $i \notin \mathbf{B}_0(\theta)$, we have

$$V_i^{\theta_1}(\rho) \geq V_i^\theta(\rho) - L\|\theta - \theta_1\| \geq \frac{2\nu_\theta}{3}.$$

Therefore, we can conclude that $\mathbf{B}_{\frac{\nu_\theta}{3}}(\theta_1) \subset \mathbf{B}_0(\theta)$ for $\theta_1 \in \mathcal{R}(\theta)$. Next, we apply Assumption B.1 on $\theta_1 \in \mathcal{R}(\theta)$, we have for each $i \in \mathbf{B}_{\frac{\nu_\theta}{3}}(\theta_1)$

$$\begin{aligned} \left\langle \frac{s_\theta}{\|s_\theta\|}, \nabla V_i^{\theta_1}(\rho) \right\rangle &= \left\langle \frac{s_\theta}{\|s_\theta\|}, \nabla V_i^\theta(\rho) \right\rangle + \left\langle \frac{s_\theta}{\|s_\theta\|}, \nabla V_i^{\theta_1}(\rho) - \nabla V_i^\theta(\rho) \right\rangle \\ &\geq \ell_\theta - \left\| \nabla V_i^{\theta_1}(\rho) - \nabla V_i^\theta(\rho) \right\| \\ &\geq \ell_\theta - M\|\theta - \theta_1\| \\ &\geq \frac{\ell_\theta}{2}. \end{aligned}$$

We further set ν_2 as

$$\nu_2 := \inf \left\{ V_i^\theta(\rho), i \in [m] \mid \theta \in \Theta \setminus \bigcup_{\theta \in \Theta'} \mathcal{R}(\theta) \right\}.$$

Notice that $\nu_2 > 0$, we set $\nu_{emf} = \min\{\nu_1, \frac{\nu_2}{2}\} > 0$. Therefore, for any $\theta_2 \in \Theta$, we have $\mathbf{B}_{\nu_{emf}}(\theta_2) \subset \bigcup_{\theta \in \Theta'} \mathbf{B}_0(\theta)$, since θ_2 must be close to one of the θ in Θ' since $V_i^\theta(\rho)$ is a continuous function for $i \in [m]$. Consequently, we have

$$\left\langle \frac{s_\theta}{\|s_\theta\|}, \nabla V_i^\theta(\rho) \right\rangle \geq \ell.$$

Therefore, Assumption 4.2 holds with such ν_{emf} and ℓ . □

Corollary B.3. *If Θ' is compact, Assumptions 3.1 and B.1 imply Assumption 4.2.*

Proof of Corollary B.3. From Proposition B.2, it is sufficient to prove that ℓ, ν_1 defined in the above proposition are positive. We will prove this by contradiction.

Let us begin by proving $\ell > 0$ by contradiction. Suppose $\ell = 0$. This implies the existence of a series of points $\{\theta_i\}_{i=1}^\infty$ in Θ' such that

$$\lim_{i \rightarrow \infty} \ell_{\theta_i} = 0.$$

Using the definition of ℓ_{θ_i} , we have

$$\lim_{i \rightarrow \infty} \sum_{j \in \mathbf{B}_0(\theta_i)} \left\| \nabla_\theta V_j^{\theta_i}(\rho) \right\| = 0.$$

Let $k_j := \sum_{i=1}^{\infty} \mathbf{1}_{j \in \mathbf{B}_0(\theta_i)}$ for $j \in [m]$. Since $\sum_{j=1}^m k_j = \sum_{i=1}^{\infty} \sum_{j=1}^m \mathbf{1}_{j \in \mathbf{B}_0(\theta_i)} = \sum_{i=1}^{\infty} |\mathbf{B}_0(\theta_i)| = \infty$, there exists a $j \in [m]$ such that $k_j = \infty$. We choose a subset of indices $\{i\}$ as $\{i_j\}$ such that $j \in \mathbf{B}_0(\theta_{i_j})$. Then, we have

$$\lim_{i_j \rightarrow \infty} \|\nabla_{\theta} V_j^{\theta_{i_j}}(\rho)\| = 0.$$

Since Θ' is a compact set, there exists a θ_{lim} such that $\lim_{i_j \rightarrow \infty} \theta_{i_j} = \theta_{lim}$. For such θ_{lim} , we have

$$\left\| \nabla_{\theta} V_j^{\theta_{lim}}(\rho) \right\| = 0 \quad \text{and} \quad V_j^{\theta_{lim}}(\rho) = 0.$$

However, this contradicts Assumption B.1. The same analysis applies for $\nu_1 = 0$, leading to a contradiction as well. \square

Discussion Notice that if there is only one constraint in the CMDP, this assumption is satisfied trivially. For the more general case with more than one constraint, in this section, we proved that the MFCQ assumption, which is weaker than Assumption 4.2, can ensure the satisfaction of Assumption 4.2 if our feasible policy parameterization set $\{\theta \mid V_i^{\theta}(\rho) \geq 0\}$ is compact. This compactness can be achieved through certain policy parameterizations. For example, we can use direct parameterization or limit the policy parameterized set to $\{\pi_{\theta} \mid \theta \in \mathbf{K}\}$, where \mathbf{K} is a compact set in \mathbf{R}^d .

B.2 Impact of Assumption 4.2

The extended MFCQ assumption ensures that for every point θ lying on the boundary, there exists a trajectory that guides θ away from the boundary. Essentially, the extended MFCQ assumption prevents the algorithm from becoming trapped at the boundary, assuming a reasonable policy exists to guide the system back within the feasible region. When the CMDP structure lacks this property, safe exploration becomes more challenging to achieve. To illustrate this point, we provide the following theorem.

Theorem B.4. *Let Assumptions 3.1, 4.1, 4.4, and 4.8 hold, and set $\eta < \nu_{emf}$, $H = \tilde{\mathcal{O}}(\frac{1}{\eta})$ and $n = \mathcal{O}\left(\exp \frac{4}{\eta} \ln \frac{1}{\delta}\right)$ and $T = \mathcal{O}\left(\exp \frac{2}{\eta}\right)$. After T iterations of the Algorithm 1, the following holds:*

1. $\mathbb{P}\left(\forall t \in [T], \min_{i \in [m]} V_i^{\theta_t}(\rho) \geq \Omega\left(\exp \frac{-1}{\eta(1-\gamma)}\right)\right) \geq 1 - mT\delta.$
2. *We can bound the regret of the objective function with a probability of at least $1 - mT\delta$ as follows:*

$$\frac{1}{T} \sum_{i=0}^{T-1} \left(V_0^{\pi^*}(\rho) - V_0^{\theta_t}(\rho) \right) \leq \mathcal{O}(\eta) + \mathcal{O}\left(\sqrt{\varepsilon_{bias}} \exp \frac{1}{\eta}\right).$$

We provide the proof of Theorem B.4 in the following section. This theorem illustrates that Algorithm 1 requires high sample complexity to achieve safe exploration and does not guarantee the optimality of the iterates simultaneously if Assumption 4.2 is not satisfied. Specifically, without Assumption 4.2, the LB-SGD iterations could be as close as $\mathcal{O}\left(\exp \frac{-1}{\eta}\right)$ to the boundary as shown in Theorem Property 1.

To illustrate this, we provide an example demonstrating that without Assumption 4.2, LB-SGD iterations might approach the boundary at a level of $\mathcal{O}(\eta^{2k+1})$ for any $k \in \mathcal{N}$. This closeness to the boundary leads to slower convergence due to smaller stepsizes and increased sample complexity as the iterates approach the boundary to maintain small bias and low variance of the estimators $\hat{V}_i^{\theta}(\rho)$ and $\hat{\nabla} V_i^{\theta}(\rho)$. Meanwhile, we cannot guarantee the optimality of the iterates as it magnifies the transfer error ε_{bias} by $\exp \frac{1}{\eta}$.

Example B.5. *We consider the problem as follows:*

$$\begin{aligned} \max_{x,y} \quad & -y \\ \text{s.t.} \quad & y^{2k+1} + x \geq 0, \\ & y^{2k+1} - 2x \geq 0, \end{aligned}$$

where $k \in \mathcal{N}$.

We can verify that the above example does not satisfy the MFCQ assumption since the constraint gradients at the point $(0, 0)$ are opposite to each other. Next, we define the log barrier function as follows:

$$B_\eta(x, y) = -y + \eta \log(y^{2k+1} + x) + \eta \log(y^{2k+1} - 2x).$$

We compute that the optimal solution for the original problem is $(x^*, y^*) = (0, 0)$, and the optimal solution for the log barrier function is $(x_\eta^*, y_\eta^*) = (-2^{2k-1}(2k+1)^{2k+1}\eta^{2k+1}, (4k+2)\eta)$. When implementing gradient ascent on the log barrier function, starting from the point $(x_0, y_0) = (0, 1)$, the trajectory follows a curve from (x_0, y_0) to the optimal point (x_η^*, y_η^*) . Meanwhile, (x_η^*, y_η^*) is at a distance of $\mathcal{O}(\eta^{2k+1})$ from the boundary. Consequently, the iterates can approach the boundary within a distance of $\mathcal{O}(\eta^{2k+1})$.

B.2.1 Proof of Theorem B.4

To prove Theorem B.4, we follow a similar structure to the proof of Theorem 4.12. For safe exploration analysis (as outlined in Lemma B.6), we make use of Assumptions 4.1 and the boundedness of the value function $V_i^\theta(\rho)$ to establish a lower bound on the distance of the iterates from the boundary and the stepsize γ_t .

By employing a stochastic gradient ascent method with an appropriate stepsize γ_t , we ensure convergence to the stationary point of the log barrier function. Using Lemma 4.10, we can measure the optimality of this stationary point.

Lemma B.6. *Let Assumptions 3.1 and 4.1 hold, and we set $n = \mathcal{O}\left(\exp\frac{4}{\eta} \ln \frac{1}{\delta}\right)$ and $H = \tilde{\mathcal{O}}\left(\frac{1}{\eta}\right)$. Then, by running T iterations of the LB-SGD algorithm, we obtain*

$$\mathbb{P}\left\{\forall t \in [T], \min_{i \in [m]} V_i^{\theta_t}(\rho) \geq c_1, \gamma_t \geq C_1 \text{ and } \|\Delta_t\| \geq \frac{\eta}{4}\right\} \geq 1 - mT\delta,$$

where $c_1 := \nu_s^m(1 - \gamma)^{m-1} \exp\frac{-1}{\eta(1-\gamma)}$, $C_1 := c_1^2 \min\left\{\frac{3}{(\sqrt{6c_1 M} + 4L)(L + m\eta L)}, \frac{1}{c_1^2 M + 20m\eta c_1 M + 32m\eta L^2}\right\}$ and $\Delta_t := \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) - \nabla_\theta B_\eta^{\theta_t}(\rho)$.

We first employ the sub-Gaussian tail bounds of the estimators $\hat{\nabla}_\theta V_i^{\theta_t}(\rho)$ and $\hat{V}_i^{\theta_t}(\rho)$ to establish concentration bounds for $\|\hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) - \nabla_\theta B_\eta^{\theta_t}(\rho)\|$ in Lemma 3.4. Additionally, as we apply the stochastic gradient ascent method with sufficient samples, $B_\eta^{\theta_t}(\rho)$ is non-decreasing. Combined with the boundedness of objective and constraint functions, we can establish a lower bound of $\mathcal{O}\left(\exp\frac{-1}{\eta}\right)$ for $V_i^{\theta_t}(\rho)$ with high probability.

Proof of Lemma D.4. First, we prove the lower bound of the value functions $V_i^\theta(\rho), i \in \{0, \dots, m\}$. If we set

$$\sigma^0(n) \leq \min \left\{ \frac{1}{16 \left(\sum_{i=1}^m \frac{L}{\alpha_i(t) \hat{\alpha}_i(t)} \right) \sqrt{\ln \frac{2}{\delta}}} \right\}, \sigma^1(n) \leq \min \left\{ \frac{\eta}{16 \left(1 + \sum_{i=1}^m \frac{\eta}{\hat{\alpha}_i(t)} \right) \sqrt{\ln \frac{e^{\frac{1}{\delta}}}{\delta}}} \right\},$$

$$b^0(H) \leq \min \left\{ \frac{1}{16 \left(\sum_{i=1}^m \frac{L}{\alpha_i(t) \hat{\alpha}_i(t)} \right)} \right\}, b^1(H) \leq \min \left\{ \frac{\eta}{16 \left(1 + \sum_{i=1}^m \frac{\eta}{\hat{\alpha}_i(t)} \right)} \right\}.$$

By Lemma 3.4, we have

$$\mathbb{P}(\Delta_t \leq \frac{\eta}{4}) \geq 1 - \delta.$$

Due to the choice of stepsize, $\mathbb{P}(\gamma_t \leq \frac{1}{M_t}) \geq 1 - \delta$, where M_t is the local smoothness constant of the log barrier function $B_\eta^\theta(\rho)$. Then, we can bound $B_\eta^{\theta_{t+1}}(\rho) - B_\eta^{\theta_t}(\rho)$ with probability at least $1 - \delta$ as follows:

$$B_\eta^{\theta_{t+1}}(\rho) - B_\eta^{\theta_t}(\rho) \geq \gamma_t \left\langle \nabla_\theta B_\eta^{\theta_t}(\rho), \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\rangle - \frac{M_t \gamma_t^2}{2} \left\| \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\|^2$$

$$\begin{aligned}
 &\geq \gamma_t \left\langle \nabla_{\theta} B_{\eta}^{\theta_t}(\rho), \hat{\nabla}_{\theta} B_{\eta}^{\theta_t}(\rho) \right\rangle - \frac{\gamma_t}{2} \left\| \hat{\nabla}_{\theta} B_{\eta}^{\theta_t}(\rho) \right\|^2 \\
 &= \gamma_t \left\langle \nabla_{\theta} B_{\eta}^{\theta_t}(\rho), \left(\hat{\nabla}_{\theta} B_{\eta}^{\theta_t}(\rho) - \nabla_{\theta} B_{\eta}^{\theta_t}(\rho) \right) + \nabla_{\theta} B_{\eta}^{\theta_t}(\rho) \right\rangle - \frac{\gamma_t}{2} \left\| \left(\hat{\nabla}_{\theta} B_{\eta}^{\theta_t}(\rho) - \nabla_{\theta} B_{\eta}^{\theta_t}(\rho) \right) + \nabla_{\theta} B_{\eta}^{\theta_t}(\rho) \right\|^2 \\
 &= \frac{\gamma_t}{2} \left\| \nabla_{\theta} B_{\eta}^{\theta_t}(\rho) \right\|^2 - \frac{\gamma_t}{2} \left\| \Delta_t \right\|^2.
 \end{aligned} \tag{6}$$

Before the break in algorithm 1 line 5, we obtain that $\left\| \nabla_{\theta} B_{\eta}^{\theta_t}(\rho) \right\| \geq \frac{\eta}{4}$ since $\left\| \Delta_t \right\| \leq \frac{\eta}{4}$. Therefore, $B_{\eta}^{\theta_{t+1}}(\rho) \geq B_{\eta}^{\theta_t}(\rho)$ which leads to $B_{\eta}^{\theta_t}(\rho) \geq B_{\eta}^{\theta_0}(\rho)$. Then,

$$\begin{aligned}
 V_0^{\theta_t}(\rho) + \eta \sum_{i \in [m]} \log V_i^{\theta_t}(\rho) &\geq V_0^{\theta_0}(\rho) + \eta \sum_{i \in [m]} \log V_i^{\theta_0}(\rho), \\
 \log V_j^{\theta_t} &\geq \frac{V_0^{\theta_0}(\rho) - V_0^{\theta_t}(\rho)}{\eta} + \sum_{i \in [m]} \log V_i^{\theta_0}(\rho) - \sum_{\substack{i \in [m] \\ i \neq j}} \log V_i^{\theta_t}, \\
 \log V_j^{\theta_t} &\geq \frac{-1}{(1-\gamma)} + m \log \nu_s + (m-1) \log(1-\gamma),
 \end{aligned}$$

where the last inequality comes from the boundness of the value functions $V_i^{\theta}(\rho), i \in \{0, \dots, m\}$. Therefore,

$$\min_{i \in [m]} V_i^{\theta_t}(\rho) \geq c_1, \forall t \in \{0, \dots, T\}, \tag{7}$$

where $c_1 := \nu_s^m (1-\gamma)^{m-1} \exp \frac{-1}{\eta(1-\gamma)} = \mathcal{O} \left(\exp \frac{-1}{\eta} \right)$. For each $i \in [m]$, if $\sigma^0(n) \leq \frac{\alpha_i(t)}{8\sqrt{\ln \frac{2}{\delta}}}$ and $b^0(H) \leq \frac{\alpha_i(t)}{8}$, we have $\mathbb{P} \left(\frac{3\alpha_i(t)}{4} \leq \hat{\alpha}_i(t) \right) \geq 1 - \delta$ using the sub-Gaussian bound in Proposition 3.3. Therefore, we need to bound the variances and biases as follows to make sure $\mathbb{P}(\Delta_t \leq \frac{\eta}{4}) \geq 1 - \delta$.

$$\begin{aligned}
 \sigma^0(n) &\leq \min \left\{ \frac{3}{64 \left(\sum_{i=1}^m \frac{L}{(\alpha_i(t))^2} \right) \sqrt{\ln \frac{2}{\delta}}}, \frac{\alpha_i(t)}{8\sqrt{\ln \frac{2}{\delta}}} \right\}, \sigma^1(n) \leq \min \left\{ \frac{3\eta}{64 \left(1 + \sum_{i=1}^m \frac{\eta}{\alpha_i(t)} \right) \sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}}}, \right\}, \\
 b^0(H) &\leq \min \left\{ \frac{3}{64 \left(\sum_{i=1}^m \frac{L}{(\alpha_i(t))^2} \right)}, \frac{\alpha_i(t)}{8} \right\}, b^1(H) \leq \min \left\{ \frac{3\eta}{64 \left(1 + \sum_{i=1}^m \frac{\eta}{\alpha_i(t)} \right)} \right\}.
 \end{aligned}$$

By the lower bound in (7) and the Proposition 3.3, the number of trajectories n and the truncated horizon H need to be set as follows:

$$H = \tilde{\mathcal{O}} \left(\frac{1}{\eta} \right), n = \mathcal{O} \left(\exp \frac{4}{\eta} \ln \frac{1}{\delta} \right).$$

Meanwhile, we can further lower bound γ_t which is

$$\gamma_t := \min \left\{ \min_{i \in [m]} \left\{ \frac{\underline{\alpha}_i(t)}{\sqrt{M_i \underline{\alpha}_i(t)} + 2|\bar{\beta}_i(t)|} \right\} \frac{1}{\left\| \hat{\nabla}_{\theta} B_{\eta}^{\theta_t}(\rho) \right\|}, \frac{1}{M + \sum_{i=1}^m \frac{10M\eta}{\underline{\alpha}_i^i} + 8\eta \sum_{i=1}^m \frac{(\bar{\beta}_i^i)^2}{(\underline{\alpha}_i^i)^2}} \right\}.$$

Since $\sigma^0(n) \leq \frac{\alpha_i(t)}{8\sqrt{\ln \frac{2}{\delta}}}$ and $b^0(H) \leq \frac{\alpha_i(t)}{8}$, we have $\mathbb{P} \left(\frac{\alpha_i(t)}{2} \leq \underline{\alpha}_i(t) \leq \frac{3}{2\alpha_i(t)} \right) \geq 1 - \delta$ using the sub-Gaussian bound in Proposition 3.3. Together with (7), we have

$$\mathbb{P}(\gamma_t \geq C_1) \geq 1 - \delta,$$

where C_1 is defined as

$$C_1 := c_1^2 \min \left\{ \frac{3}{(\sqrt{6c_1 M} + 4L)(L + m\eta L)}, \frac{1}{c_1^2 M + 20m\eta c_1 M + 32m\eta L^2} \right\}$$

which is at the level of $\mathcal{O} \left(\exp \frac{-2}{\eta} \right)$. \square

Proof of Theorem B.4. We set the values for n , H , and η to satisfy the conditions outlined in Lemma B.6. Due to our choice of stepsize, we have $\mathbb{P}\left(\gamma_t \leq \frac{1}{M_t}\right) \geq 1 - \delta$, where M_t represents the local smoothness constant of the log barrier function $B_\eta^\theta(\rho)$. With this, we can bound $B_\eta^{\theta_{t+1}}(\rho) - B_\eta^{\theta_t}(\rho)$ with a probability of at least $1 - \delta$ as

$$\begin{aligned} B_\eta^{\theta_{t+1}}(\rho) - B_\eta^{\theta_t}(\rho) &\geq \gamma_t \left\langle \nabla_\theta B_\eta^{\theta_t}(\rho), \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\rangle - \frac{M_t \gamma_t^2}{2} \left\| \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\|^2 \\ &\geq \gamma_t \left\langle \nabla_\theta B_\eta^{\theta_t}(\rho), \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\rangle - \frac{\gamma_t}{2} \left\| \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\|^2 \\ &= \gamma_t \left\langle \nabla_\theta B_\eta^{\theta_t}(\rho), \Delta_t + \nabla_\theta B_\eta^{\theta_t}(\rho) \right\rangle - \frac{\gamma_t}{2} \left\| \Delta_t + \nabla_\theta B_\eta^{\theta_t}(\rho) \right\|^2 \\ &= \frac{\gamma_t}{2} \left\| \nabla_\theta B_\eta^{\theta_t}(\rho) \right\|^2 - \frac{\gamma_t}{2} \left\| \Delta_t \right\|^2. \end{aligned} \quad (8)$$

We divide the analysis into two cases based on the **if condition** in algorithm 1 line 5.

Case 1: If $\left\| \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\| \geq \frac{\eta}{2}$, then $\left\| \nabla_\theta B_\eta^{\theta_t}(\rho) \right\| \geq \frac{\eta}{4}$ because $\left\| \Delta_t \right\| \leq \frac{\eta}{4}$ by Proposition 4.3. We can bound (8) as

$$B_\eta^{\theta_{t+1}}(\rho) - B_\eta^{\theta_t}(\rho) \geq \frac{C_1 \eta}{8} \left\| \nabla_\theta B_\eta^{\theta_t}(\rho) \right\| - \frac{C_1 \eta^2}{32}, \quad (9)$$

where we plug in $\gamma_t \geq C_1$ in the last step. Summing inequality (9) from $t = 0$ to $t = T - 1$, we have

$$\begin{aligned} B_\eta^{\theta_T}(\rho) - B_\eta^{\theta_0}(\rho) + \frac{C_1 \eta^2 T}{32} &\geq \sum_{t=0}^{T-1} \frac{C_1 \eta}{8} \left\| \nabla_\theta B_\eta^{\theta_t}(\rho) \right\| \\ \frac{8(B_\eta^{\theta_T}(\rho) - B_\eta^{\theta_0}(\rho))}{C_1 \eta T} + \frac{\eta}{4} &\geq \frac{1}{T} \sum_{t=0}^{T-1} \left\| \nabla_\theta B_\eta^{\theta_t}(\rho) \right\|. \end{aligned}$$

Since the value functions $V_i^\theta(\rho)$ are upper bounded by $\frac{1}{1-\gamma}$, we can further bound the above inequality as

$$\frac{1}{T} \sum_{t=0}^{T-1} \left\| \nabla_\theta B_\eta^{\theta_t}(\rho) \right\| \leq \frac{8 \left(\frac{1}{1-\gamma} - m\eta \log(1-\gamma) - B_\eta^{\theta_0}(\rho) \right)}{C_1 \eta T} + \frac{\eta}{4}$$

By setting $T = \mathcal{O}\left(\frac{1}{C_1 \eta^2}\right) = \mathcal{O}\left(\exp \frac{2}{\eta}\right)$, we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \left\| \nabla_\theta B_\eta^{\theta_t}(\rho) \right\| \leq \mathcal{O}(\eta) \quad (10)$$

Case 2: If $\left\| \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\| \leq \frac{\eta}{2}$, we have

$$\left\| \nabla_\theta B_\eta^{\theta_t}(\rho) \right\| \leq \left\| \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\| + \left\| \Delta_t \right\| \leq \frac{3\eta}{4}. \quad (11)$$

Applying Lemma 4.10 on (10) and (11), we have

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \left(V_0^{\pi^*}(\rho) - V_0^{\theta_t}(\rho) \right) &\leq m\eta + \frac{1}{T} \sum_{t=0}^{T-1} \left(\frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} \left(1 + \sum_{i=1}^m \frac{\eta}{V_i^{\theta_t}(\rho)} \right) \right) + \mathcal{O}(\eta) \\ &= \mathcal{O}(\eta) + \mathcal{O}\left(\sqrt{\varepsilon_{bias}} \exp \frac{1}{\eta}\right), \end{aligned}$$

where we use $V_i^\theta(\rho) \geq c_1 = \mathcal{O}(\exp \frac{-1}{\eta})$ in the last step. Therefore, we can conclude the following: after $T = \mathcal{O}\left(\exp \frac{2}{\varepsilon}\right)$ iterations of the LB-SGD algorithm with $\eta = \varepsilon$, we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \left(V_0^{\pi^*}(\rho) - V_0^{\theta_t}(\rho) \right) \leq \mathcal{O}(\varepsilon) + \mathcal{O}\left(\sqrt{\varepsilon_{bias}} \exp \frac{1}{\varepsilon}\right),$$

while safe exploration is ensured with a probability of at least $1 - mT\delta$. \square

C POLICY PARAMETERIZATION

In this section, we delve deeper into widely accepted assumptions and derive conditions under which they may or may not hold. Specifically, we investigate the relationship between the (relaxed) Fisher non-degenerate assumption and the bounded transfer error assumption, especially concerning commonly used tabular policy parameterizations. These parameterizations include softmax, log-linear, and neural softmax policies. These policies are defined as follows:

$$\pi_{\theta}(a|s) = \frac{\exp f_{\theta}(s, a)}{\sum_{a' \in \mathcal{A}} \exp f_{\theta}(s, a')}.$$

1. For softmax policy, $f_{\theta}(s, a) = \theta(s, a)$.
2. For log-linear policy, $f_{\theta}(s, a) = \theta^T \cdot \phi(s, a)$, with $\theta \in \mathbb{R}^d$ and $\phi(s, a) \in \mathbb{R}^d$.
3. For neural softmax policy, $f_{\theta}(s, a)$ is parameterized using neural networks.

We first introduce two critical concepts related to policy parameterization: the ε -deterministic policy and richness in the policy parameterization.

Definition C.1 (ε -deterministic Policy). *We define a policy, π_{θ} , as an ε -deterministic policy if $\pi_{\theta} \in \Pi_{\varepsilon} := \{\pi_{\theta} \mid \text{for every state } s, \text{ there exists } a_{i_s} \in \mathcal{A} \text{ such that } \pi_{\theta}(a_{i_s}|s) \geq 1 - \varepsilon\}$. As ε approaches zero, the policy is said to approach a deterministic policy.*

Definition C.2 (Richness of Policy Parameterization). *Define Π as the closure of all stochastically parameterized policies, denoted by $\mathcal{CI}\{\pi_{\theta} \mid \theta \in \mathbb{R}^d\}$. If we have another policy parameterization Π' , we say Π' is a richer parameterization compared to Π if $\Pi' \subsetneq \Pi$.*

In the following section, we examine the commonly employed assumptions of (relaxed) Fisher non-degeneracy and bounded transfer error with tabular policy parameterizations as defined above. In Section C.1, we demonstrate that softmax parameterization cannot satisfy Fisher non-degeneracy, and softmax, log-linear, and neural softmax parameterizations fail to meet relaxed Fisher non-degeneracy as the policy approaches a deterministic policy. This reveals the relationship between the relaxed Fisher non-degenerate assumption and the exploration rate of the policy. In Section C.2, we prove that the transfer error bound can be reduced by increasing the richness of the policy set for log-linear and neural softmax policy parameterizations, indicating the relationship between the transfer error bound assumption and the richness of the policy parameterization.

C.1 Discussion on Assumption 4.4

In this section, we divide the analysis into two parts. In Section C.1.1, we prove that the softmax parameterization cannot satisfy the Fisher non-degeneracy assumption, and fails to satisfy relaxed Fisher non-degeneracy as the policy approaches determinism. Then, in Section C.1.2, we demonstrate that log-linear and neural softmax parameterizations also fail to satisfy relaxed Fisher non-degeneracy when the policy approaches a deterministic state. This section reveals the relationship between the relaxed Fisher non-degeneracy assumption and the exploration rate of the policy, highlighting the limitations of algorithms that depend on satisfying this relaxed assumption.

C.1.1 Softmax Parameterization

The authors of (Ding et al., 2022b) claim that the softmax parameterization fails to satisfy Fisher non-degeneracy when the policy approaches a deterministic policy. In this section, we prove a stronger version of this claim,

Proposition C.3. • *Softmax parameterization does not satisfy Fisher non-degeneracy.*

- *Softmax fails to satisfy relaxed Fisher non-degeneracy as the policy approaches determinism.*

Proof. The first property may seem obvious by (Metelli et al., 2022, Appendix), as the softmax parameterization requires $|S||A|$ parameters, which can result in rank deficiency in the Fisher information matrix due to the redundancy of $|S|$ parameters. For completeness, we still provide a formal proof of this statement below.

Note that the Fisher information matrix is computed as:

$$F^\theta(\rho) = \mathbb{E}_{(s,a) \sim d_\rho^\theta} [\nabla_\theta \log \pi_\theta(a|s) (\nabla_\theta \log \pi_\theta(a|s))^T]$$

Therefore, the image of the Fisher information matrix is the span of the linear space $\{\nabla_\theta \log \pi_\theta(a|s), \forall a \in \mathcal{A}, s \in \mathcal{S}\}$ if the state-action occupancy measure d_ρ^θ is non-zero for every state-action pair (s, a) .

For softmax parameterization, $\nabla_{\theta_{s'}} \log \pi_\theta(a|s)$ can be computed as:

$$\nabla_{\theta_{s'}} \log \pi_\theta(a|s) = \mathbf{1}_{s'=s} (e_a - \pi(\cdot|s)),$$

where $\theta_{s'} := \{\theta(s', a_1), \dots, \theta(s', a_{|\mathcal{A}|})\}$. Here, $e_a \in \mathbb{R}^{|\mathcal{A}|}$ is an elementary vector, with zeros everywhere except in the a th position, and $\pi(\cdot|s) := (\pi_\theta(a_1|s), \dots, \pi_\theta(a_{|\mathcal{A}|}|s))^T$. Therefore, $\{\nabla_{\theta_{s'}} \log \pi_\theta(a|s), \forall a \in \mathcal{A}\} = \{\mathbf{0}\}$ if $s' \neq s$ and

$$\{\nabla_{\theta_s} \log \pi_\theta(a|s), \forall a \in \mathcal{A}\} = \{e_a - (\pi(\cdot|s), \forall a \in \mathcal{A})\} = \{e_{a_1} - (\pi(\cdot|s), e_{a_1} - e_{a_i}, \forall i \in [|\mathcal{A}|])\},$$

where the first equality is written by definition and second equality is computed by $(e_a - (\pi(\cdot|s)) - (e_{a'} - (\pi(\cdot|s)))$. The rank of $\{\nabla_{\theta_s} \log \pi_\theta(a|s), \forall a \in \mathcal{A}\}$ is $|\mathcal{A}| - 1$ since the rank of $\{e_{a_1} - e_{a_i}, \forall i \in [|\mathcal{A}|]\}$ is $|\mathcal{A}| - 1$. Additionally, one vector $\mathbf{1} \in \mathbb{R}^{|\mathcal{A}|}$ is orthogonal to $\{\nabla_{\theta_s} \log \pi_\theta(a|s), \forall a \in \mathcal{A}\}$ due to:

$$\mathbf{1}^T (e_a - \pi(\cdot|s)) = 1 - \sum_{a \in \mathcal{A}} \pi(a|s) = 0.$$

In conclusion, $F^\theta(\rho)$ has rank at most $(|\mathcal{A}| - 1)|\mathcal{S}|$. This indicates that the Fisher information matrix is not a full-rank matrix, and therefore unable to satisfy the Fisher non-degeneracy.

To prove the second part of the proposition, we show that the trace of the Fisher information matrix, denoted as $\text{tr}(F^\theta(\rho))$, approaches zero as the policy approaches determinism. Consequently, the smallest non-zero eigenvalues of the Fisher information matrix also converge to zero, thereby fails to satisfy the relaxed Fisher non-degeneracy assumption.

We compute $\text{tr}(F^\theta(\rho))$ as follows:

$$\text{tr}(F^\theta(\rho)) = \sum_s d_\rho^\theta(s) \sum_{a \in \mathcal{A}} \pi(a|s) \left(\sum_{a' \in \mathcal{A}, a' \neq a} \pi(a'|s)^2 + (1 - \pi(a|s))^2 \right). \quad (12)$$

Given a policy $\pi_\theta \in \Pi_\varepsilon$, for any state $s \in \mathcal{S}$, there exists an action $a_{i_s} \in \mathcal{A}$ such that $\pi_\theta(a_{i_s}|s) \geq 1 - \varepsilon$. Therefore, we have

$$\sum_{a' \in \mathcal{A}, a' \neq a_{i_s}} \pi(a'|s)^2 + (1 - \pi(a_{i_s}|s))^2 \leq \left(\sum_{a' \in \mathcal{A}, a' \neq a_{i_s}} \pi(a'|s) + 1 - \pi(a_{i_s}|s) \right)^2 \leq 4\varepsilon^2.$$

Plugging the above inequality into inequality (12), we have

$$\begin{aligned} \text{tr}(F^\theta(\rho)) &= \sum_s d_\rho^\theta(s) \sum_{a \in \mathcal{A}} \pi(a|s) \left(\sum_{a' \in \mathcal{A}, a' \neq a} \pi(a'|s)^2 + (1 - \pi(a|s))^2 \right) \\ &\leq \sum_s d_\rho^\theta(s) \sum_{a \in \mathcal{A}, a \neq a_{i_s}} \pi(a|s) \left(\sum_{a' \in \mathcal{A}, a' \neq a} \pi(a'|s)^2 + (1 - \pi(a|s))^2 \right) + 4\varepsilon^2 \sum_s d_\rho^\theta(s) \pi(a_{i_s}|s) \\ &\leq \sum_s d_\rho^\theta(s) \sum_{a \in \mathcal{A}, a \neq a_{i_s}} \pi(a|s) \left(\sum_{a' \in \mathcal{A}, a' \neq a} \pi(a'|s)^2 + (1 - \pi(a|s))^2 \right) + 4\varepsilon^2 \\ &\leq \sum_s d_\rho^\theta(s) \sum_{a \in \mathcal{A}, a \neq a_{i_s}} \pi(a|s) \left(\sum_{a' \in \mathcal{A}, a' \neq a} \pi(a'|s) + 1 - \pi(a|s) \right)^2 + 4\varepsilon^2 \end{aligned}$$

$$\begin{aligned}
 &\leq 4 \sum_s d_\rho^\theta(s) \sum_{a \in \mathcal{A}, a \neq a_{i_s}} \pi(a|s) + 4\varepsilon^2 \\
 &\leq 4\varepsilon + 4\varepsilon^2.
 \end{aligned}$$

Therefore, when a policy π_θ approaches a deterministic policy, the trace of Fisher information matrix approaches 0. \square

C.1.2 Log-Linear and Neural Softmax Parameterizations

It is known that for MDPs, the optimal policy can be deterministic. For CMDPs, the optimal policy can be deterministic if none of the constraints are active. Furthermore, during the implementation of policy gradient-based algorithms, the algorithm can be trapped at a stationary point, which can be deterministic. While log-linear and neural softmax parameterizations are commonly employed for discrete state and action spaces, this section reveals their failure to exhibit relaxed Fisher non-degeneracy when approaching a deterministic policy.

Proposition C.4. *Let $\|\nabla_\theta f_\theta(s, a)\| \leq M$ hold for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, Log-linear and neural softmax parameterizations fail to meet Assumption 4.4 as the policy approaches a deterministic policy.*

Proof. For a policy $\pi_\theta \in \Pi_\varepsilon$, which is ε -close to a deterministic policy, we first compute $\nabla_\theta \log \pi_\theta(a|s)$ as follows:

$$\nabla_\theta \log \pi_\theta(a|s) = \nabla f_\theta(s, a) - \sum_{a' \in \mathcal{A}} \nabla f_\theta(s, a') \pi_\theta(a'|s) = \sum_{a' \in \mathcal{A}, a' \neq a} (\nabla f_\theta(s, a) - \nabla f_\theta(s, a')) \pi_\theta(a'|s).$$

To satisfy Assumption 3.1 for log-linear and neural softmax parameterizations, it is commonly assumed that $\|\nabla f_\theta(s, a)\| \leq M$ for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$. Under this condition, we can bound $\nabla_\theta \log \pi_\theta(a_i|s)$ into two cases. For $i = i_s$, we know that $\sum_{a' \in \mathcal{A}, a' \neq a_{i_s}} \pi_\theta(a'|s) \leq \varepsilon$, therefore

$$\|\nabla_\theta \log \pi_\theta(a_{i_s}|s)\| = \left\| \sum_{a' \in \mathcal{A}, a' \neq a_{i_s}} (\nabla f_\theta(a_i|s) - \nabla f_\theta(a'|s)) \pi(a'|s) \right\| \leq 2M\varepsilon.$$

For $i \neq i_s$, we know that $\sum_{a' \in \mathcal{A}, a' \neq a_i} \pi_\theta(a'|s) \leq 1$, therefore

$$\|\nabla_\theta \log \pi_\theta(a_i|s)\| = \left\| \sum_{a' \in \mathcal{A}, a' \neq a_i} (\nabla f_\theta(a_i|s) - \nabla f_\theta(a'|s)) \pi(a'|s) \right\| \leq 2M.$$

Using the above two inequalities, we can upper bound the norm of the Fisher information matrix for the policy π_θ as

$$\begin{aligned}
 \|F^\theta(\rho)\| &= \left\| \sum_{s \in \mathcal{S}} d_\rho^\theta(s) \sum_{i=1}^{|\mathcal{A}|} \pi_\theta(a_{i_s}|s) \nabla_\theta \log \pi_\theta(a_{i_s}|s) (\nabla_\theta \log \pi_\theta(a_{i_s}|s))^T \right\| \\
 &\leq \left\| \sum_{s \in \mathcal{S}} d_\rho^\theta(s) \pi_\theta(a_{i_s}|s) \nabla_\theta \log \pi_\theta(a_{i_s}|s) (\nabla_\theta \log \pi_\theta(a_{i_s}|s))^T \right\| \\
 &\quad + \left\| \sum_{s \in \mathcal{S}} d_\rho^\theta(s) \sum_{a \neq a_{i_s}, a \in \mathcal{A}} \pi(a|s) \nabla \log \pi_\theta(a|s) (\nabla \log \pi_\theta(a|s))^T \right\| \\
 &\leq 4M^2\varepsilon^2 \left\| \sum_{s \in \mathcal{S}} d_\rho^\theta(s) \right\| + 4M^2 \left\| \sum_{s \in \mathcal{S}} d_\rho^\theta(s) \sum_{a_i \neq a_{i_s}, a_i \in \mathcal{A}} \pi_\theta(a_i|s) \right\| \\
 &\leq 4\varepsilon^2 M^2 + 4\varepsilon M^2.
 \end{aligned}$$

Therefore, when a policy π_θ approaches a deterministic policy, the norm of Fisher information matrix approaches 0. Then, we cannot find a positive constant μ_F such that the smallest non-zero eigenvalue of $F^\theta(\rho)$ is lower bounded by μ_F . \square

Discussion For continuous action space, Assumption 4.4 is satisfied by Gaussian policies $\pi_\theta(\cdot|s) = \mathcal{N}(\mu_\theta(s), \Sigma)$ when the parameterized mean $\mu_\theta(s)$ has a full row rank Jacobian and the covariance matrix Σ is fixed. Cauchy policies also satisfy these conditions.

For discrete action space, we proved that softmax, log-linear and neural softmax parameterizations fail to satisfy Assumption 4.4 as the policy becomes more deterministic. Building upon this fact, to ensure softmax, log-linear and neural softmax parameterizations satisfy relaxed Fisher non-degeneracy, we can constrain the parameter space to a compact set, preventing the policy from approaching determinism.

C.2 Discussion on Assumption 4.8

Inspired by (Wang et al., 2020), where the authors demonstrated that employing a rich two-layer neural-network parameterization can yield small ε_{bias} values, we generalize their result to demonstrate that increasing the richness of the policy set can lead to a reduction in the transfer error ε_{bias} for the log-linear and neural softmax policy parameterizations.

We consider the log-linear and neural softmax policy parameterization given by

$$\pi_\theta(a|s) = \frac{\exp f_\theta(s, a)}{\sum_{a' \in \mathcal{A}} \exp f_\theta(s, a')}.$$

Proposition C.5. *For log-linear and neural softmax policy parameterizations, increasing the dimension of the set $\{f'(s, a), s \in \mathcal{S}, a \in \mathcal{A}\}$ such that $\{f(s, a), s \in \mathcal{S}, a \in \mathcal{A}\} \subsetneq \{f'(s, a), s \in \mathcal{S}, a \in \mathcal{A}\}$ results in a richer parameterization. This richer parameterization leads to a decrease in the transfer error ε_{bias} .*

Proof. We first upper bound the transfer error as shown in (Agarwal et al., 2021, page 29):

$$L(\mu_i^*, \theta, d_\rho^{\pi^*}) \leq \max_{s \in \mathcal{S}} \frac{\sum_{a \in \mathcal{A}} d_\rho^{\pi^*}(s, a)}{(1 - \gamma)\rho(s)} L(\mu_i^*, \theta, d_\rho^\theta),$$

where $L(\mu_i^*, \theta, d_\rho^\theta) := \mathbb{E}_{(s, a) \sim d_\rho^\theta} [(A_i^\theta(s, a) - (1 - \gamma)\mu_i^{*T} \nabla_\theta \log \pi_\theta(a|s))^2]$ and $\mu_i^* := (F^\theta(\rho))^{-1} \nabla_\theta V_i^\theta(\rho)$. In the following, we demonstrate that $L(\mu_i^*, \theta, d_\rho^\theta)$ can be reduced due to the richer parametrization.

For every policy $\pi_\theta \in \Pi_f$, we set a vector-valued function $\mathbb{A} : \Pi_f \rightarrow \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as

$$\mathbb{A}_i^{\pi_\theta} := \begin{bmatrix} \sqrt{d_\rho^\pi(s_1, a_1)} A_i^\pi(s_1, a_1) \\ \vdots \\ \sqrt{d_\rho^\pi(s_{|\mathcal{S}|}, a_{|\mathcal{A}|})} A_i^\pi(s_{|\mathcal{S}|}, a_{|\mathcal{A}|}) \end{bmatrix}$$

and set \mathbb{B} as

$$\mathbb{B}_{\pi_\theta} := \begin{bmatrix} \sqrt{d_\rho^\pi(s_1, a_1)} \nabla_\theta \log \pi_\theta(a_1|s_1) \\ \vdots \\ \sqrt{d_\rho^\pi(s_{|\mathcal{S}|}, a_{|\mathcal{A}|})} \nabla_\theta \log \pi_\theta(a_{|\mathcal{A}|}|s_{|\mathcal{S}|}) \end{bmatrix}^T.$$

Notice that

$$\nabla_\theta \log \pi_\theta(a|s) = \sum_{a' \in \mathcal{A}, a' \neq a} (\nabla f_\theta(s, a) - \nabla f_\theta(s, a')) \pi_\theta(a'|s).$$

Then the column space of \mathbb{B}_{π_θ} is the span of $\{\nabla_\theta f(s, a)\}_{s \in \mathcal{S}, a \in \mathcal{A}}$. Using the above notations, we write Fisher information matrix as

$$F^{\pi_\theta}(\rho) = \mathbb{E}_{(s, a) \sim d_\rho^\pi} [\nabla_\theta \log \pi_\theta(a|s) (\nabla_\theta \log \pi_\theta(a|s))^T] = \mathbb{B}_{\pi_\theta} \mathbb{B}_{\pi_\theta}^T,$$

and the gradient of value function as

$$\nabla_\theta V_i^\theta(\rho) = \frac{1}{1 - \gamma} \mathbb{E}_{(s, a) \sim d_\rho^\pi} [\nabla_\theta \log \pi_\theta(a|s) A_i^\theta(s, a)] = \frac{1}{1 - \gamma} \mathbb{B}_{\pi_\theta} \mathbb{A}_i^{\pi_\theta}.$$

and $\mu_i^* = F^{\pi_\theta}(\rho)^{-1} \nabla_\theta V_i^\theta(\rho) = \frac{1}{1-\gamma} (\mathbb{B}_\pi \mathbb{B}_\pi^T)^{-1} \mathbb{B}_\pi \mathbb{A}_i^\pi$. Therefore, we have

$$\begin{aligned} L(\mu_i^*, \pi, d_\rho^{\pi_\theta}) &= \mathbb{E}_{(s,a) \sim d_\rho^\pi} \left[\left(A_i^{\pi_\theta}(s, a) - (1-\gamma) \mu_i^{*T} \nabla_\theta \log \pi_\theta(a|s) \right)^2 \right] \\ &= \sum_{s,a} \left(\sqrt{d_\rho^\pi(s, a)} (\mathbb{B}_{\pi_\theta} \mathbb{B}_{\pi_\theta}^T)^\dagger \mathbb{B}_{\pi_\theta} \mathbb{A}_i^{\pi_\theta} \nabla_\theta \log \pi_\theta(a|s) - \sqrt{d_\rho^\pi(s, a)} A_i^{\pi_\theta}(s, a) \right)^2 \\ &= \left\| \mathbb{A}_i^{\pi_\theta} - \mathbb{B}_{\pi_\theta}^T (\mathbb{B}_{\pi_\theta} \mathbb{B}_{\pi_\theta}^T)^\dagger \mathbb{B}_{\pi_\theta} \mathbb{A}_i^{\pi_\theta} \right\|_2^2 \\ &= \left\| \mathbb{A}_i^{\pi_\theta} - \mathbf{P}_{\mathbb{B}_{\pi_\theta}} \mathbb{A}_i^{\pi_\theta} \right\|_2^2, \end{aligned}$$

where $\mathbf{P}_{\mathbb{B}_{\pi_\theta}}$ is the orthogonal projection onto the row space of $\mathbb{B}_{\pi_\theta} \in \mathbb{R}^{d \times |\mathcal{S}||\mathcal{A}|}$. If we increase the dimension of $\{f'(s, a), s \in \mathcal{S}, a \in \mathcal{A}\}$ such that $\{f'(s, a), s \in \mathcal{S}, a \in \mathcal{A}\} \subsetneq \{f'(s, a), s \in \mathcal{S}, a \in \mathcal{A}\}$, it results in a more richer parameterization and a decrease in $L(\mu_i^*, \pi, d_\rho^\pi)$ since the column rank of \mathbb{B}_{π_θ} has increased. \square

D PROOFS

D.1 Proof of Proposition 3.3

To set up sub-Gaussian bounds for the gradient estimates in the RL case, we require the following lemma.

Lemma D.1. Vector Bernstein Inequality (Kohler and Lucchi, 2017, Lemma 18): Let $x_i \in \mathbb{R}^d$ be independent vector-valued random variables for $i \in [n]$. If there exist constants $B, \sigma \geq 0$ such that $\mathbb{E}[x_i] = 0$, $\|x_i\| \leq B$ and $\mathbb{E}[\|x_i\|^2] \leq \sigma^2$, the following inequality holds:

$$\mathbb{P} \left(\left\| \frac{\sum_{i=1}^n x_i}{n} \right\| \geq \varepsilon \right) \leq \exp \left(\frac{1}{4} - \frac{n\varepsilon^2}{8\sigma^2} \right),$$

where $\varepsilon \in (0, \frac{\sigma^2}{B})$.

Proof of Proposition 3.3. The first property has been proven in (Bai et al., 2023, Lemma 2) and (Yuan et al., 2022, Lemma 4.4). The bias bound $b^1(H)$ has been established in (Yuan et al., 2022, Lemma 4.5).

To establish the upper bound of $b^0(H)$ in the second property, we consider $i \in \{0, \dots, m\}$:

$$\left| \mathbb{E} [\hat{V}_i^\theta(\rho)] - V_i^\theta(\rho) \right| \leq \frac{1}{n} \left| \mathbb{E} \left[\sum_{j=1}^n \sum_{t=0}^{H-1} \gamma^t r_i(s_t^j, a_t^j) - \sum_{j=1}^n \sum_{t=0}^{\infty} \gamma^t r_i(s_t^j, a_t^j) \right] \right| \leq \sum_{t=H}^{\infty} \gamma^t = \frac{\gamma^H}{1-\gamma}.$$

Next, we prove that the value function estimator $\hat{V}_i^\theta(\rho)$ has a sub-Gaussian bound. For $i \in \{0, \dots, m\}$, we note that $\hat{V}_i^\theta(\rho)$ is bounded in the interval $[\frac{-1}{1-\gamma}, \frac{1}{1-\gamma}]$. Using Hoeffding's inequality, we have for any $\varepsilon > 0$,

$$\mathbb{P} \left(\left| \hat{V}_i^\theta(\rho) - \mathbb{E} [\hat{V}_i^\theta(\rho)] \right| \geq \varepsilon \right) \leq 2 \exp \left(-\frac{n\varepsilon^2(1-\gamma)^2}{2} \right).$$

We can rewrite the above inequality as:

$$\mathbb{P} \left(\left| \hat{V}_i^\theta(\rho) - \mathbb{E} [\hat{V}_i^\theta(\rho)] \right| \leq \sigma^0(n) \sqrt{\ln \frac{2}{\delta}} \right) \geq 1 - \delta,$$

for any $\delta \in (0, 1)$, where $\sigma^0(n) := \frac{\sqrt{2}}{\sqrt{n(1-\gamma)}}$.

Finally, we prove that the gradient estimator $\hat{\nabla} V_i^\theta(\rho)$ has a sub-Gaussian bound. From (Yuan et al., 2022, Lemma 4.2), we have

$$\text{Var} \left[\left(\hat{\nabla}_\theta V_i^\theta(\rho) \right)_j \right] \leq \frac{M_g^2}{(1-\gamma)^3}.$$

We conclude that $\left\| \left(\hat{\nabla}_\theta V_i^\theta(\rho) \right)_j \right\| \leq \frac{M_g}{(1-\gamma)^2}$ from (Xu et al., 2020, Proposition 4.2) and $\left\| \mathbb{E} \left(\hat{\nabla}_\theta V_i^\theta(\rho) \right)_j \right\| \leq \frac{M_g}{(1-\gamma)^2}$ from (Liu et al., 2020b, Lemma B.1). Therefore, we have

$$\left\| \left(\hat{\nabla}_\theta V_i^\theta(\rho) \right)_j - \mathbb{E} \left[\left(\hat{\nabla}_\theta V_i^\theta(\rho) \right)_j \right] \right\| \leq \frac{2M_g}{(1-\gamma)^2}.$$

We can apply Lemma D.1 to the estimator $\hat{\nabla}_\theta V_i^\theta(\rho) := \frac{1}{n} \sum_{j=1}^n \left(\hat{\nabla}_\theta V_i^\theta(\rho) \right)_j$, yielding the inequality

$$\mathbb{P} \left(\left\| \hat{\nabla}_\theta V_i^\theta(\rho) - \mathbb{E} \left[\hat{\nabla}_\theta V_i^\theta(\rho) \right] \right\| \geq \varepsilon \right) \leq \exp \left(\frac{1}{4} - \frac{n\varepsilon^2(1-\gamma)^3}{8M_g^2} \right),$$

where $\varepsilon \in (0, \frac{M_g}{2(1-\gamma)})$. Therefore, for any $\delta \in (0, 1)$ and $n \geq 8 \ln \frac{\exp \frac{1}{4}}{\delta}$, we have

$$\mathbb{P} \left(\left\| \hat{\nabla}_\theta V_i^\theta(\rho) - \mathbb{E} \left[\hat{\nabla}_\theta V_i^\theta(\rho) \right] \right\| \leq \sigma^1(n) \sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}} \right) \geq 1 - \delta,$$

where $\sigma^1(n) := \frac{2\sqrt{2}M_g}{\sqrt{n(1-\gamma)^{\frac{3}{2}}}}$. □

D.2 Proof of Lemma 3.4

For the rest of the paper, we first define $\Delta_t := \hat{\nabla}_\theta B_{\eta^t}^{\theta_t}(\rho) - \nabla_\theta B_{\eta^t}^{\theta_t}(\rho)$. Now, we start the proof of Lemma 3.4.

Proof of Lemma 3.4. we can bound $\|\Delta_t\|$ as

$$\begin{aligned} & \|\Delta_t\| \\ &= \left\| \nabla_\theta V_0^{\theta_t}(\rho) - \hat{\nabla}_\theta V_0^{\theta_t}(\rho) + \sum_{i=1}^m \left[\eta \frac{\hat{\nabla}_\theta V_i^{\theta_t}(\rho) - \nabla_\theta V_i^{\theta_t}(\rho)}{\hat{\alpha}_i(t)} + \eta \nabla_\theta V_i^{\theta_t}(\rho) \left(\frac{1}{\hat{\alpha}_i(t)} - \frac{1}{\alpha_i(t)} \right) \right] \right\| \end{aligned} \quad (13)$$

$$\leq \left\| \nabla_\theta V_0^{\theta_t}(\rho) - \hat{\nabla}_\theta V_0^{\theta_t}(\rho) \right\| + \sum_{i=1}^m \left[\eta \frac{\left\| \hat{\nabla}_\theta V_i^{\theta_t}(\rho) - \nabla_\theta V_i^{\theta_t}(\rho) \right\|}{\hat{\alpha}_i(t)} + \eta \left\| \nabla_\theta V_i^{\theta_t}(\rho) \right\| \left| \frac{1}{\hat{\alpha}_i(t)} - \frac{1}{\alpha_i(t)} \right| \right]. \quad (14)$$

Using the sub-Gaussian bound in Proposition 3.3, we have

$$\begin{aligned} & \mathbb{P} \left\{ |\hat{\alpha}_i(t) - \alpha_i(t)| \leq b^0(H) + \sigma^0(n) \sqrt{\ln \frac{2}{\delta}} \right\} \geq 1 - \delta, \\ & \mathbb{P} \left\{ \left\| \hat{\nabla}_\theta V_i^{\theta_t}(\rho) - \nabla_\theta V_i^{\theta_t}(\rho) \right\| \leq b^1(H) + \sigma^1(n) \sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}} \right\} \geq 1 - \delta. \end{aligned}$$

Also, we know $\|\nabla_\theta V_i^{\theta_t}(\rho)\| \leq L$ by Proposition 3.3. Combining these properties into inequality (14), we have:

$$\mathbb{P} \left(\left\| \hat{\nabla}_\theta B_{\eta^t}^{\theta_t}(\rho) - \nabla_\theta B_{\eta^t}^{\theta_t}(\rho) \right\| \leq \left(1 + \sum_{i=1}^m \frac{\eta}{\hat{V}_i^{\theta_t}(\rho)} \right) (b^1(H) + \sigma^1(n) \sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}}) + \sum_{i=1}^m \frac{\eta L}{\hat{V}_i^{\theta_t}(\rho) V_i^{\theta_t}(\rho)} (b^0(H) + \sigma^0(n) \sqrt{\ln \frac{2}{\delta}}) \right) \geq 1 - \delta.$$

Meanwhile, for each $i \in [m]$, we have $\mathbb{P} \left(\frac{3V_i^{\theta_t}(\rho)}{4} \leq \hat{V}_i^{\theta_t}(\rho) \right) \geq 1 - \delta$ if $\sigma^0(n) \leq \frac{V_i^{\theta_t}(\rho)}{8\sqrt{\ln \frac{2}{\delta}}}$ and $b^0(H) \leq \frac{V_i^{\theta_t}(\rho)}{8}$, using the sub-Gaussian bound from Proposition 3.3. Therefore, to bound the gradient estimation error $\|\hat{\nabla}_\theta B_{\eta^t}^{\theta_t}(\rho) - \nabla_\theta B_{\eta^t}^{\theta_t}(\rho)\|$ by ε with probability $1 - \delta$, we need to set $\sigma^0(n)$, $\sigma^1(n)$, $b^0(H)$ and $b^1(H)$ as follows:

$$\sigma^0(n) \leq \min_{i \in [m]} \left\{ \frac{V_i^{\theta_t}(\rho)}{8\sqrt{\ln \frac{2}{\delta}}}, \frac{3\varepsilon V_i^{\theta_t}(\rho)^2}{16m\eta L \sqrt{\ln \frac{2}{\delta}}} \right\}, \quad \sigma^1(n) \leq \min_{i \in [m]} \left\{ \frac{3\varepsilon V_i^{\theta_t}(\rho)}{4(4m\eta + 3V_i^{\theta_t}(\rho)) \sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}}} \right\},$$

$$b^0(H) \leq \min_{i \in [m]} \left\{ \frac{V_i^\theta(\rho)}{8}, \frac{3\varepsilon V_i^\theta(\rho)^2}{16m\eta L} \right\}, \quad b^1(H) \leq \min_{i \in [m]} \left\{ \frac{3\varepsilon V_i^\theta(\rho)}{4(4m\eta + 3V_i^\theta(\rho))} \right\}.$$

To satisfy the above inequality, we can choose n and H as follows using Proposition 3.3:

$$n = \mathcal{O} \left(\frac{\eta^2 \ln \frac{1}{\delta}}{(1-\gamma)^6 \varepsilon^2 \min_{i \in [m]} V_i^\theta(\rho)^4} \right), \quad H = \mathcal{O} \left(\ln \frac{1}{\varepsilon \min_{i \in [m]} V_i^\theta(\rho)} \right).$$

□

D.3 Determining Stepsize and Proof of Proposition 4.3

D.3.1 Determining Stepsize Using Local Smoothness

In this section, we discuss the choice of stepsize γ_t as introduced in Algorithm 1, specifically in line 10. For simplicity, we denote

$$\alpha_i(t) := V_i^{\theta_t}(\rho), \quad \hat{\alpha}_i(t) := \hat{V}_i^{\theta_t}(\rho), \quad \underline{\alpha}_i(t) := \hat{V}_i^{\theta_t}(\rho) - b^0(H) - \sigma^0(n) \sqrt{\ln \frac{2}{\delta}}.$$

Recall that the gradient of the log barrier function is defined as $\nabla_\theta B_\eta^\theta(\rho) = \nabla V_0^\theta(\rho) + \eta \sum_{i=1}^m \frac{\nabla V_i^\theta(\rho)}{V_i^\theta(\rho)}$. The log barrier function is non-smooth because the norm of the gradient exhibits unbounded growth when θ approaches the boundary of the feasible domain. However, it has been proven that within a small region around each iterate θ_t , denoted as $\mathcal{R}(\theta_t) = \{\theta \in \Theta \mid V_i^\theta(\rho) \geq \frac{V_i^{\theta_t}(\rho)}{2}, i \in [m]\}$, the log barrier function is M_t locally smooth around θ_t which is defined as

$$\|\nabla B_\eta^\theta(\rho) - \nabla B_\eta^{\theta'}(\rho)\| \leq M_t \|\theta - \theta'\|, \quad \forall \theta, \theta' \in \mathcal{R}(\theta_t).$$

Therefore, we first need to carefully choose the stepsize γ_t to ensure that the next iterate θ_{t+1} remains inside the region $\mathcal{R}(\theta_t)$. In (Usmanova et al., 2024, Lemma 3), the authors utilize the smoothness of the constraint functions to provide suggestions for choosing γ_t . We restate the lemma as follows:

Lemma D.2. *Under Assumption 3.1, if*

$$\gamma_t \leq \min_{i \in [m]} \left\{ \frac{\alpha_i(t)}{\sqrt{M\alpha_i(t)} + 2|\beta_i(t)|} \right\} \frac{1}{\|\hat{\nabla}_\theta B_\eta^{\theta_t}(\rho)\|},$$

we have

$$V_i^{\theta_{t+1}}(\rho) \geq \frac{V_i^{\theta_t}(\rho)}{2}.$$

Next, the authors proved the existence of such M_t (Usmanova et al., 2024, Lemma 2), and we restate it as follows:

Lemma D.3. *Let Assumption 3.1 hold, the log barrier function $B_\eta^\theta(\rho)$ is M_t locally smooth for $\theta \in \mathcal{R}(\theta_t)$, where*

$$M_t = M + \sum_{i=1}^m \frac{2M\eta}{\alpha_i(t)} + 4\eta \sum_{i=1}^m \frac{\left\langle \nabla_\theta V_i^{\theta_{t+1}}(\rho), \frac{\nabla_\theta B_\eta^{\theta_t}(\rho)}{\|\nabla_\theta B_\eta^{\theta_t}(\rho)\|} \right\rangle^2}{(\alpha_i(t))^2}.$$

Moreover, if $\gamma_t \leq \min_{i \in [m]} \left\{ \frac{\alpha_i(t)}{\sqrt{M\alpha_i(t)} + 2|\beta_i(t)|} \right\} \frac{1}{\|\hat{\nabla}_\theta B_\eta^{\theta_t}(\rho)\|}$, then

$$M_t = M + \sum_{i=1}^m \frac{10M\eta}{\alpha_i(t)} + 8\eta \sum_{i=1}^m \frac{(\beta_i(t))^2}{(\alpha_i(t))^2},$$

where $\beta_i(t) = \langle \nabla V_i^{\theta_t}(\rho), \frac{\nabla_\theta B_\eta^{\theta_t}(\rho)}{\|\nabla_\theta B_\eta^{\theta_t}(\rho)\|} \rangle$.

Using above lemmas, we set γ_t as

$$\gamma_t := \min \left\{ \min_{i \in [m]} \left\{ \frac{\alpha_i(t)}{\sqrt{M\alpha_i(t)} + 2|\beta_i(t)|} \right\}, \frac{1}{\|\hat{\nabla}_\theta B_{\eta^t}(\rho)\|}, \frac{1}{M_t} \right\}. \quad (15)$$

Therefore, we can ensure that the next iterate θ_{t+1} always remains within the region $\mathcal{R}(\theta_t)$, and prevent over-shooting by utilizing the local smoothness property. Notice that we only have estimates of $\alpha_i(t)$ and $\beta_i(t)$, therefore we replace $\alpha_i(t)$ with its lower bound of the estimates as $\underline{\alpha}_i(t)$ and $\beta_i(t)$ with its upper bound of the estimates as $\bar{\beta}_i(t) := \left| \langle \hat{\nabla}_\theta V_i^{\theta_t}(\rho), \frac{\hat{\nabla}_\theta B_{\eta^t}(\rho)}{\|\hat{\nabla}_\theta B_{\eta^t}(\rho)\|} \rangle \right| + \sigma^1(n) \sqrt{\ln \frac{e^{\frac{1}{\delta}}}{\delta}} + b^1(H)$, $i \in [m]$. Because of the sub-Gaussian bound established in Proposition 3.3, $\alpha_i(t)$ is lower bounded by $\underline{\alpha}_i(t)$ and $\beta_i(t)$ is upper by $\bar{\beta}_i(t)$ with high probability. Therefore, we choose the lower bound of (15) to set γ_t as

$$\gamma_t := \min \left\{ \min_{i \in [m]} \left\{ \frac{\underline{\alpha}_i(t)}{\sqrt{M\underline{\alpha}_i(t)} + 2|\bar{\beta}_i(t)|} \right\}, \frac{1}{\|\hat{\nabla}_\theta B_{\eta^t}(\rho)\|}, \frac{1}{M + \sum_{i=1}^m \frac{10M\eta}{\underline{\alpha}_i^t} + 8\eta \sum_{i=1}^m \frac{(\bar{\beta}_i^t)^2}{(\underline{\alpha}_i^t)^2}} \right\}.$$

D.3.2 Proof of Proposition 4.3

To prove Proposition 4.3, we establish the following more general lemma, which will be essential for the proof of Theorem 4.12.

Lemma D.4. *Define the events \mathcal{A} , \mathcal{B} , and \mathcal{C} as follows:*

$$\mathcal{A} := \left\{ \forall t \in [T], \min_{i \in [m]} V_i^{\theta_t}(\rho) \geq c\eta \right\}, \mathcal{B} := \left\{ \forall t \in [T], \min_{i \in [m]} \gamma_t \geq C\eta \right\}, \mathcal{C} := \left\{ \forall t \in [T], \|\Delta_t\| \geq \frac{\eta}{4} \right\},$$

where constants c and C are defined in Equations (24) and (25). Let Assumptions 3.1, 4.1, and 4.2 hold, and set $\eta \leq \nu_{emf}$, $n = \mathcal{O}\left(\frac{\ln \frac{1}{\delta}}{(1-\gamma)^{6+8m} \ell^{4m} \eta^4}\right)$ where $0 < \delta < 1$ and $H = \mathcal{O}\left(\log \frac{1}{(1-\gamma)^{3+4m} \ell^{2m} \eta^2}\right)$, we have

$$\mathbb{P}\{\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}\} \geq 1 - mT\delta.$$

Our approach to prove Lemma D.4 is as follows: 1) First, we establish $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$ by considering suitably small variances $\sigma^0(n)$, $\sigma^1(n)$, and biases $b^0(H)$ and $b^1(H)$; 2) Second, we guarantee the event \mathcal{B} based on the construction of γ_t in Section D.3.1; 3) Third, the sub-Gaussian bounds in Proposition 3.3 enable us to establish $\mathbb{P}(\mathcal{C}) \geq 1 - \delta$, again by sufficiently bounding the variances $\sigma^0(n)$, $\sigma^1(n)$, and biases $b^0(H)$ and $b^1(H)$. By combining all of these results, we can determine the requirements for the variances $\sigma^0(n)$, $\sigma^1(n)$, and the biases $b^0(H)$ and $b^1(H)$ to satisfy

$$\mathbb{P}\{\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}\} \geq 1 - \delta.$$

All of these factors can be controlled by n and H .

To establish 1) above, we first show that the product of the values of $V_i^{\theta_t}(\rho)$, where $i \in \mathbf{B}_\eta(\theta_t)$, does not decrease in the next iteration, as shown in Lemma D.5. Lemma D.5 implies that if one of these constraint values in $\mathbf{B}_\eta(\theta_t)$ decreases in the next step, then at least one of the other constraint values in $\mathbf{B}_\eta(\theta_t)$ will increase in the next step. Therefore, Lemma D.5 prevents the constraint values from continuously decreasing during the learning process. Furthermore, due to the chosen stepsize, each $V_i^{\theta_{t+1}}(\rho)$ for $i \in [m]$ cannot decrease significantly, as it is always lower bounded by $\frac{V_i^{\theta_t}(\rho)}{2}$. Therefore, we can establish 1) as demonstrated in (Usmanova et al., 2024, Lemma 6).

Lemma D.5. *Let Assumptions 3.1, 4.1, and 4.2 hold with $\eta \leq \nu_{emf}$, and set*

$$\sigma^0(n) \leq \frac{\alpha_i(t) \min\{2\alpha_i(t), \eta\}}{8\eta \sqrt{\ln \frac{2}{\delta}}}, \sigma^1(n) \leq \frac{L\alpha_i(t)}{3\eta \sqrt{\ln \frac{1}{\delta}}}, b^0(H) \leq \frac{\alpha_i(t) \min\{2\alpha_i(t), \eta\}}{8\eta}, b^1(H) \leq \frac{L\alpha_i(t)}{3\eta}.$$

If at iteration t we have $\min_{i \in [m]} \alpha_i(t) \leq \frac{\ell\eta}{L(1+\frac{4m}{3})}$, for the next iteration we have $\mathbb{P}(\prod_{i \in \mathbf{B}} \alpha_i(t+1) \geq \prod_{i \in \mathbf{B}} \alpha_i(t)) \geq 1 - \delta$ for any \mathbf{B} such that $\mathbf{B}_\eta(\theta_t) \subset \mathbf{B}$, where $\mathbf{B}_\eta(\theta_t) := \{i \in [m] \mid \alpha_i(t) \leq \eta\}$.

Proof of Lemma D.5. Due to the choice of stepsize, $\mathbb{P}(\gamma_t \leq \frac{1}{M_t}) \geq 1 - \delta$, where M_t is the local smoothness constant of the log barrier function $B_\eta^\theta(\rho)$. We have

$$\begin{aligned}
 & \eta \sum_{i \in \mathbf{B}_\eta(\theta_t)} \log \alpha_i(t+1) - \eta \sum_{i \in \mathbf{B}_\eta(\theta_t)} \log \alpha_i(t) \\
 & \geq \gamma_t \left\langle \eta \sum_{i \in \mathbf{B}_\eta(\theta_t)} \frac{\nabla_\theta V_i^{\theta_t}(\rho)}{\alpha_i(t)}, \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\rangle - \frac{M_t \gamma_t^2}{2} \left\| \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\|^2 \\
 & \geq \gamma_t \left(\left\langle \eta \sum_{i \in \mathbf{B}_\eta(\theta_t)} \frac{\nabla_\theta V_i^{\theta_t}(\rho)}{\alpha_i(t)}, \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\rangle - \frac{1}{2} \left\| \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\|^2 \right) \\
 & = \frac{\gamma_t \eta^2}{2} (2 \langle D_1, D_1 + D_2 \rangle - \|D_1 + D_2\|^2) \\
 & = \frac{\gamma_t \eta^2}{2} (\|D_1\|^2 - \|D_2\|^2), \tag{16}
 \end{aligned}$$

where $D_1 := \sum_{i \in \mathbf{B}_\eta(\theta_t)} \frac{\nabla_\theta V_i^{\theta_t}(\rho)}{\alpha_i(t)}$ and $D_2 := \frac{\hat{\nabla}_\theta B_\eta^{\theta_t}(\rho)}{\eta} - \sum_{i \in \mathbf{B}_\eta(\theta_t)} \frac{\nabla_\theta V_i^{\theta_t}(\rho)}{\alpha_i(t)}$. Under Assumption 4.2, we have

$$\|D_1\| = \left\| \sum_{i \in \mathbf{B}_\eta(\theta_t)} \frac{\nabla_\theta V_i^{\theta_t}(\rho)}{\alpha_i(t)} \right\| \geq \left\langle \sum_{i \in \mathbf{B}_\eta(\theta_t)} \frac{\nabla_\theta V_i^{\theta_t}(\rho)}{\alpha_i(t)}, s_\theta \right\rangle \geq \sum_{i \in \mathbf{B}_\eta(\theta_t)} \frac{\ell}{\alpha_i(t)} \geq \frac{L(1 + \frac{4m}{3})}{\eta}, \tag{17}$$

where we use $\min_{i \in [m]} \alpha_i(t) \leq \frac{\ell \eta}{L(1 + \frac{4m}{3})}$ in the last step. For each $i \in [m]$, since $\sigma^0(n) \leq \frac{\alpha_i(t)}{8\sqrt{\ln \frac{2}{\delta}}}$ and $b^0(H) \leq \frac{\alpha_i(t)}{8}$, we have $\mathbb{P}\left(\frac{3\alpha_i(t)}{4} \leq \hat{\alpha}_i(t)\right) \geq 1 - \delta$ using the sub-Gaussian bound in Proposition 3.3. Therefore, $\mathbb{P}\left(\frac{3\eta}{4} \leq \hat{\alpha}_i(t)\right) \geq 1 - \delta$ for $i \notin \mathbf{B}_\eta(\theta_t)$. Then, we can upper bound $\|D_2\|$ with probability at least $1 - \delta$ as follows:

$$\|D_2\| \tag{18}$$

$$= \left\| \frac{\hat{\nabla}_\theta V_0^{\theta_t}(\rho)}{\eta} + \sum_{i \notin \mathbf{B}_\eta(\theta_t)} \frac{\hat{\nabla}_\theta V_i^{\theta_t}(\rho)}{\hat{\alpha}_i(t)} + \sum_{i \in \mathbf{B}_\eta(\theta_t)} \frac{\hat{\nabla}_\theta V_i^{\theta_t}(\rho)}{\hat{\alpha}_i(t)} - \sum_{i \in \mathbf{B}_\eta(\theta_t)} \frac{\nabla_\theta V_i^{\theta_t}(\rho)}{\alpha_i(t)} \right\| \tag{19}$$

$$\leq \frac{L}{\eta} \left(1 + \frac{4}{3} (m - |\mathbf{B}_\eta(\theta_t)|) \right) + \left\| \sum_{i \in \mathbf{B}_\eta(\theta_t)} \left(\frac{\hat{\nabla}_\theta V_i^{\theta_t}(\rho)}{\hat{\alpha}_i(t)} - \frac{\hat{\nabla}_\theta V_i^{\theta_t}(\rho)}{\alpha_i(t)} + \frac{\hat{\nabla}_\theta V_i^{\theta_t}(\rho)}{\alpha_i(t)} - \frac{\nabla_\theta V_i^{\theta_t}(\rho)}{\alpha_i(t)} \right) \right\| \tag{20}$$

$$\leq \sum_{i \in \mathbf{B}_\eta(\theta_t)} \left(\left\| \frac{\hat{\nabla}_\theta V_i^{\theta_t}(\rho) - \nabla_\theta V_i^{\theta_t}(\rho)}{\alpha_i(t)} \right\| + \left| \frac{1}{\hat{\alpha}_i(t)} - \frac{1}{\alpha_i(t)} \right| \left\| \hat{\nabla}_\theta V_i^{\theta_t}(\rho) \right\| \right) + \frac{L}{\eta} \left(1 + \frac{4}{3} (m - |\mathbf{B}_\eta(\theta_t)|) \right) \tag{21}$$

$$\leq \sum_{i \in \mathbf{B}_\eta(\theta_t)} \left(\frac{\sigma^1(n) \sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}} + b^1(H)}{\alpha_i(t)} + L \frac{\sigma^0(n) \sqrt{\ln \frac{2}{\delta}} + b^0(H)}{\hat{\alpha}_i(t) \alpha_i(t)} \right) + \frac{L}{\eta} \left(1 + \frac{4}{3} (m - |\mathbf{B}_\eta(\theta_t)|) \right) \tag{22}$$

$$\leq \sum_{i \in \mathbf{B}_\eta(\theta_t)} \left(\frac{\sigma^1(n) \sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}} + b^1(H)}{\alpha_i(t)} + 4L \frac{\sigma^0(n) \sqrt{\ln \frac{2}{\delta}} + b^0(H)}{3(\alpha_i(t))^2} \right) + \frac{L}{\eta} \left(1 + \frac{4}{3} (m - |\mathbf{B}_\eta(\theta_t)|) \right). \tag{23}$$

From (19) to (20), we use $\frac{3\eta}{4} \leq \hat{\alpha}_i(t)$ and $\|\hat{\nabla}_\theta V_0^{\theta_t}(\rho)\| \leq L$ by (Xu et al., 2020, Proposition 4.2). From (21) to (22), we use the sub-Gaussian bound in Proposition 3.3 and $\|\hat{\nabla}_\theta V_i^{\theta_t}(\rho)\| \leq L$ for $i \in [m]$. From (22) to (23), we use $\frac{3\alpha_i(t)}{4} \leq \hat{\alpha}_i(t)$. Further, if we set the variances and biases in (23) as

$$\sigma^0(n) \leq \frac{(\alpha_i(t))^2}{4\eta \sqrt{\ln \frac{2}{\delta}}}, \sigma^1(n) \leq \frac{L\alpha_i(t)}{3\eta \sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}}}, b^0(H) \leq \frac{(\alpha_i(t))^2}{4\eta}, b^1(H) \leq \frac{L\alpha_i(t)}{3\eta},$$

then we can have $\mathbb{P}\left(\|D_2\| \leq \frac{L(1 + \frac{4m}{3})}{\eta}\right) \geq 1 - \delta$. Combining this property with (17), we have $\mathbb{P}(\|D_2\| \leq \|D_1\|) \geq$

$1 - \delta$. Taking this relation into (16), we have

$$\mathbb{P} \left(\prod_{i \in \mathbf{B}_\eta(\theta_t)} \alpha_i(t+1) \geq \prod_{i \in \mathbf{B}_\eta(\theta_t)} \alpha_i(t) \right) \geq 1 - \delta.$$

Same result if we replace $\mathbf{B}_\eta(\theta_t)$ with any \mathbf{B} such that $\mathbf{B}_\eta(\theta_t) \subset \mathbf{B}$. \square

With the above lemma in place, we are ready to prove Lemma D.4.

Proof of Lemma D.4. First, we need to choose $\eta \leq \nu_{emf}$ and set

$$\sigma^0(n) \leq \frac{\alpha_i(t) \min \{2\alpha_i(t), \eta\}}{8\eta \sqrt{\ln \frac{2}{\delta}}}, \sigma^1(n) \leq \frac{L\alpha_i(t)}{3\eta \sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}}}, b^0(H) \leq \frac{\alpha_i(t) \min \{2\alpha_i(t), \eta\}}{8\eta}, b^1(H) \leq \frac{L\alpha_i(t)}{3\eta}$$

to satisfy the conditions in Lemma D.5. Then, we can combine the result from Lemma D.5 with the result from (Usmanova et al., 2024, Lemma 6), and we have

$$\mathbb{P} \left\{ \forall t \in [T], \min_{i \in [m]} \alpha_i(t) \geq c\eta, \min_{i \in [m]} \hat{\alpha}_i(t) \geq \frac{3}{8}c\eta \text{ and } \min_{i \in [m]} \underline{\alpha}_i(t) \geq \frac{c\eta}{2} \right\} \geq 1 - mT\delta, c = \left(\frac{\ell(1-\gamma)^2}{4M_g(1 + \frac{4m}{3})} \right)^m. \quad (24)$$

Based on the lower bound of $\hat{\alpha}_i(t)$, we can further bound γ_t by (Usmanova et al., 2024, Lemma 6) as

$$\gamma_t \geq C\eta, C := \frac{c}{2L^2(1 + \frac{m}{c}) \max \left\{ 4 + \frac{5Mc}{L^2}, 1 + \sqrt{\frac{Mc}{4L^2}} \right\}}. \quad (25)$$

Further, if we set

$$\begin{aligned} \sigma^0(n) &\leq \min \left\{ \frac{\alpha_i(t) \min \{2\alpha_i(t), \eta\}}{8\eta \sqrt{\ln \frac{2}{\delta}}}, \frac{1}{\left(\sum_{i=1}^m \frac{16L}{\alpha_i(t)\hat{\alpha}_i(t)} \right) \sqrt{\ln \frac{2}{\delta}}} \right\}, \\ \sigma^1(n) &\leq \min \left\{ \frac{L\alpha_i(t)}{3\eta \sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}}}, \frac{\eta}{16 \left(1 + \sum_{i=1}^m \frac{\eta}{\hat{\alpha}_i(t)} \right) \sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}}} \right\}, \\ b^0(H) &\leq \min \left\{ \frac{\alpha_i(t) \min \{2\alpha_i(t), \eta\}}{8\eta}, \frac{1}{16 \left(\sum_{i=1}^m \frac{L}{\alpha_i(t)\hat{\alpha}_i(t)} \right)} \right\}, b^1(H) \leq \min \left\{ \frac{L\alpha_i(t)}{3\eta}, \frac{\eta}{16 \left(1 + \sum_{i=1}^m \frac{\eta}{\hat{\alpha}_i(t)} \right)} \right\}. \end{aligned} \quad (26)$$

By Lemma 3.4, we have

$$\mathbb{P} \left(\|\Delta_t\| \geq \frac{\eta}{4} \right), \quad (28)$$

Combing the results from (24) and (25) with inequality (28), we have

$$\mathbb{P} \left\{ \forall t \in [T], \min_{i \in [m]} V_i^{\theta_t}(\rho) \geq c\eta, \gamma_t \geq C\eta \text{ and } \|\Delta_t\| \geq \frac{\eta}{4} \right\} \geq 1 - mT\delta.$$

Based on the above result regarding the lower bound on $\alpha_i(t)$ and $\hat{\alpha}_i(t)$, we can further set the variances and biases in (27) as follows:

$$\sigma^0(n) \leq \min \left\{ \frac{c\eta \min \{4c, 1\}}{4\sqrt{\ln \frac{2}{\delta}}}, \frac{3c^2\eta^2}{32L\sqrt{\ln \frac{2}{\delta}}} \right\}, \sigma^1(n) \leq \min \left\{ \frac{2cL}{3\sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}}}, \frac{\eta}{16 \left(1 + \frac{4m}{3c} \right) \sqrt{\ln \frac{e^{\frac{1}{4}}}{\delta}}} \right\},$$

$$b^0(H) \leq \min \left\{ \frac{c\eta \min\{4c, 1\}}{4}, \frac{3c^2\eta^2}{32L} \right\}, b^1(H) \leq \min \left\{ \frac{2cL}{3}, \frac{\eta}{16(1 + \frac{4m}{3c})} \right\}.$$

According to Proposition 3.3, the number of trajectories n and the truncated horizon H need to be set as follows:

$$n := \max \left\{ \frac{2048L^2 \ln \frac{2}{\delta}}{9(1-\gamma)^2 c^4 \eta^4}, \frac{32 \ln \frac{2}{\delta}}{c^2(1-\gamma)^2 \eta^2 \min\{16c^2, 1\}}, \frac{2048(1 + \frac{4m}{3c})^2 M_g^2 \ln \frac{e^{\frac{1}{4}}}{\delta}}{\eta^2(1-\gamma)^3} \frac{18M_g^2 \ln \frac{e^{\frac{1}{4}}}{\delta}}{c^2 L^2 (1-\gamma)^3} \right\},$$

$$H := \max \left\{ \log_\gamma \left(\frac{3(1-\gamma)c^2\eta^2}{32L} \right), \log_\gamma \left(\frac{c(1-\gamma)\eta \min\{4c, 1\}}{4} \right), \mathcal{O} \left(\log_\gamma \frac{(1-\gamma)\eta}{16(1 + \frac{4m}{3c})M_g} \right), \right. \quad (29)$$

$$\left. \mathcal{O} \left(\log_\gamma \frac{2(1-\gamma)cL}{3M_g} \right) \right\}, \quad (\text{N-H})$$

where the first condition for H is to satisfy $b^0(H) \leq \min \left\{ \frac{c \min\{4c, 1\}}{4}, \frac{3c^2}{32L} \right\}$ and the second condition for H is to satisfy $b^1(H) \leq \min \left\{ \frac{2cL}{3}, \frac{\eta}{16(1 + \frac{4m}{3c})} \right\}$. Therefore, we need to set the number of trajectories n of the order $\mathcal{O} \left(\frac{\ln \frac{1}{\delta}}{(1-\gamma)^{6+8m} \ell^4 m \eta^4} \right)$ and the truncated horizon H of the order $\mathcal{O} \left(\ln \frac{1}{\ell \eta} \right)$. \square

D.4 Proof of Lemma 4.10

The proof of Lemma 4.10 is based on the performance difference lemma, provided below for completeness.

Theorem D.6 (The performance difference lemma (Sutton et al., 1999)). $\forall \theta, \theta' \in \mathbb{R}^d, \forall i \in \{0, \dots, m\}$, we have

$$V_i^{\theta'}(\rho) - V_i^{\theta}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_\rho^\theta} \left[A_i^{\theta'}(s, a) \right].$$

Proof of Lemma 4.10. We can derive the following equality by using the performance difference lemma,

$$V_0^{\pi^*}(\rho) - V_0^\theta(\rho) + \eta \sum_{i=1}^m \left(\frac{V_i^{\pi^*}(\rho) - V_i^\theta(\rho)}{V_i^\theta(\rho)} \right) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_\rho^{\pi^*}} \left[A_0^\theta(s, a) + \eta \sum_{i=1}^m \frac{A_i^\theta(s, a)}{V_i^\theta(\rho)} \right]. \quad (30)$$

Applying Jensen's inequality to Assumption 4.8, we obtain $\forall i \in \{0, \dots, m\}$,

$$\mathbb{E}_{(s,a) \sim d_\rho^{\pi^*}} \left[A_i^\theta(s, a) - (1-\gamma) \mu_i^{*T} \nabla_\theta \log \pi_\theta(a|s) \right] \leq \sqrt{\varepsilon_{bias}}. \quad (31)$$

Plugging inequality (31) into (30), we get

$$\begin{aligned} & V_0^{\pi^*}(\rho) - V_0^\theta(\rho) + \eta \sum_{i=1}^m \left(\frac{V_i^{\pi^*}(\rho) - V_i^\theta(\rho)}{V_i^\theta(\rho)} \right) \\ & \leq \mathbb{E}_{(s,a) \sim d_\rho^{\pi^*}} \left[\left(\mu_0^* + \sum_{i=1}^m \frac{\eta \mu_i^*}{V_i^\theta(\rho)} \right)^T \nabla_\theta \log \pi_\theta(a|s) \right] + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} \left(\sum_{i=1}^m \frac{\eta}{V_i^\theta(\rho)} + 1 \right) \\ & = \mathbb{E}_{(s,a) \sim d_\rho^{\pi^*}} \left[(\nabla_\theta B_\eta^\theta(\rho))^T (F^\theta(\rho))^\dagger \nabla_\theta \log \pi_\theta(a|s) \right] + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} \left(1 + \sum_{i=1}^m \frac{\eta}{V_i^\theta(\rho)} \right) \\ & = \mathbb{E}_{(s,a) \sim d_\rho^{\pi^*}} \left[(\mathbf{P}_{\text{Ker}(F^\theta(\rho))} \nabla_\theta B_\eta^\theta(\rho) + \mathbf{P}_{\text{Im}(F^\theta(\rho))} \nabla_\theta B_\eta^\theta(\rho))^T (F^\theta(\rho))^\dagger \nabla_\theta \log \pi_\theta(a|s) \right] + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} \left(1 + \sum_{i=1}^m \frac{\eta}{V_i^\theta(\rho)} \right) \\ & = \mathbb{E}_{(s,a) \sim d_\rho^{\pi^*}} \left[(\mathbf{P}_{\text{Im}(F^\theta(\rho))} \nabla_\theta B_\eta^\theta(\rho))^T (F^\theta(\rho))^\dagger \nabla_\theta \log \pi_\theta(a|s) \right] + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} \left(1 + \sum_{i=1}^m \frac{\eta}{V_i^\theta(\rho)} \right) \\ & \leq \frac{M_h}{\mu_F} \left\| \mathbf{P}_{\text{Im}(F^\theta(\rho))} \nabla_\theta B_\eta^\theta(\rho) \right\| + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} \left(1 + \sum_{i=1}^m \frac{\eta}{V_i^\theta(\rho)} \right), \end{aligned} \quad (32)$$

where we use Assumption 4.4 to bound $\left\| \left(\mathbf{P}_{\mathbf{Im}(F^\theta(\rho))} \nabla_\theta B_\eta^\theta(\rho) \right)^T (F^\theta(\rho))^\dagger \right\|$ by $\|\mathbf{P}_{\mathbf{Im}(F^\theta(\rho))} \nabla_\theta B_\eta^\theta(\rho)\|/\mu_F$ and Assumption 3.1 to bound $\|\nabla \log \pi_\theta(a|s)\|$ by M_h in the last step. Rearranging the above inequality, we have

$$\begin{aligned} V_0^{\pi^*}(\rho) - V_0^\theta(\rho) &\leq m\eta + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} \left(1 + \sum_{i=1}^m \frac{\eta}{V_i^\theta(\rho)} \right) + \frac{M_h}{\mu_F} \|\mathbf{P}_{\mathbf{Im}(F^\theta(\rho))} \nabla_\theta B_\eta^\theta(\rho)\| \\ &\leq m\eta + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} \left(1 + \sum_{i=1}^m \frac{\eta}{V_i^\theta(\rho)} \right) + \frac{M_h}{\mu_F} \|\nabla_\theta B_\eta^\theta(\rho)\|. \end{aligned}$$

□

D.5 Proof of Lemma 4.11

In this section, we prove that the stationary points of the log barrier function is at most $\Omega(\nu_{emf} + \eta)$ close to the boundary.

Proof of Lemma 4.11. Since θ_{st} is the stationary point, we have

$$\nabla_\theta B_\eta^{\theta_{st}}(\rho) = \nabla_\theta V_0^{\theta_{st}}(\rho) + \eta \sum_{i=1}^m \frac{\nabla_\theta V_i^{\theta_{st}}(\rho)}{V_i^{\theta_{st}}(\rho)} = 0.$$

Rearranging the terms in the above equation, we obtain

$$\sum_{i \notin \mathbf{B}_{\nu_{emf}}(\theta_{st})} \frac{\nabla_\theta V_i^{\theta_{st}}(\rho)}{V_i^{\theta_{st}}(\rho)} + \sum_{i \in \mathbf{B}_{\nu_{emf}}(\theta_{st})} \frac{\nabla_\theta V_i^{\theta_{st}}(\rho)}{V_i^{\theta_{st}}(\rho)} = \frac{-\nabla_\theta V_0^{\theta_{st}}(\rho)}{\eta}. \quad (33)$$

If $\mathbf{B}_{\nu_{emf}}(\theta_{st})$ is an empty set, then

$$\min_{i \in [m]} \{V_i^{\theta_{st}}(\rho)\} \geq \nu_{emf}.$$

Otherwise, by Assumption 4.2, there exists a unit vector $s_{\theta_{st}} \in \mathbb{R}^d$ such that for $i \in \mathbf{B}_{\nu_{emf}}(\theta_{st})$, we have

$$\langle s_{\theta_{st}}, \nabla_\theta V_i^{\theta_{st}}(\rho) \rangle > \ell.$$

Taking the dot product of both sides of equation (33) with $s_{\theta_{st}}$ and using Lipschitz continuity, we obtain

$$\begin{aligned} \frac{\ell}{\min_{i \in [m]} \{V_i^{\theta_{st}}(\rho)\}} &\leq \frac{\langle s_{\theta_{st}}, \nabla_\theta V_i^{\theta_{st}}(\rho) \rangle}{\min_{i \in [m]} \{V_i^{\theta_{st}}(\rho)\}} \sum_{i \in \mathbf{B}_{\nu_{emf}}(\theta_{st})} \frac{\min_{i \in [m]} \{V_i^{\theta_{st}}(\rho)\}}{V_i^{\theta_{st}}(\rho)} \\ &= \frac{\langle s_{\theta_{st}}, -\nabla_\theta V_0^{\theta_{st}}(\rho) \rangle}{\eta} - \sum_{i \notin \mathbf{B}_{\nu_{emf}}(\theta_{st})} \frac{\langle s_{\theta_{st}}, \nabla_\theta V_i^{\theta_{st}}(\rho) \rangle}{V_i^{\theta_{st}}(\rho)} \geq \frac{mL}{\min\{\eta, \nu_{emf}\}}, \end{aligned}$$

Therefore,

$$\min_{i \in [m]} \{V_i^{\theta_{st}}(\rho)\} \geq \frac{\min\{\eta, \nu_{emf}\} \ell}{mL}.$$

In conclusion, we have $\min_{i \in [m]} \{V_i^{\theta_{st}}(\rho)\} \geq \min \left\{ \frac{\nu_{emf}, \min\{\eta, \nu_{emf}\} \ell}{mL} \right\}$. □

D.6 Proof of Theorem 4.12

Proof of Theorem 4.12: We set the values for n , H , and η to satisfy the conditions outlined in Proposition 4.3. Due to our choice of stepsize, we have $\mathbb{P}\left(\gamma_t \leq \frac{1}{M_t}\right) \geq 1 - \delta$ as showed in Proposition 4.3, where M_t represents the local smoothness constant of the log barrier function $B_\eta^\theta(\rho)$. With this, we can bound $B_\eta^{\theta_{t+1}}(\rho) - B_\eta^{\theta_t}(\rho)$ with probability at least $1 - \delta$ as

$$B_\eta^{\theta_{t+1}}(\rho) - B_\eta^{\theta_t}(\rho) \geq \gamma_t \left\langle \nabla_\theta B_\eta^{\theta_t}(\rho), \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\rangle - \frac{M_t \gamma_t^2}{2} \left\| \hat{\nabla}_\theta B_\eta^{\theta_t}(\rho) \right\|^2$$

$$\begin{aligned}
 &\geq \gamma_t \left\langle \nabla_{\theta} B_{\eta}^{\theta_t}(\rho), \hat{\nabla}_{\theta} B_{\eta}^{\theta_t}(\rho) \right\rangle - \frac{\gamma_t}{2} \left\| \hat{\nabla}_{\theta} B_{\eta}^{\theta_t}(\rho) \right\|^2 \\
 &= \gamma_t \left\langle \nabla_{\theta} B_{\eta}^{\theta_t}(\rho), \Delta_t + \nabla_{\theta} B_{\eta}^{\theta_t}(\rho) \right\rangle - \frac{\gamma_t}{2} \left\| \Delta_t + \nabla_{\theta} B_{\eta}^{\theta_t}(\rho) \right\|^2 \\
 &= \frac{\gamma_t}{2} \left\| \nabla_{\theta} B_{\eta}^{\theta_t}(\rho) \right\|^2 - \frac{\gamma_t}{2} \left\| \Delta_t \right\|^2.
 \end{aligned} \tag{34}$$

We divide the analysis into two cases based on the **if condition** in algorithm 1 line 5.

Case 1: If $\left\| \hat{\nabla}_{\theta} B_{\eta}^{\theta_t}(\rho) \right\| \geq \frac{\eta}{2}$, then $\left\| \nabla_{\theta} B_{\eta}^{\theta_t}(\rho) \right\| \geq \frac{\eta}{4}$ since $\left\| \Delta_t \right\| \leq \frac{\eta}{4}$ by Proposition 4.3. We can further write (34) as

$$B_{\eta}^{\theta_{t+1}}(\rho) - B_{\eta}^{\theta_t}(\rho) \geq \frac{C\eta^2}{8} \left\| \nabla_{\theta} B_{\eta}^{\theta_t}(\rho) \right\| - \frac{C\eta^3}{32}, \tag{35}$$

where we plug in $\gamma_t \geq C\eta$ in the last step. Since $\min_{i \in [m]} V_i^{\theta_t}(\rho) \geq c\eta$, we can rewrite Lemma 4.10 as:

$$V_0^{\pi^*}(\rho) - V_0^{\theta_t}(\rho) \leq a + \frac{M_h}{\mu_F} \left\| \nabla_{\theta} B_{\eta}^{\theta_t}(\rho) \right\|, \tag{36}$$

where $a = m\eta + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} \left(1 + \frac{m}{c}\right)$. Plugging (36) into (35), we get

$$B_{\eta}^{\theta_{t+1}}(\rho) - B_{\eta}^{\theta_t}(\rho) \geq \frac{C\mu_F\eta^2}{8M_h} \left(V_0^{\pi^*}(\rho) - V_0^{\theta_t}(\rho) \right) - \frac{C\eta^3}{32} - \frac{aC\mu_F\eta^2}{8M_h},$$

which can be further simplified to:

$$V_0^{\pi^*}(\rho) - V_0^{\theta_{t+1}}(\rho) \leq \left(1 - \frac{C\mu_F\eta^2}{8M_h}\right) \left(V_0^{\pi^*}(\rho) - V_0^{\theta_t}(\rho) \right) + \frac{C\eta^3}{32} + \frac{aC\mu_F\eta^2}{8M_h} + \eta \sum_{i=1}^m \log \frac{V_i^{\theta_{t+1}}(\rho)}{V_i^{\theta_t}(\rho)}.$$

By recursively applying the above inequality and setting $\frac{C\mu_F\eta^2}{8M_h} < 1$, we obtain

$$\begin{aligned}
 &V_0^{\pi^*}(\rho) - V_0^{\theta_{t+1}}(\rho) \\
 &\leq \left(1 - \frac{C\mu_F\eta^2}{8M_h}\right)^{t+1} \left(V_0^{\pi^*}(\rho) - V_0^{\theta_0}(\rho) \right) + \left(\frac{aC\mu_F\eta^2}{8M_h} + \frac{C\eta^3}{32} \right) \sum_{i=0}^t \left(1 - \frac{C\mu_F\eta^2}{8M_h}\right)^i \\
 &\quad + \eta \sum_{i=1}^m \log V_i^{\theta_{t+1}}(\rho) - \eta \left(1 - \frac{C\mu_F\eta^2}{8M_h}\right)^t \sum_{i=1}^m \log V_i^{\theta_0}(\rho) - \frac{C\mu_F\eta^3}{8M_h} \sum_{i=1}^m \sum_{j=1}^t \left(1 - \frac{C\mu_F\eta^2}{8M_h}\right)^{t-j} \log V_i^{\theta_j}(\rho) \\
 &\leq \left(1 - \frac{C\mu_F\eta^2}{8M_h}\right)^{t+1} \left(V_0^{\pi^*}(\rho) - V_0^{\theta_0}(\rho) \right) + a + \frac{M_h\eta}{4\mu_F} + m\eta \log \frac{1}{1-\gamma} \\
 &\quad - m\eta \left(1 - \frac{C\mu_F\eta^2}{8M_h}\right)^t \log \nu_s - \frac{C\mu_F\eta^3}{8M_h} \sum_{i=1}^m \sum_{j=1}^t \left(1 - \frac{C\mu_F\eta^2}{8M_h}\right)^{t-j} \log(c\eta) \\
 &\leq \left(1 - \frac{C\mu_F\eta^2}{8M_h}\right)^{t+1} \left(V_0^{\pi^*}(\rho) - V_0^{\theta_0}(\rho) \right) + a + \frac{M_h\eta}{4\mu_F} + m\eta \log \frac{1}{c\eta(1-\gamma)} - m\eta \left(1 - \frac{C\mu_F\eta^2}{8M_h}\right)^t \log \nu_s.
 \end{aligned} \tag{37}$$

Case 2: If $\left\| \hat{\nabla}_{\theta} B_{\eta}^{\theta_t}(\rho) \right\| \leq \frac{\eta}{2}$, we have $\left\| \nabla_{\theta} B_{\eta}^{\theta_t}(\rho) \right\| \leq \left\| \hat{\nabla}_{\theta} B_{\eta}^{\theta_t}(\rho) \right\| + \left\| \Delta_t \right\| \leq \frac{3\eta}{4}$. Applying (36), we have

$$V_0^{\pi^*}(\rho) - V_0^{\theta_t}(\rho) \leq \frac{3M_h\eta}{4\mu_F} + a. \tag{38}$$

Combining the inequalities (37) and (38), we conclude that after T iterations of the Algorithm 1, the output policy $\pi_{\theta_{out}}$ satisfies

$$\begin{aligned}
 &V_0^{\pi^*}(\rho) - V_0^{\theta_{out}}(\rho) \\
 &\leq \left(1 - \frac{C\mu_F\eta^2}{8M_h}\right)^T \left(V_0^{\pi^*}(\rho) - V_0^{\theta_0}(\rho) \right) + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} \left(1 + \frac{m}{c}\right)
 \end{aligned}$$

$$\begin{aligned}
 & + m\eta \left(\frac{3M_h}{4\mu_F m} + 1 + \log \frac{1}{c\eta(1-\gamma)} - \left(1 - \frac{C\mu_F\eta^2}{8M_h} \right)^{T-1} \log \nu_s \right) \\
 & = \mathcal{O} \left(\exp(-CT\mu_F\eta^2) \right) \left(V_0^{\pi^*}(\rho) - V_0^{\theta_0}(\rho) \right) + \mathcal{O} \left(\frac{\sqrt{\varepsilon_{\text{bias}}}}{(1-\gamma)^{2m+1}\ell^m} \right) + \mathcal{O} \left(\frac{\eta}{\mu_F} \right) + \mathcal{O} \left(\eta \ln \frac{1}{\eta} \right)
 \end{aligned}$$

with a probability of at least $1 - mT\delta$. \square

D.7 Proof of Corollary 4.13

Proof of Corollary 4.13. By analyzing the inequalities, namely (37) and (38), provided in Section D.6, we can prove Corollary 4.13 as follows: setting $n = \mathcal{O} \left(\frac{\ln \frac{1}{\delta}}{(1-\gamma)^{6+8m}\ell^4 m \varepsilon^4} \right)$ and $H = \mathcal{O} \left(\log \frac{1}{\varepsilon} \right)$, $\eta = \varepsilon$ and $T = \mathcal{O} \left(\frac{\log \frac{1}{\varepsilon}}{\mu_F \varepsilon^2 \ell^{2m} (1-\gamma)^{4m+4}} \right)$. After T iterations of the LB-SGD Algorithm, the output θ_{out} satisfies

$$V_0^{\pi^*}(\rho) - V_0^{\theta_{\text{out}}}(\rho) \leq \mathcal{O} \left(\varepsilon \log \frac{1}{\varepsilon} \right) + \mathcal{O} \left(\frac{\varepsilon}{\mu_F} \right) + \mathcal{O} \left(\frac{\sqrt{\varepsilon_{\text{bias}}}}{(1-\gamma)^{2m+1}\ell^m} \right), \quad (39)$$

while safe exploration is ensured with a probability of at least $1 - mT\delta$. To ensure satisfaction of (39) and maintain safe exploration with a probability of at least $1 - \beta$, we need set $\delta = \frac{\beta}{mT} = \mathcal{O}(\beta\varepsilon^2)$ and sample size $\mathcal{O} \left(\frac{\log^3 \frac{1}{\varepsilon} \log \frac{1}{\ell}}{\mu_F (1-\gamma)^{10+12m} \ell^6 m \varepsilon^6} \right)$. \square

E EXPERIMENT

We conducted experiments ³ in a 6×6 gridworld environment introduced by (Sutton and Barto, 2018) (see Figure 1). We aim to reach the rewarded cell while controlling the time spent visiting the red rectangles under a certain threshold. We define the CMDP as follows: the environment involves four actions: *up*, *right*, *down*, and *left*. The agent moves in the specified direction with a 0.9 probability and randomly selects another direction with a 0.1 probability after taking an action. The constraints are defined as follows: If the agent hits the second-row or the fourth-row red rectangles, the reward functions $r_1(s, a)$ and $r_2(s, a)$ receive -10 respectively. Once the agent reaches a rewarded cell, it remains there indefinitely, receiving a reward of 1 per iteration. We set the discount factor γ to 0.7. We define the CMDP we solve as follows:

$$\max_{\pi_\theta} V_0^{\pi_\theta}(\rho) \quad \text{subject to} \quad V_i^{\pi_\theta}(\rho) \geq -2, \quad i \in [2],$$

utilizing the softmax policy parameterization.

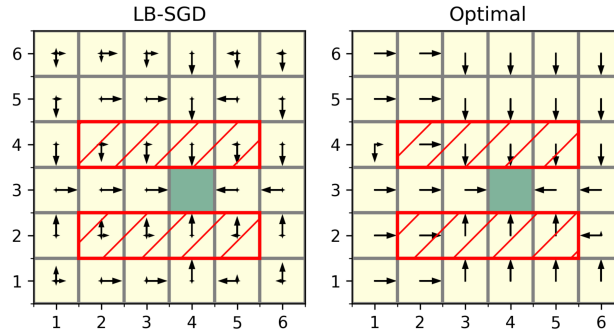


Figure 1: Gridworld Environment: The green block denotes the reward, the arrows represent the policies, and the two red-hatched rectangles indicate the constrained states.

The optimal policy, depicted as arrows in Figure 1, is computed using linear programming in occupancy-measure space. Notice that the agent needs to learn the optimal policy within the regions highlighted by red rectangles, as there is always a small probability of ending up there.

³All the experiments in this subsection were carried out on a MacBook Pro with an Apple M1 Pro chip and 32 GB of RAM. Our code is developed based on (Schlaginhaufen and Kamgarpour, 2023, <https://github.com/andrschl/cirl>).

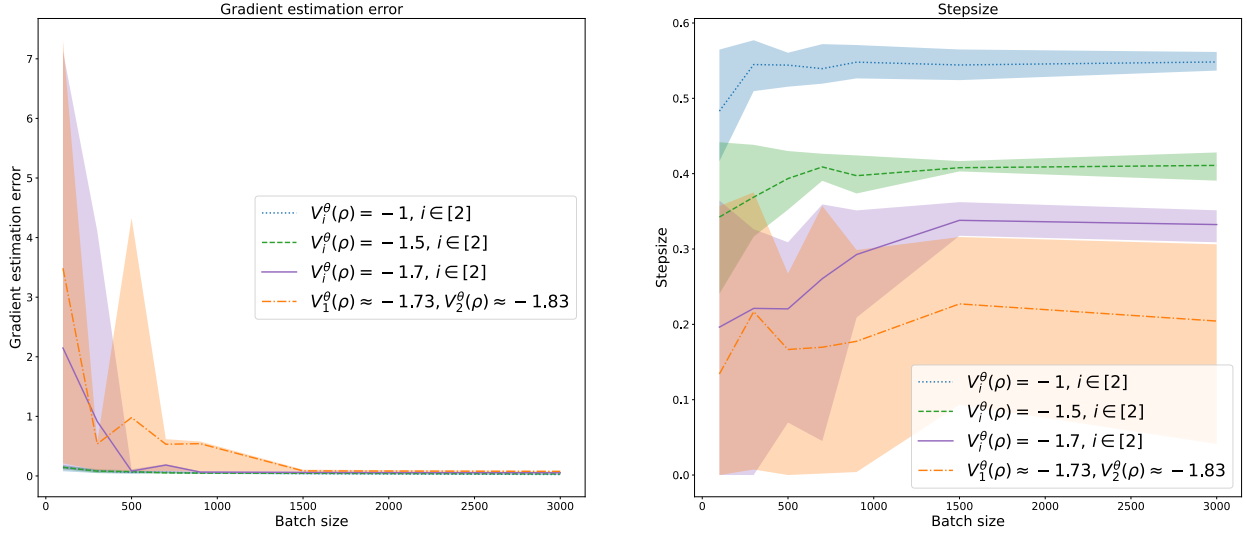


Figure 2: The gradient estimation error for the log barrier function and the value of stepsize for LB-SGD algorithm with sample sizes of 100, 300, 500, 700, 900, 1500, and 3000 at varying distances from the boundary. In the figure above, the lines indicate the median values obtained from 10 independent experiments, while the shaded areas represent the 10% and 90% percentiles calculated from 10 different random seeds.

Since our paper emphasizes the feasibility of iterates rather than bounding average constraint violations, we compare with two typical CMDP learning algorithms: (1) the IPO algorithm (Liu et al., 2020a), which combines the log barrier method with a policy gradient approach and fixed stepsize, and (2) the RPG-PD algorithm (Ding et al., 2024), a primal-dual method that guarantees the feasibility of the last iterate, using an entropy-regularized policy gradient and quadratic-regularized gradient ascent for the dual variable.

E.1 Hyperparameter Tuning

In this section, we outline the parameter choices for the three algorithms: LB-SGD, IPO, and RPG-PD.

For the LB-SGD algorithm, the Lipschitz constants M_g and the smoothness parameter M_h of the function $\log \pi_\theta(a|s)$ are required to compute the smoothness of the value functions $V_i^\theta(\rho)$, which are critical for determining stepsizes. For direct, softmax, or log-linear parameterizations, these values can be computed directly. For other parameterizations, they can be estimated from sampled trajectories (see Appendix F). Although M_g and M_h are both 1 for the softmax parameterization used in our experiment, we validated the effectiveness of the estimation approach described in Appendix F. To implement the LB-SGD algorithm, we further need to choose the parameter η , which needs to satisfy $\eta \leq \nu_{\text{emf}}$ according to Theorem 3.5. However, calculating ν_{emf} is generally infeasible. In practice, we do not need its exact value to select η . If $\eta \geq \nu_{\text{emf}}$, the iterates remain within a distance of $\Omega(\nu_{\text{emf}})$ from the boundary rather than $\Omega(\eta)$. Therefore, LB-SGD ensures safe exploration for any η , though smaller η values bring the iterates closer to the boundary, expanding the search area for policy optimization. A smaller η leads to a final policy closer to the optimal one, but this requires more samples to ensure accurate estimators, especially near the boundary. In our case, we selected $\eta = 0.01$ to keep the output policy close to optimal. After fixing η , we need to choose the number of trajectories per iteration for calculating the stepsize and the log barrier gradient. In Figure 2, we plot the gradient estimation error for the log barrier function and the stepsize returned by the LB-SGD algorithm, using different sample sizes at different points that vary the distance to the boundary. As shown in Figure 2, when the point is closer to the boundary, our algorithm requires a higher number of samples per iteration to obtain accurate estimates of stepsizes and log barrier gradients. Inadequate sampling leads to relatively smaller estimates of stepsizes with higher variance in both gradient and stepsize estimations. Therefore, we used 3,000 trajectories per iteration to accurately estimate the gradient and stepsize.

Since the IPO algorithm also uses the log barrier method combined with a policy gradient approach, similar to LB-SGD, we chose the same value for η and used 3,000 trajectories per iteration to estimate the gradient. The main difference lies in IPO’s manually chosen fixed stepsize. To ensure safe exploration and avoid unsafe behavior from larger stepsizes, we selected stepsizes of 1.5, 1, and 0.5.

The RPG-PD algorithm requires tuning three parameters: the regularization parameter τ , the stepsize, and ε_0 for projecting the policy onto the probability simplex. We selected the stepsize based on (Ding et al., 2024, Theorems 2), which provides an optimal choice of η , determined by τ , ε_0 , and the Slater parameter ν_s which can be computed since we are given a safe initial starting point. To ensure the safety of the last iterate, RPG-PD learns from more conservative constraints, $V_i^\pi(\rho) \geq b - 2$ with $b > 0$. We fine-tuned the parameters τ , ε_0 and b using exact information and selected $\tau = \varepsilon_0 = b = 0.1$. We plotted the performance of RPG-PD with these parameters in Figure 3, labeled as learning curve RPG-PD_exact. For the stochastic version of RPG-PD, we used the same parameters as the exact version and employed 800 trajectories to estimate gradient information.

E.2 Comparison and Conclusion

In this section, we analyze the performance of the three algorithms—LB-SGD, IPO, and RPG-PD, using parameters as described in Section E.1, and evaluated with stochastic information.

In terms of ensuring safe exploration, all three algorithms successfully achieve this, as shown in Figure 3. However, for the IPO and RPG-PD algorithms, stepsize tuning is crucial to avoid constraint violations during learning, as larger stepsizes can lead to unsafe behavior. In contrast, the LB-SGD algorithm, which adapts its stepsizes based on sample estimates, eliminates the need for manual tuning.

Regarding convergence guarantees, LB-SGD proves to be more sample-efficient in finding the optimal policy compared to RPG-PD. By iteration 4000, the RPG-PD algorithm’s iterate remains approximately 0.5 away from the boundary, whereas LB-SGD’s iterate is only about 0.02 away. This result is consistent with theoretical results: RPG-PD requires $\mathcal{O}(\varepsilon^{-14})$ samples to find an $\mathcal{O}(\varepsilon)$ -optimal policy, while LB-SGD requires only $\tilde{\mathcal{O}}(\varepsilon^{-6})$ samples. However, the IPO algorithm, with well-tuned stepsizes, converges faster and achieves closer proximity to the optimal reward value compared to LB-SGD. This is due to LB-SGD’s more conservative choice of stepsizes to ensure safe exploration. Specifically, the variation in constraint 2 values is smaller with LB-SGD than with IPO, particularly when the initial point is near the boundary.

Our experiment confirmed that LB-SGD achieves safe exploration while converging to the optimal policy efficiently. Meanwhile, in Figure 1, we depict the policy obtained from our algorithm. Although it bears a resemblance to the optimal policy, it is less deterministic to circumvent the red rectangles. As expected, ensuring these guarantees necessitates a higher number of samples per iteration near the boundary for accurate estimates (see Figure 2). It would be interesting to determine whether this sample complexity is inherent to our algorithm and its analysis or if it is a consequence of the safe exploration requirement.

E.3 Discussion

Due to approximation errors, all algorithms—including IPO, RPG-PD, and LB-SGD—may occasionally take a bad step and produce an infeasible iterate π_{θ_t} . While sometimes the gradient update remains feasible, allowing the algorithm to recover from its bad step automatically, in other cases, a recovery method becomes necessary. In our experiments, we implement recovery by reverting to previous iterates and increasing the sample complexity while simultaneously decreasing the stepsize at previous iterates. This ensures the safe exploration of subsequent iterates with high probability.

F ESTIMATION OF SMOOTHNESS PARAMETER

In (Yuan et al., 2022, Proof of Lemma 4.4), the second order of the value functions $\nabla^2 V_i^\theta(\rho)$ is computed as

$$\begin{aligned} & \nabla^2 V_i^\theta(\rho) \\ &= \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) \left(\left(\sum_{k=0}^t \nabla_\theta^2 \log \pi_\theta(a_k | s_k) \right) + \left(\sum_{k=0}^t \nabla_\theta \log \pi_\theta(a_k | s_k) \right) \left(\sum_{k=0}^t \nabla_\theta \log \pi_\theta(a_k | s_k) \right)^T \right) \right]. \end{aligned}$$

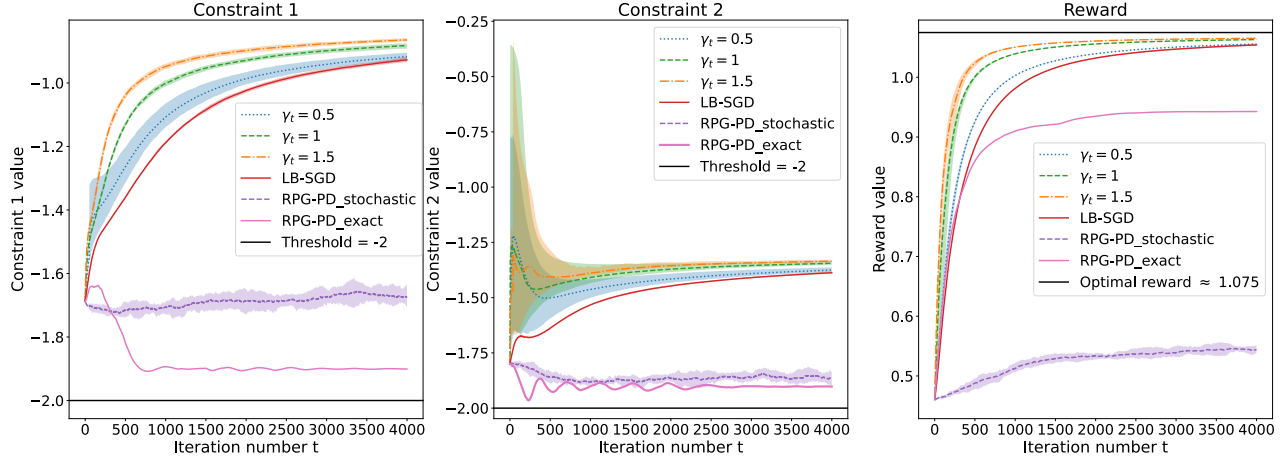


Figure 3: The average performance comparison between the IPO algorithm (Liu et al., 2020a) with stepsizes $\gamma_t = 0.5, 1, 1.5$, RPG-PD using a regularization parameter $\tau = 0.1$ and stepsize chosen according to (Ding et al., 2024, Theorem 2) with $b = 0.1$, and LB-SGD is shown below. The lines represent the median values from 10 independent experiments, while the shaded areas illustrate the 10% and 90% percentiles calculated from 10 different random seeds.

Using the information of n truncated trajectories, with a fixed horizon H , denoted as $\tau_j := \left(s_t^j, a_t^j, \left\{r_i(s_t^j, a_t^j)\right\}_{i=0}^m\right)_{t=0}^{H-1}$, we can estimate $\nabla^2 V_i^\theta(\rho)$ by the Monte-Carlo method as

$$\begin{aligned} & \hat{\nabla}^2 V_i^\theta(\rho) \\ &= \frac{1}{n} \sum_{j=1}^n \left[\sum_{t=0}^H \gamma^t r_i(s_t^j, a_t^j) \left(\left(\sum_{k=0}^t \nabla_\theta^2 \log \pi_\theta(a_k^j | s_k^j) \right) + \left(\sum_{k=0}^t \nabla_\theta \log \pi_\theta(a_k^j | s_k^j) \right) \left(\sum_{k=0}^t \nabla_\theta \log \pi_\theta(a_k^j | s_k^j) \right)^T \right) \right]. \end{aligned}$$

We estimate the smoothness parameter M_i for the value function $V_i^\theta(\rho)$ as

$$M_i = \|\hat{\nabla}^2 V_i^\theta(\rho)\|.$$