A Hessian-Free Actor-Critic Algorithm for Bi-Level Reinforcement Learning with Applications to LLM Fine-Tuning

Sihan Zeng* **Sujay Bhatt*** **Sumitra Ganesh*** **Alec Koppel†**

*JPMorgan AI Research † Johns Hopkins University Applied Physics Laboratory

Abstract

We study a structured bi-level optimization problem where the upper-level objective is a smooth function and the lower-level problem is policy optimization in a Markov decision process (MDP). The upper-level decision variable parameterizes the reward of the lower-level MDP, and the upper-level objective depends on the optimal induced policy. Existing methods for bi-level optimization and RL often require second-order information, impose strong regularization at the lower level, or inefficiently use samples through nested-loop procedures. In this work, we propose a single-loop, first-order actor-critic algorithm that optimizes the bi-level objective via a penalty-based reformulation. We introduce into the lower-level RL objective an attenuating entropy regularization, which enables asymptotically unbiased upper-level hyper-gradient estimation without solving the unregularized RL problem exactly. We establish the finite-time and finite-sample convergence of the proposed algorithm to a stationary point of the original, unregularized bi-level optimization problem through a novel lower-level residual analysis under a special type of Polyak–Łojasiewicz condition. We validate the performance of our method through experiments on a GridWorld goal position problem and on happy tweet generation through reinforcement learning from human feedback (RLHF).

1 INTRODUCTION

We study bi-level reinforcement learning (RL), a structured bi-level optimization program in which the upper-level decision variable determines the reward function of a lower-level RL problem, and the upper-level objective is evaluated under the lower-level optimal policy. This framework abstracts a wide range of applications where the reward must be tuned to achieve high-level goals while the underlying policy adapts to the reward. Examples include reward shaping [Ng et al., 1999, Hu et al., 2020], inverse RL [Zhang et al., 2025], multi-agent incentive design [Ma et al., 2025], contract design [Zhu et al., 2023], synthetic data generation [Wu et al., 2024], and, notably, reinforcement learning from human feedback (RLHF), one of the central paradigms for training and fine-tuning large language models [Chakraborty et al., 2024, Ye et al., 2025].

Despite a recent surge of interest in bi-level optimization, bi-level RL remains challenging both in theory and practice, due to the non-convexity at the lower level and the difficulty of estimating the upper-level hyper-gradient. Existing gradient-based approaches to bi-level optimization largely fall into two categories. The first leverages the implicit function theorem to derive the upper-level hyper-gradient [Ghadimi and Wang, 2018] and then applies iterative gradient descent along this direction. However, since the hyper-gradient depends on the Jacobian and Hessian of the lower-level objective, these methods are difficult to apply in bi-level RL, where second-order information either requires oracle access to the transition model or is prohibitively expensive to estimate from trajectory samples.

Another approach replaces the lower-level optimality condition with an explicit penalty term [Kwon et al., 2023, Shen and Chen, 2023], enabling an alternative expression of the hyper-gradient that depends only on first-order information. While this approach bypasses Hessian and Jacobian estimation, existing convergence analyses require strong structural assumptions on the lower-level problem, most commonly strong convexity or the Polyak–Łojasiewicz (PL) condition. In RL, however, these assumptions generally fail: the policy optimization objective may satisfy a weaker "gradient domination" property [Agarwal et al., 2021, Mei et al., 2020], but strong convexity and the PL condition do not hold under common policy parameterizations. As a result, existing guarantees in bi-level optimization apply

Work done while Alec Koppel was at JPMorgan AI Resarch.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

	Lower-Level Structure	Single Loop	Only Using First-Order Information	Sample Complexity	Anytime Valid
Kwon et al. [2023]	Strong convexity	✓	✓	$\widetilde{\mathcal{O}}(\epsilon^{-3.5})^{\dagger}$	✓
Shen and Chen [2023]	PL condition	✗	✓	N/A (Iteration complexity derived)	-
Xiao and Chen [2024]	PL condition	✓	✓	N/A (Deterministic setting studied)	-
Shen et al. [2024]	Regularized RL	✗	✓	N/A (Iteration complexity derived)	-
Chakraborty et al. [2024]	Regularized RL	✗	✗	N/A (Iteration complexity derived)	-
Thoma et al. [2024]	Regularized RL	✗	✓	$\widetilde{\mathcal{O}}(\epsilon^{-4})$	✗
Xiao et al. [2025]	Strong convexity	✗	✓	N/A	✗
Chen et al. [2024]	PL condition	✗	✓	$\widetilde{\mathcal{O}}(\epsilon^{-3})$	✗
Yang et al. [2025]	Regularized RL	✗	✓	$\widetilde{\mathcal{O}}(\epsilon^{-3.5})$	✗
Gaur et al. [2025]	Regularized RL	✗	✓	$\widetilde{\mathcal{O}}(\epsilon^{-3})$	✗
This work	**Regularized RL**	✓	✓	$\widetilde{\mathcal{O}}(\epsilon^{-3})$	✓
This work	**Original RL**	✓	✓	$\widetilde{\mathcal{O}}(\epsilon^{-10})$	✓

Table 1: Assumption, structure, and sample complexity (measured by squared gradient norm) of existing algorithms for bi-level optimization and RL. †This is the complexity of the standard F^{2}SA (Fully First-order Stochastic Approximation) algorithm. Kwon et al. [2023] also proposes a momentum-based algorithm (F^{3}SA) that achieves a better rate. For a fair comparison with other non-momentum algorithms in the table, we report the complexity of F^{2}SA. Notably, Yang et al. [2023] further improve the design and analysis of momentum-based algorithms, achieving a complexity of $\mathcal{O}(\epsilon^{-1.5})$, where the key innovation is estimating the lower-level Hessian via finite differences.

only to regularized or restricted settings of RL, leaving open the question of whether we can provably solve original, unregularized bi-level RL. Moreover, existing penalty-based methods are often implemented in a nested-loop fashion, repeatedly solving the lower-level problem to high precision before each upper-level update to ensure stability. The price of this stability is inefficient use of samples in practice and limited scalability.

Our work addresses this research gap. We advance penalty-based approaches in the context of RL and propose a single-loop actor-critic algorithm that provably converges to a stationary point of the bi-level RL objective. The core idea is to enforce the PL condition at the lower level through entropy regularization, and then gradually adjust its weight so that the regularized problem asymptotically recovers the original one. Two important technical innovations enable the finite-time and finite-sample analysis for the algorithm. First, we tightly characterize the iteration-wise convergence of the lower-level optimality error for the regularized RL objective, under a time-varying upper-level decision variable which itself is updated in the same loop on a slower time scale. Our innovation is a novel lower-level residual decomposition scheme under the PL condition, which generalizes the technique used in Kwon et al. [2023] under lower-level strong convexity. Combined with the penalty reformulation, this innovation allows us to establish the convergence of a fully single-loop penalty-based algorithm in the lower-level PL/regularized RL setting, achieving a convergence rate that surpasses the best known rate derived in Kwon et al. [2023] under lower-level strong convexity.

Our goal is to optimize the original, unregularized bi-level RL objective. While a large regularization accelerates the solution of the regularized problem, it also enlarges the discrepancy between the regularized and unregularized optima. Our second innovation addresses this trade-off by dynamically decaying the regularization weight, allowing the algorithm to track the regularized optima as the regularized problem gradually approaches the original one. A key challenge arises here from the time-varying lower-level landscape, which we overcome via a careful multi-time-scale stochastic approximation analysis that balances the change rate of the landscape with the algorithm step sizes. Below we detail and re-iterate our innovation and main contributions.

Main Contributions

• We establish several fundamental structural properties of bi-level RL, linking the problem to its entropy-regularized counterpart. In particular, we show that as the regularization weight decays, the optimizer of the entropy-regularized RL objective converges to the unique entropy-maximizing policy within the set of optimizers of the unregularized objective, and we provide a bound on this rate of convergence. We also provide bounds on the difference between the bi-level objective (and its gradient) with and without lower-level regularization. Absent in the prior work, this type of structural analysis lays the foundation for using the regularized objective as a faithful surrogate for the original formulation, and offers insights of potential independent interest to the broader fields of RL and bi-level optimization.

• We present a sample-based, single-loop bi-level RL algorithm and characterize its finite-sample complexity. The algorithm optimizes a regularized bi-level RL objective, while the regularization weight dynamically decays over time. We prove that this algorithm converges to a stationary point of the original bi-level objective with a sample complexity of

$\mathcal{O}(\epsilon^{-10})$ through a novel five-time-scale analysis, carefully balancing the regularization weight with the update rates of the dual variable, upper-level decision variable, policy iterates, and value function estimates. To our knowledge, this is the first algorithm that provably solves the unregularized bi-level RL problem, and the first that enjoys finite-time and finite-sample guarantees.

• We show that our algorithm can be instantiated with a constant regularization weight to solve the corresponding regularized bi-level RL problem with sample complexity $\mathcal{O}(\epsilon^{-3})$. This matches the state-of-the-art complexity of a comparable nested-loop method under lower-level PL condition [Chen et al., 2024], and improves over the $\mathcal{O}(\epsilon^{-3.5})$ complexity of a single-loop algorithm under strong convexity Kwon et al. [2023].

• We verify convergence of the proposed algorithm through experiments in a synthetic RL environment and in finetuning a (small) language model for happy tweet generation, demonstrating that it converges to a substantially better solution than comparable baselines. We also evaluate against a variant of the proposed algorithm with a fixed regularization. The results show that in general solving the regularized problem leads to sub-optimal solutions of the original one.

1.1 Related Work

Our work contributes to the increasing volume of literature on bi-level optimization. Here we discuss the most relevant papers on first-order methods to give context to our contributions, and provide a complete literature comparison with other first- and second-order methods in Table 1.

Kwon et al. [2023], Shen and Chen [2023] first introduced the penalty reformulation in bi-level optimization to bypass the estimation of second-order terms when evaluating the bi-level hyper-gradient. Leveraging the reformulation, Kwon et al. [2023] developed a fully first-order algorithm and showed that, under lower-level strong convexity, it converges to a stationary point of the bi-level objective with sample complexity $\tilde{\mathcal{O}}(\epsilon^{-3.5})$. Shen and Chen [2023] considered lower-level objectives satisfying the PL condition, relaxing the strong convexity assumption. The authors proposed a nested-loop algorithm and established its convergence rate as a function of outer-loop iterations, but the complexity with respect to the total number of iterations or samples remains unknown.

The penalty reformulation has inspired several works in bi-level RL [Shen et al., 2024, Yang et al., 2025, Gaur et al., 2025], which derived first-order expressions for the hyper-gradient and proposed ways of estimating them directly from environment samples. For technical tractability, these works focused on regularized bi-level RL, where the (fixed) regularization induces the PL condition for the policy optimization objective. Their analyses largely mirror those of generic bi-level optimization under a lower-level PL condition and do not solve the original unregularized problem. The quality of a regularized solution may be highly undesirable when evaluated under the original objective, which we illustrate with an example in Section 5. In addition, due to their nested-loop structure, the algorithms in these prior works require a target precision to be specified in advance to determine the number of inner-loop iterations, making the convergence guarantees not *anytime valid*: there is no guarantee that the optimality gap decreases monotonically after every iteration, and running the algorithm beyond the prescribed number of iterations does not further reduce the gap. Our work exactly addresses these limitations by proposing a single-loop actor–critic algorithm that converges to a stationary point of the unregularized bi-level RL problem through appropriate attenuation of the regularization, and we establish anytime valid finite-sample guarantees without requiring a pre-specified target precision.

2 FORMULATION

Consider an infinite-horizon discounted-reward MDP defined as $\mathcal{M}_x = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r_x, \gamma)$, where $x \in \mathbb{R}^d$ is an exogenous control parameter. Under a fixed x, $\mathcal{M}_x$ is a standard MDP. The state space $\mathcal{S}$ and action space $\mathcal{A}$ are assumed to be finite. The transition kernel is denoted by $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, and we use $\mathcal{P}(s' \mid s, a)$ to represent the probability that the next state is s' when an agent takes action a in state s. The reward function $r_x : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a function of x. The discount factor is denoted by $\gamma \in (0, 1)$. Our paper considers the setting where the transition kernel is independent of x, motivated by applications such as RLHF and reward shaping, where the exogenous variable modulates the reward function but not the system dynamics.

An agent learning in this MDP may not directly observe x, and takes actions according to a policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$, which we can represent as a table $\Delta_{\mathcal{A}}^{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$. Given a control-policy pair (x, π), we measure its performance in state s by the value function

$$V^{x,\pi}(s) \triangleq \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k r_x(s_k, a_k) \mid s_0 = s\Big],$$

where the expectation in the first equation is taken over the trajectory $a_k \sim \pi(\cdot \mid s_k), s_{k+1} \sim \mathcal{P}(\cdot \mid s_k, a_k)$ (we similarly omit a_k, s_{k+1} in the expectations below).

Let $d_s^\pi \in \Delta_{\mathcal{S}}$ and $d_\rho^\pi \triangleq \mathbb{E}_{s\sim\rho}[d_s^\pi] \in \Delta_{\mathcal{S}}$ denote the discounted visitation distributions under initial state s and the initial state distribution $\rho \in \Delta_{\mathcal{S}}$, respectively

$$d_s^\pi(s') \triangleq (1-\gamma)\mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k \mathbf{1}(s_k = s') \mid s_0 = s\Big].$$

We define the expected cumulative reward under (x, π)

$$J(x, \pi) \triangleq \mathbb{E}_{s\sim\rho}[V^{x,\pi}(s)] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^\pi, a\sim\pi(\cdot|s)}[r_x(s, a)].$$

If x were fixed, our goal would be to find a policy that maximizes $J(x, \pi)$. In the meantime, the exogenous controller has its own objective to optimize, anticipating the best response from the policy optimization agent. We denote the controller's objective by $f : \mathbb{R}^d \times \Delta_{\mathcal{A}}^{\mathcal{S}} \to \mathbb{R}$. Let $\Pi^\star(x)$ be the set of optimal policies under control x, which we note may not be singleton, and g be a function that maps $\Pi^\star(x)$ to a unique optimal policy within the set (we will shortly introduce g). The controller's optimization problem can then be formulated as the following bi-level program

$$\begin{aligned} &\min_{x \in \mathbb{R}^d} f(x, g(\Pi^\star(x))) && \textbf{Upper-Level} \\ &\text{s.t. } \Pi^\star(x) \triangleq \operatorname*{argmax}_{\pi} J(x, \pi) && \textbf{Lower-Level RL} \end{aligned} \tag{1}$$

Our goal in this paper is to solve (1). This is a challenging problem, as the lower-level objective lacks strong structural properties and may not admit a unique solution. To introduce additional structure and enhance the solvability, we add entropy regularization into the lower-level objective, which leads to solution uniqueness and a strong form of "gradient domination". We stress that the entropy-regularized formulation serves only as an intermediate tool – our ultimate aim remains to solve the original, unregularized problem in (1).

Remark 1 *We discuss how the bi-level RL framework models RLHF, along with experimental results, in Section 5. More applications of bi-level RL with concrete problem formulations are presented in Appendix B.*

2.1 Entropy Regularization

We discuss the regularized objective and its structural properties. Given (x, π) and regularization weight τ, we define the regularized value function $V_\tau^{x,\pi} \in \mathbb{R}^{|\mathcal{S}|}$ and expected cumulative reward J_τ

$$\begin{aligned} V_\tau^{x,\pi}(s) &\triangleq \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k (r_x(s_k, a_k) - \tau \log \pi(a_k \mid s_k)) \,|\, s_0 = s\Big] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_s^\pi,\, a' \sim \pi(\cdot|s')}[r_x(s', a') + \tau E(\pi, s')], \end{aligned} \tag{2}$$

$$J_\tau(x, \pi) \triangleq \mathbb{E}_{s \sim \rho}[V_\tau^{x,\pi}(s)], \tag{3}$$

where $E(\pi, s) = -\sum_a \pi(a \mid s) \log \pi(a \mid s)$ is the entropy function. Under regularization weight $\tau \leq 1$, we have $|V_\tau^{x,\pi}(s)| \leq B_V$ for all x, π, s, where $B_V = \frac{1+\log|\mathcal{A}|}{1-\gamma}$.

If the initial state distribution has a full support, an assumption we will shortly introduce and impose throughout the paper, then the optimizer of $J_\tau(x, \cdot)$ is unique for any $\tau > 0$. We define the operator $\pi_\tau^\star : \mathbb{R}^d \to \Delta_{\mathcal{A}}^{\mathcal{S}}$, which maps a control variable to the optimal policy that it induces

$$\pi_\tau^\star(x) \triangleq \operatorname*{argmax}_{\pi} J_\tau(x, \pi), \quad \forall x \in \mathbb{R}^d. \tag{4}$$

As we use the regularized RL problem to approximate the original one, it is important to understand how $\pi_\tau^\star(x)$ relates to $\Pi^\star(x)$. We make the connection in Lemma 1, under the following assumption on initial state distribution and ergodicity. The assumption is commonly made in the RL literature to guarantee that the Markov chain of states under any policy has a unique, well-defined stationary distribution [Mei et al., 2020, Wu et al., 2020, Khodadadian et al., 2022].

Assumption 1 (Sufficient Exploration) *The initial state distribution ρ is bounded away from zero, i.e., there exists a constant $\rho_{\min} > 0$ such that $\rho(s) \geq \rho_{\min}$ for all $s \in \mathcal{S}$. Additionally, for any π, the Markov chain $\{s_t\}$ generated by P^π following $s_{t+1} \sim P^\pi(\cdot \mid s_t)$ is ergodic.*

Lemma 1 *Let $\pi^\star(x)$ denote the optimal policy for the unregularized MDP with the largest (weighted) entropy*

$$\pi^\star(x) \triangleq \operatorname*{argmax}_{\pi \in \Pi^\star(x)} \mathbb{E}_{s \sim d_\rho^\pi}[E(\pi, s)]. \tag{5}$$

Then, under Assumption 1, it holds that $\pi^\star(x)$ is unique for all x and is the limit point of $\{\pi_\tau^\star(x)\}_\tau$

$$\pi^\star(x) = \lim_{\tau \to 0} \pi_\tau^\star(x).$$

The uniqueness of $\pi^\star(x)$ is not obvious and does not directly follow from known results in convex optimization, since $\Pi^\star$ is not a convex set and the weighted entropy objective in (5) is non-concave (see Lemma 3.1 of Hazan et al. [2019] for a proof of non-concavity). Our proof of Lemma 1, presented in Appendix E.1, exploits the strict concavity of the unweighted entropy function $E(\cdot, s)$ in the interior of the simplex, as well as the fact that $\Pi^\star(x)$, though non-convex, is a connected set with special structure [Zeng et al., 2023].

Our objective in this work is to solve the problem below, which corresponds to (1) with g mapping a set to the (unique) element of the set maximizing the weighted entropy

$$\min_x \ \Phi(x) \triangleq f(x, \pi^\star(x)). \tag{6}$$

Directly solving (6) is challenging, as the lack of strong structure in $J(x, \cdot)$ makes it difficult to find $\pi^\star(x)$. To bypass the challenge, we introduce the following regularized version of the objective, which serves as an important intermediary in our analysis. Conceptually, the algorithm to be introduced optimizes the regularized objective Φ_τ as the regularization attenuation drives Φ_τ toward Φ.

$$\min_x \Phi_\tau(x) \triangleq f(x, \pi_\tau^\star(x)). \tag{7}$$

3 ALGORITHM DEVELOPMENT

In this section we develop a single-loop first-order algorithm that optimizes Φ_τ while gradually decaying τ to zero, thereby recovering the solution to (6). The algorithm operates under stochastic gradient samples of the upper-level objective, as well as state transition and reward samples from

the lower-level MDP. We design the algorithm based on a penalty reformulation recently developed in [Kwon et al., 2023] for solving bi-level problems with lower-level strong convexity. Our analysis of the algorithm overcomes the unaddressed challenges specific to bi-level RL, namely, the lower-level non-convexity and the biased, non-i.i.d. stochastic gradients estimated from state-action-reward samples from an MDP. We begin by presenting an overview of the reformulation in our context.

3.1 Preliminaries – Penalty Reformulation

Our goal is to solve (7) via (stochastic) gradient descent. By the implicit function theorem [Ghadimi and Wang, 2018], $\nabla_x \Phi_\tau(x)$ admits the closed-form expression below when $\nabla^2_{\pi,\pi} J_\tau(x, \pi^\star(x))$ is invertible.

$$\nabla_x \Phi_\tau(x) = \nabla_x f(x, \pi^\star_\tau(x)) + \nabla_\pi f(x, \pi^\star_\tau(x)) \frac{\partial \pi^\star_\tau(x)}{\partial x} \quad (8)$$

$$= \nabla_x f(x, \pi^\star_\tau(x)) - \nabla^2_{x,\pi} J_\tau(x, \pi^\star_\tau(x)) \cdot \nabla^2_{\pi,\pi} J_\tau(x, \pi^\star_\tau(x))^{-1} \cdot \nabla_\pi f(x, \pi^\star_\tau(x)) \quad (9)$$

Obtaining unbiased samples of $\nabla_x \Phi_\tau(x)$ based on (9), however, poses significant challenges, as the expression depends on second-order Jacobian and Hessian terms that cannot be efficiently estimated from state–reward samples. The penalty reformulation is designed to provide an alternative approach of obtaining (asymptotically) unbiased gradient estimates, only requiring first-order information.

Recall the definition of $\pi^\star_\tau(x)$ in (4). We can re-write (7) as follows by introducing a constraint

$$\min_{x,\pi} f(x, \pi) \quad \text{s.t. } J_\tau(x, \pi^\star_\tau(x)) - J_\tau(x, \pi) \le 0. \quad (10)$$

Given a positive constant w, we define

$$\mathcal{L}_{w,\tau}(x, \pi) \triangleq f(x, \pi) + \frac{1}{w}\big(J_\tau(x, \pi^\star_\tau(x)) - J_\tau(x, \pi)\big), \quad (11)$$

$$\Phi_{w,\tau}(x) \triangleq \min_\pi \mathcal{L}_{w,\tau}(x, \pi) = \min_\pi f(x, \pi) + \frac{1}{w}\big(J_\tau(x, \pi^\star_\tau(x)) - J_\tau(x, \pi)\big). \quad (12)$$

We can regard $\mathcal{L}_{w,\tau}$ as the Lagrangian associated with (10), in which $1/w$ plays the role of the dual variable. To solve (10), it may be tempting to find a minimax saddle point of the Lagrangian using gradient descent ascent. However, as pointed out in Kwon et al. [2023], the solution of (10) is only attained in the limit as the dual variable becomes infinitely large (i.e., $w = 0$). This motivates us to treat w as a parameter governed by a prescribed decay schedule towards zero, rather than as a dual variable updated via gradient ascent. It is known from Kwon et al. [2023][Lemma 3.1] that $\nabla_x \Phi_{w,\tau}(x)$ admits the following expression involving only first-order terms[1]

$$\nabla_x \Phi_{w,\tau}(x) = \nabla_x \mathcal{L}_{w,\tau}(x, \pi^\star_{w,\tau}(x))$$

$$= \nabla_x f(x, \pi^\star_{w,\tau}(x)) + \frac{1}{w}\big(\nabla_x J_\tau(x, \pi^\star_\tau(x)) - \nabla_x J_\tau(x, \pi^\star_{w,\tau}(x))\big), \quad (13)$$

[1]We follow the convention and use $\nabla_x \mathcal{L}_{w,\tau}(x, \pi^\star_{w,\tau}(x))$ to denote the partial gradient with respect to x evaluated at $(x, \pi^\star_{w,\tau}(x))$, i.e., $\nabla_x \mathcal{L}_{w,\tau}(x, \pi^\star_{w,\tau}(x)) = \nabla_x \mathcal{L}_{w,\tau}(x, \pi) \mid_{\pi = \pi^\star_{w,\tau}(x)}$. The same principle will be used for other functions, such as f and J_τ.

Algorithm 1 Single-Loop Actor-Critic Algorithm for Bi-Level RL

1: **Initialize:** control variable x_0, policy parameters θ_0 and $\theta_0^{\mathcal{L}}$, value function estimates $\hat{V}_0, \hat{V}_0^{\mathcal{L}} \in \mathbb{R}^{|\mathcal{S}|}$
2: **for** iteration $k = 0, 1, 2, \ldots$ **do**
3: **Trajectory 1**: With probability $1-\gamma$, restart the trajectory by taking $s_{k+1} \sim \rho$. With probability γ, continue following the current trajectory. Take action $a_k \sim \pi_{\theta_k}(\cdot \mid s_k)$, receive rewards $r_{x_k}(s_k, a_k)$, and observe the next state $s_{k+1} \sim \mathcal{P}(\cdot \mid s_k, a_k)$.
4: **Trajectory 2**: With probability $1-\gamma$, restart the trajectory by taking $\bar{s}_{k+1} \sim \rho$. With probability γ, continue following the current trajectory. Take action $\bar{a}_k \sim \pi_{\theta_k^{\mathcal{L}}}(\cdot \mid \bar{s}_k)$, receive rewards $r_{x_k}(\bar{s}_k, \bar{a}_k)$, and observe the next state $\bar{s}_{k+1} \sim \mathcal{P}(\cdot \mid \bar{s}_k, \bar{a}_k)$.
5: Observe/Obtain $\xi_k \sim \mu$.
6: Control variable update: **Upper-Level Update**

$$x_{k+1} = x_k - \zeta_k \Big(\widetilde{\nabla}_x f(x_k, \pi_{\theta_k^{\mathcal{L}}}, \xi_k) + \frac{1}{w_k}\big(\nabla_x r_{x_k}(s_k, a_k) - \nabla_x r_{x_k}(\bar{s}_k, \bar{a}_k)\big)\Big).$$

7: Policy update: **Lower-Level Update**

$$\theta_{k+1} = \theta_k + \alpha_k \nabla_\theta \log \pi_{\theta_k}(a_k \mid s_k) \cdot \big(r_{x_k}(s_k, a_k) - \tau_k \log \pi_{\theta_k}(a_k \mid s_k) + \gamma \hat{V}_k(s_{k+1})\big),$$

$$\theta_{k+1}^{\mathcal{L}} = \theta_k^{\mathcal{L}} + \alpha_k \Big(- w_k \widetilde{\nabla}_\theta f(x_k, \pi_{\theta_k^{\mathcal{L}}}, \xi_k) + \nabla_\theta \log \pi_{\theta_k^{\mathcal{L}}}(\bar{a}_k \mid \bar{s}_k) \cdot \big(r_{x_k}(\bar{s}_k, \bar{a}_k) - \tau_k \log \pi^{\mathcal{L}}_{\theta_k}(\bar{a}_k \mid \bar{s}_k) + \gamma \hat{V}_k^{\mathcal{L}}(\bar{s}_{k+1})\big)\Big),$$

$$\pi_{k+1} = \pi_{\theta_{k+1}}, \quad \pi^{\mathcal{L}}_{k+1} = \pi_{\theta^{\mathcal{L}}_{k+1}}.$$

8: Value function update:

$$\hat{V}_{k+1} = \Pi_{B_V}\Big(\hat{V}_k + \beta_k e_{s_k}\big(r_{x_k}(s_k, a_k) - \tau_k \log \pi_{\theta_k}(a_k \mid s_k) + \gamma \hat{V}_k(s_{k+1}) - \hat{V}_k(s_k)\big)\Big),$$

$$\hat{V}^{\mathcal{L}}_{k+1} = \Pi_{B_V}\Big(\hat{V}^{\mathcal{L}}_k + \beta_k e_{\bar{s}_k}\big(r_{x_k}(\bar{s}_k, \bar{a}_k) - \tau_k \log \pi^{\mathcal{L}}_{\theta_k}(\bar{a}_k \mid \bar{s}_k) + \gamma \hat{V}^{\mathcal{L}}_k(\bar{s}_{k+1}) - \hat{V}^{\mathcal{L}}_k(\bar{s}_k)\big)\Big).$$

9: **end for**

where we define $\pi^\star_{w,\tau}(x) \triangleq \operatorname{argmin}_\pi \mathcal{L}_{w,\tau}(x,\pi)$ for all $w,\tau>0$. Importantly, $\nabla_x \Phi_{w,\tau}(x)$ closely tracks $\nabla_x \Phi_\tau(x)$ – the distance between $\nabla_x \Phi_{w,\tau}(x)$ and $\nabla_x \Phi_\tau(x)$ scales linearly in w, a result which we establish later in Lemma 11.

3.2 Single-Loop Actor-Critic Algorithm

We introduce x_k as an estimate of the solution to the bi-level objective (k is the iteration index) and design an algorithm that iteratively carries out stochastic gradient descent on x_k in an approximate direction of $\nabla_x \Phi_{w_k,\tau_k}(x_k)$, estimated using online samples from the MDP. Here w_k and τ_k are time-varying penalty and regularization weights. As w_k, τ_k decay according to carefully designed schedules, the surrogate objective $\Phi_{w_k,\tau_k}(x_k)$ increasingly approximates $\Phi(x_k)$, allowing us to solve the original bi-level problem.

To estimate $\nabla_x \Phi_{w_k,\tau_k}$ based on (13), we need estimates of $\nabla_x J_{\tau_k}(x_k, \pi^\star_{\tau_k}(x_k))$ and $\nabla_x J_{\tau_k}(x_k, \pi^\star_{w_k,\tau_k}(x_k))$. Note that $\nabla_x J_\tau(x,\pi)$ can be expressed in the simple form below

$$\nabla_x J_\tau(x,\pi) = \mathbb{E}_{s\sim d_\rho^\pi,\, a\sim\pi(\cdot|s)}[\nabla_x r_x(s,a)]. \tag{14}$$

Given (14), if we had access to an oracle that generates $\pi = \pi^\star_{\tau_k}(x)$ for any x, we could obtain asymptotically unbiased samples of $\nabla_x J_{\tau_k}(x_k, \pi^\star_{\tau_k}(x_k))$ by simply generating a Markovian chain $\{s_k, a_k\}$ under $\pi = \pi^\star_{\tau_k}(x_k)$ and evaluating $\nabla_x r_{x_k}(s_k,a_k)$ along the trajectory. The same can be done to estimate $\nabla_x J_{\tau_k}(x_k, \pi^\star_{w_k,\tau_k}(x_k))$ if $\pi^\star_{w_k,\tau_k}(x_k)$ were available. However, $\pi = \pi^\star_{\tau_k}(x_k)$ and $\pi^\star_{w_k,\tau_k}(x_k)$ are solutions to (augmented) lower-level RL problems and cannot be directly accessed. To overcome the oracle unavailability, we introduce the iterates $\pi_k, \pi_k^{\mathcal{L}}$ as approximations of $\pi^\star_{\tau_k}(x_k), \pi^\star_{w_k,\tau_k}(x_k)$, and update them via another layer of stochastic gradient ascent.

Existing bi-level RL methods [Yang et al., 2025, Gaur et al., 2025] typically introduce a nested-loop algorithmic structure when estimating these optimal policies, ensuring that $\pi_k, \pi_k^{\mathcal{L}}$ from the inner loop fully converges to $\pi^\star_{\tau_k}(x_k), \pi^\star_{w_k,\tau_k}(x_k)$ up to a desired precision. However, such nested-loop algorithms are usually inconvenient to implement in practice and require setting the target precision in advance. While the convergence of nested-loop optimization algorithms may generally be (near) optimal in theory under carefully selected step sizes and inner-loop iterations, how to choose such parameters is often unclear in practical applications, making nested-loop algorithms substantially less sample efficient than their single-loop counterparts.

Our algorithm is completely single-loop and updates $\pi_k, \pi_k^{\mathcal{L}}$ at the same time as x_k, with a larger step size (i.e., on a faster time scale) that approximates a nested-loop procedure. Specifically, we maintain policy parameters $\theta_k, \theta_k^{\mathcal{L}}$ that encode $\pi_k, \pi_k^{\mathcal{L}}$ (with notation $\pi_k = \pi_{\theta_k}, \pi_k^{\mathcal{L}} = \pi_{\theta_k^{\mathcal{L}}}$) and iteratively refine them according to

$$\theta_{k+1} = \theta_k + \alpha_k \widetilde{\nabla}_\theta J_{\tau_k}(x_k, \pi_{\theta_k}),$$

$$\theta_{k+1}^{\mathcal{L}} = \theta_k^{\mathcal{L}} + \alpha_k \Big(\widetilde{\nabla}_\theta J_{\tau_k}(x_k, \pi_{\theta_k^{\mathcal{L}}}) - w_k \widetilde{\nabla}_\theta f(x_k, \pi_{\theta_k^{\mathcal{L}}}) \Big).$$

Here α_k is a step size properly balanced with the decay rates of w_k and τ_k, and $\widetilde{\nabla}_\theta f, \widetilde{\nabla}_\theta J_{\tau_k}$ denote stochastic samples of the true gradients with exact forms presented in line 7 of Algorithm 1. Note that $\nabla_\theta J_{\tau_k}$ admits the following closed-form expression, and can be estimated in an asymptotically unbiased way via an actor-critic approach.

$$\nabla_\theta J_\tau(x,\pi_\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\rho^{\pi_\theta}, a\sim\pi_\theta(\cdot|s), s'\sim\mathcal{P}(\cdot|s,a)} \tag{15}$$
$$\Big[\big(r_x(s,a) - \tau \log \pi_\theta(a|s) + \gamma V_\tau^{x,\pi_\theta}(s') \big) \nabla_\theta \log \pi_\theta(a|s) \Big]$$

Actor-critic methods sample stochastic gradients according to (15), replacing the unobservable value function with an estimate updated on an even faster time scale via temporal difference learning. Specifically, we introduce two variables $\hat{V}_k, \hat{V}_k^{\mathcal{L}}$ to track $V_{\tau_k}^{x_k,\pi_k}, V_{\tau_k}^{x_k,\pi_k^{\mathcal{L}}}$ and present their update rules in line 8 of Algorithm 1, where $\Pi_{B_V} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ denotes the element-wise projection of a vector to the interval $[0, B_V]$. The projection guarantees the stability of the value function estimates, and the interval contains the true (regularization) value function under weight $\tau_k \le 1$.

The algorithm can be described at an abstract level as follows. We perform stochastic gradient descent on x_k along the hyper-gradient direction. The hyper-gradient estimation relies on the solutions of lower-level RL problem and the penalty-augmented RL objective, which we obtain via a single-loop actor-critic method. Highlighted in blue in Algorithm 1, our updates of the lower-level variables $\theta_k, \theta_k^{\mathcal{L}}, \hat{V}_k, \hat{V}_k^{\mathcal{L}}$ follow the standard actor–critic procedure, with the key distinction that we incorporate entropy regularization and gradually attenuate its weight over time. Note that despite the resemblance of our actor–critic updates to existing algorithms, the analysis is significantly more challenging in the bi-level setting. In particular, the learning targets for the lower-level policies are non-stationary, evolving both with the penalty and regularization schedules and with the updates of the upper-level variable. We overcome this challenge with a novel error decomposition scheme that tightly links the sub-optimality gap of the lower-level RL problem to that of the bi-level objective, under the shifting landscape which becomes less structured over time as the penalty and regularization weights decay.

We present our method in Algorithm 1, where we represent the policies through tabular softmax parameterization[2], i.e., the parameter $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ encodes the policy π_θ as below

$$\pi_\theta(a \mid s) = \frac{\exp(\theta(s,a))}{\sum_{a'} \exp(\theta(s,a'))}.$$

[2]We consider the softmax parameterization for the purpose of mathematical analysis. The algorithm is compatible with any function approximation in practical implementations.

Algorithm 1 employs three step size parameters and two penalty/regularization weights, which are all time-decaying sequences: step size ζ_k for upper-level variable update, step size α_k for policy update, step size β_k for value function update, penalty weight w_k, and regularization τ_k. The step sizes are associated with the primal variable (x_k) update, and we need to choose $\zeta_k \ll \alpha_k \ll \beta_k$ to approximate the nested-loop dynamics, where we run a large number of value function updates per policy update and a large number of policy updates per upper-level variable update.

Remark 2 *As introduced in Section 2, our work assumes that the upper-level variable influences the lower-level objective J only through the reward function, while the transition kernel is fixed and independent of x. We make this assumption for the ease of expressing and estimating the gradient of the lower-level objective with respect to x as in* (14). *If the transition kernel $\mathcal{P}$ were a function of x, the penalty-based method discussed prior to* (14) *would still remain valid, but the gradient $\nabla_x J_\tau(x,\pi)$ would become more challenging to derive and estimate using samples. Extending the algorithm and analysis to the setting of x-dependent transition kernel can be an important future work.*

4 CONVERGENCE ANALYSIS

We start the section by introducing the technical assumptions.

Assumption 2 *The Hessian $\nabla^2_{\theta,\theta} J_\tau(x,\pi_\theta)$ is invertible for all x,θ and $\tau \geq 0$.*

Recall that the Hessian inverse is a component of the hypergradient (9). Although our algorithm does not directly involve the Hessian, the assumption importantly guarantees the differentiability of Φ_τ. Let $\sigma_{\min}(\cdot)$ denote the smallest *singular value* of a matrix. Assumption 2 implies that there exists a constant $\underline{\sigma} > 0$ such that

$$\sigma_{\min}\Big(\nabla^2_{\theta,\theta} J_\tau(x,\pi_\theta)\Big) \geq \underline{\sigma}, \quad \forall x,\theta,\tau. \tag{16}$$

Without loss of generality, we let $\underline{\sigma} \leq 1$ for the convenience of combining terms in the analysis. Note that (16) should not be confused with the strong convexity of $J_\tau(x,\pi_\theta)$ with respect to θ, which requires $\nabla^2_{\theta,\theta} J_\tau(x,\pi_\theta)$ to be positive definite, i.e. its smallest *eigenvalue* is positive.

Assumption 3 *The function f is differentiable, and we have access to unbiased stochastic gradient operators $\widetilde{\nabla}_x f(x,\pi,\xi), \widetilde{\nabla}_\pi f(x,\pi,\xi)$ and i.i.d. samples ξ from a distribution μ such that*

$$\mathbb{E}_{\xi\sim\mu}[\widetilde{\nabla}_x f(x,\pi,\xi)] = \nabla_x f(x,\pi),$$
$$\mathbb{E}_{\xi\sim\mu}[\widetilde{\nabla}_\pi f(x,\pi,\xi)] = \nabla_\pi f(x,\pi).$$

Also, there exists a constant $L_f < \infty$ such that $\forall x,x',\pi,\pi',\xi$

$$\|\widetilde{\nabla}_x f(x,\pi,\xi)\| \leq L_f, \quad \|\widetilde{\nabla}_\pi f(x,\pi,\xi)\| \leq L_f,$$
$$\|\widetilde{\nabla}_x f(x,\pi,\xi) - \widetilde{\nabla}_x f(x',\pi',\xi)\| \leq L_f(\|x-x'\| + \|\pi-\pi'\|),$$
$$\|\widetilde{\nabla}_\pi f(x,\pi,\xi) - \widetilde{\nabla}_\pi f(x',\pi',\xi)\| \leq L_f(\|x-x'\| + \|\pi-\pi'\|).$$

We also assume that the minimizer of $f(\cdot,\pi)$ exists, i.e., $f(x,\pi)$ never blows up to $-\infty$. Without loss of generality, we can shift the function such that $f(x,\pi) \geq 0$, $\forall x,\pi$.

Assumption 4 *There is a constant L_r such that $\forall s,a,x_1,x_2$*

$$|r_{x_1}(s,a) - r_{x_2}(s,a)| \leq L_r\|x_1 - x_2\|,$$
$$\|\nabla_x r_{x_1}(s,a) - \nabla_x r_{x_2}(s,a)\| \leq L_r\|x_1 - x_2\|,$$
$$\|\nabla^2_{x,x} r_{x_1}(s,a) - \nabla^2_{x,x} r_{x_2}(s,a)\| \leq L_r\|x_1 - x_2\|.$$

Assumptions 3 and 4 are standard regularity assumptions in the bi-level RL literature [Chakraborty et al., 2024, Gaur et al., 2025]. Assumption 3 imposes upper-level Lipschitz continuity, smoothness, and the ability to obtain reliable gradient samples, and Assumption 4 requires the reward function to be Lipschitz along with its gradient and Hessian. Comparable conditions on upper- and lower-level objectives are also commonly imposed by works on generic bi-level optimization [Kwon et al., 2023, Shen and Chen, 2023].

Assumption 5 (PL-Type Condition) *Recall the definition of $\pi^\star_{w,\tau}$ after* (13). *We assume that the minimizer $\pi^\star_{w,\tau}(x)$ is unique for all x and $w,\tau > 0$, and that there exist constants $C_L, \bar{w} > 0$ such that for all $\tau > 0$, $w \leq \bar{w}$, and x,θ*

$$\|\nabla_\theta \mathcal{L}_{w,\tau}(x,\pi_\theta)\|^2 \geq \frac{C_L\tau}{w}\Big(\mathcal{L}_{w,\tau}(x,\pi_\theta) - \mathcal{L}_{w,\tau}(x,\pi^\star_{w,\tau}(x))\Big). \tag{17}$$

This structural condition plays an important role and states that the Lagrangian defined in (12) satisfies the PL condition with respect to the policy parameter θ, while the PL constant attenuates as the regularization weight becomes smaller. While we directly impose (17) for convenience and for potential broader applicability of our analysis beyond bi-level RL, we note that the condition can be derived in the tabular setting for a proper range of w under assumptions of initial state distribution coverage (Assumption 1) and exploratory policy (i.e., there exists a constant $\underline{\pi} \in (0,1)$ such that $\pi_\theta(a \mid s) \geq \underline{\pi}$ for all θ that we will consider in this work). To see this, note that as $w \to 0$, $L_{w,\tau}(x,\pi)$ approaches a $1/w$-scaled (and shifted) version of $J_\tau(x,\pi)$, which is known to satisfy the PL condition with $C_L = \mathcal{O}\big(\rho^2_{\min}(\min_{s,a}\pi_\theta(a \mid s))^2\big)$ (see Mei et al. [2020][Lemma 15]). The scaling explains the dependence of the right-hand side of (17) on $1/w$. For sufficiently small w, the contribution of the f term in $L_{w,\tau}$ remains negligible, so the PL condition on J_τ continues to dominate and allows (17) to hold.

Theorem 1 *Suppose Assumptions 1-5 hold. Consider the iterates of Algorithm 1 under the step sizes and weights*

$$\zeta_k = \frac{\zeta_0}{(k+1)^{c_\zeta}}, \quad \alpha_k = \frac{\alpha_0}{(k+1)^{c_\alpha}},$$

$$\beta_k = \frac{\beta_0}{(k+1)^{c_\beta}}, \quad w_k = \frac{w_0}{(k+1)^{c_w}}, \quad \tau_k = \frac{\tau_0}{(k+1)^{c_\tau}},$$

with $c_\zeta = \frac{9}{10}, c_\alpha = \frac{1}{2}, c_\beta = \frac{1}{2}, c_w = \frac{3}{20}, c_\tau = \frac{1}{20}$ *and properly chosen* $\zeta_0, \alpha_0, \beta_0, w_0, \tau_0$. *We have for all* $k \geq 0$

$$\min_{t<k} \mathbb{E}[\|\nabla_x \Phi(x_t)\|^2] \leq \widetilde{\mathcal{O}}\left(\frac{1}{(k+1)^{1/10}}\right). \tag{18}$$

Theorem 1 establishes the convergence of the best iterate of Algorithm 1 to a stationary point of the bi-level RL objective, with rate $\widetilde{\mathcal{O}}(k^{-1/10})$. The algorithm draws two samples in each iteration, so (1) translates to a sample complexity of the same order. This is the first time an algorithm has been shown to provably solve the original, unregularized bi-level RL problem. The key technical insight and novelty enabling our analysis is 1) that we recognize the RL objective as one observing a regularization-dependent PL condition with the PL constant diminishing as regularization approaches zero, 2) a multi-time-scale stochastic approximation analysis that balances the decay of step sizes and w_k with that of τ_k, allowing the algorithm's convergence to be established under the challenge of a time-varying optimization landscape.

We may also instantiate Algorithm 1 with a constant regularization weight, thereby optimizing the regularized objective instead of the original one. In this regime, as the lower-level problem satisfies the PL condition with a fixed PL constant, we are able to establish a faster convergence rate.

Theorem 2 *Given any fixed regularization weight* τ_0*, i.e.,* $\tau_k = \tau_0$ *for all* $k \geq 0$*, consider the iterates of Algorithm 1 under the step sizes and penalty weight*

$$\zeta_k = \frac{\zeta_0}{(k+1)^{c_\zeta}}, \quad \alpha_k = \frac{\alpha_0}{(k+1)^{c_\alpha}},$$
$$\beta_k = \frac{\beta_0}{(k+1)^{c_\beta}}, \quad w_k = \frac{w_0}{(k+1)^{c_w}},$$

with $c_\zeta = \frac{2}{3}, c_\alpha = \frac{1}{2}, c_\beta = \frac{1}{2}, c_w = \frac{1}{6}$ *and properly chosen* $\zeta_0, \alpha_0, \beta_0, w_0$. *Under Assumptions 1-5, we have for all* k

$$\min_{t<k} \mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_t)\|^2] \leq \widetilde{\mathcal{O}}\left(\frac{1}{(k+1)^{1/3}}\right). \tag{19}$$

Theorem 2 shows that Algorithm 1 under a constant τ_0 converges to a stationary point of the regularized bi-level objective, and again (19) implies a sample complexity of the same order. Importantly, this rate surpasses that of the F^{2}SA algorithm in Kwon et al. [2023], which is $\widetilde{\mathcal{O}}(k^{-1/3.5})$ derived under the stronger assumption of lower-level strong convexity. We achieve the rate improvement under weaker lower-level structure by designing a novel error decomposition scheme, which allows us to tightly bound the residuals in the policy iterates based on the PL condition. Also note that the complexity of Algorithm 1 matches the best-known complexity of a nested-loop algorithm developed in Gaur et al. [2025] for solving the regularized bi-level RL problem.

5 EXPERIMENTAL RESULTS

We numerically verify the convergence of Algorithm 1 in two environments. The first is a synthetic bi-level RL problem of goal placement in a 10×10 grid, which is small-scale with fully known reward and transition kernel, allowing us to precisely evaluate $\Phi(x_k)$ along the upper-level variable trajectory $\{x_k\}$ for various algorithm. The second problem is to finetune a language model to generate tweets with positive sentiment, which is large-scale and involves neural networks as the function approximation.

GridWorld Goal Placement. The reward for the MDP is the negated distance between the current state and a goal position (which encourages the lower-level RL agent to reach the goal in fewest possible steps), whereas the goal is placed by the upper level decision variable. We design the upper-level objective, as a function of upper-level decision variable x and the optimal policy $\pi^\star(x)$, to penalize deviations of the goal from the center of the grid, while encouraging $\pi^\star(x)$ to have direction biases towards moving down and right. The environment is small enough that we can exactly compute the bi-level objective $\Phi(x)$ for any x, and we report $\Phi(x_k)$ produced by the algorithms. Details of the experimental setup are deferred to Appendix F.

In this first environment, we exactly implement the proposed method according to Algorithm 1 with carefully selected diminishing weights w_k, τ_k, with the purpose of verifying that the proposed algorithm indeed exhibits faster convergence than comparable first-order baselines in the tabular softmax setting where our theoretical guarantees are known to hold (given that the assumptions are satisfied).

Language Model Finetuning. The goal in this experiment is to train a language model to generate tweets of positive sentiment given pairwise human preference feedback. The problem can be formulated as a bi-level RL objective, where the upper-level variable x parameterizes a reward model and the lower-level variable θ corresponds to the parameters of a language model [Chakraborty et al., 2024].

$$\begin{aligned} \min_x \quad & \mathbb{E}_{(p_0,p_1,y)\sim\psi(\pi^\star(x))}[y \log P_x(p_0 > p_1) \\ & \qquad + (1-y)\log P_x(p_0 < p_1)] \\ \text{s.t.} \quad & \pi^\star(x) = \arg\max_\pi J(x,\pi) \end{aligned}$$

Here $(p_0, p_1, y) \sim \psi(\pi^\star(x))$ denotes the two sample trajectories p_0, p_1 (completed texts) generated independently using the policy $\pi^\star(x)$, on which a human determines a preference $y \in \{0, 1\}$, with 1 indicating p_0 is preferred over p_1 and 0 indicating the opposite. We model the preference probability P_x using the Bradley-Terry model [Bradley and Terry, 1952], calculated from the reward function as follows

$$P_x(p_0 > p_1) = \frac{\exp(r_x(p_0))}{\exp(r_x(p_0)) + \exp(r_x(p_1))}.$$

Conceptually, this objective means that the language model

is optimized to generate paragraphs of positive sentiment that achieve high rewards under the reward model, while the upper-level variable updates the reward model to align with pairwise human preferences.

We simulate human preference by a rule-based sentiment analyzer [Hutto and Gilbert, 2014], which assigns scores in the range $[-1, 1]$ to a completed body of text, with a larger value corresponding to more positive sentiment. We determine that $y = 1$ if the score of the first response is higher than the second, and $y = 0$ otherwise. To scale up the proposed algorithm we employ function approximation (small 4-layer transformer-based neural networks) to parameterize both the reward and language models. We set the regularization weight to 0 in this set of experiments for simplicity.

The bi-level objective Φ is not easily computable in this environment, and we report as a proxy the negated sentiment score (given by the rule-based analyzer) of paragraphs generated under policy π_{θ_k}.

In terms of baselines, a first natural choice is an alternating (partial) gradient descent–ascent algorithm, which is commonly adopted by empirical works in bi-level RL [Zhang et al., 2020, Hu et al., 2024]: we maintain iterates (x_k, θ_k) and update them in the direction of $\nabla_x f(x_k, \pi_{\theta_k})$ and $\nabla_\theta J(x_k, \pi_{\theta_k})$, respectively, in an alternating fashion. We refer to this approach as "Partial SGD", as the update direction for x_k is a partial component of the full hyper-gradient $\nabla_x \Phi(x_k)$. In the context of the second experiment, Partial SGD resembles the standard RLHF procedure: according to standard RLHF, we draw a large number of text samples from a pre-finetuned language model to train the reward model and then train the language model with the reward model frozen; in comparison, Partial SGD alternates between reward model and language model updates with every paired text sample draw. Note that Partial SGD may be stuck at sub-optimal solutions when the partial gradient is misaligned with the true hyper-gradient. This phenomenon is evident in Figure 1, where Partial SGD fails to achieve the minimum objective value. Standard RLHF also exhibits a substantial performance gap compared with the proposed algorithm, as it fails to properly account for the coupling between reward model training and policy optimization, an observation also made in Chakraborty et al. [2024].

We also compare against an algorithm that performs stochastic gradient descent with gradients estimated from the chain-rule expression (8), where $\frac{\partial \pi^\star(x)}{\partial x}$ is approximated via finite differencing. Specifically, for a scalar x, we approximate $\frac{\partial \pi^\star(x)}{\partial x} \approx \frac{\pi^\star(x+\epsilon) - \pi^\star(x-\epsilon)}{2\epsilon}$, where $\pi^\star(x+\epsilon)$ and $\pi^\star(x-\epsilon)$ are approximated by running an actor-critic algorithm within a large number of inner-loop iterations. This can be extended to vector or tensor x with simultaneous perturbation stochastic approximation [Spall, 2002]. We refer to this method as "Finite-Difference Approximation" in Figure 1 and note that it is sample-inefficient, owing to both

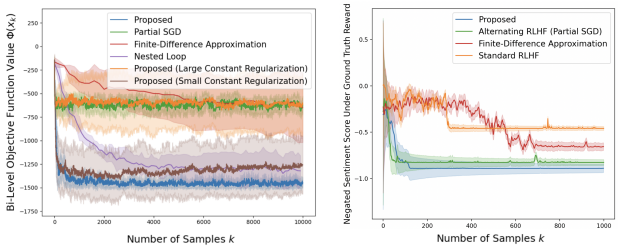

Figure 1: Algorithm Performance on GridWorld Goal Placement (Left) and Language Model Finetuning (Right)

the computational overhead and the inaccuracy inherent in finite-difference-based gradient estimation.

In the first experiment, we additionally compare with a nested-loop variant of Algorithm 1, where we run a large number of inner-loop iterations to approximately compute $\pi^\star_{w_k,\tau_k}(x_k), \pi^\star_{\tau_k}(x_k)$ as x_k is updated much less frequently in the outer loop. In spirit this nested-loop procedure resembles the algorithm proposed in Gaur et al. [2025]. It has been commonly observed in the literature that single-loop algorithms often outperform their nested-loop alternatives due to more efficient use of inner-loop samples, and we see a similar relative performance gap in our setting as well.

Finally, in the first experiment, we implement the proposed algorithm under fixed regularization weights, with the purpose of showing that in general solving the regularized problem leads to a sub-optimal solution of the original problem. Specifically, we test two variants: one with a relatively large regularization weight, which overly restricts the solution and prevents progress toward the true optimum, and another with a smaller regularization weight, which still falls short of the performance attained with a decaying regularization weight. These two variants are labeled "Proposed (Large Constant Regularization)" and "Proposed (Small Constant Regularization)" in Figure 1 (left).

6 CONCLUSION & FUTURE WORK

We present a novel single-loop, first-order actor-critic algorithm for bi-level RL. By introducing the entropy regularization, we enforce a special type of PL condition (with diminishing strength) at the lower level, enabling the algorithm to converge to a stationary point of the original, unregularized objective with a sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-1/10})$. A future direction is to investigate whether the convergence rate can be improved by incorporating variance reduction and/or momentum, under which the decision variables are updated along an averaged stochastic gradient direction rather than the instantaneous one. Momentum has been shown to improve the convergence rate of penalty-based bi-level optimization algorithms under lower-level strong convexity, giving promise that it may also lead to convergence acceleration under our weaker version of PL condition.

Disclaimer

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan") and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Souradip Chakraborty, Amrit S Bedi, Alec Koppel, Huazheng Wang, Dinesh Manocha, Mengdi Wang, and Furong Huang. Parl: A unified framework for policy alignment in reinforcement learning from human feedback. In *ICLR*, 2024.

Lesi Chen, Jing Xu, and Jingzhao Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 947–980. PMLR, 2024.

Mudit Gaur, Amrit Singh Bedi, Raghu Pasupathu, and Vaneet Aggarwal. On the sample complexity bounds in bilevel reinforcement learning. *arXiv preprint arXiv:2503.17644*, 2025.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.

Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, and Changjie Fan. Learning to utilize shaping rewards: A new approach of reward shaping. *Advances in Neural Information Processing Systems*, 33:15931–15941, 2020.

Zechen Hu, Daigo Shishika, Xuesu Xiao, and Xuan Wang. Bi-cl: A reinforcement learning framework for robots coordination through bi-level optimization. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 581–586. IEEE, 2024.

Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

Sajad Khodadadian, Thinh T Doan, Justin Romberg, and Siva Theja Maguluri. Finite-sample analysis of two-time-scale natural actor–critic algorithm. *IEEE Transactions on Automatic Control*, 68(6):3273–3284, 2022.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pages 18083–18113. PMLR, 2023.

Haoxiang Ma, Shuo Han, Ahmed Hemida, Jie Fu, et al. Adaptive incentive design for markov decision processes with unknown rewards. *Transactions on Machine Learning Research*, 2025.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*, pages 6820–6829. PMLR, 2020.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287. Citeseer, 1999.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, pages 30992–31015. PMLR, 2023.

Han Shen, Zhuoran Yang, and Tianyi Chen. Principled penalty-based methods for bilevel reinforcement learning and rlhf. In *International Conference on Machine Learning*, pages 44774–44799. PMLR, 2024.

Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, and Chao Mi. Hessian aided policy gradient. In *International conference on machine learning*, pages 5729–5738. PMLR, 2019.

James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3):332–341, 2002.

Vinzenz Thoma, Barna Pásztor, Andreas Krause, Giorgia Ramponi, and Yifan Hu. Contextual bilevel reinforcement learning for incentive alignment. *Advances in Neural Information Processing Systems*, 37:127369–127435, 2024.

Chia-Yuan Wu, Frank E Curtis, and Daniel P Robinson. Using synthetic data to mitigate unfairness and preserve

privacy in collaborative machine learning. *arXiv preprint arXiv:2409.09532*, 2024.

Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

Quan Xiao and Tianyi Chen. Unlocking global optimality in bilevel optimization: A pilot study. *arXiv preprint arXiv:2408.16087*, 2024.

Quan Xiao, Hui Yuan, AFM Saif, Gaowen Liu, Ramana Kompella, Mengdi Wang, and Tianyi Chen. A first-order generative bilevel optimization framework for diffusion models. *arXiv preprint arXiv:2502.08808*, 2025.

Yan Yang, Bin Gao, and Ya-xiang Yuan. Bilevel reinforcement learning via the development of hyper-gradient without lower-level convexity. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Achieving $\mathcal{O}(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. *Advances in Neural Information Processing Systems*, 36:39491–39503, 2023.

Ziyu Ye, Rishabh Agarwal, Tianqi Liu, Rishabh Joshi, Sarmishta Velury, Quoc V Le, Qijun Tan, and Yuan Liu. Reward-guided prompt evolving in reinforcement learning for llms. In *Forty-second International Conference on Machine Learning*, 2025.

Sihan Zeng, Malik Aqeel Anwar, Thinh T Doan, Arijit Raychowdhury, and Justin Romberg. A decentralized policy gradient approach to multi-task reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 1002–1012. PMLR, 2021.

Sihan Zeng, Thinh Doan, and Justin Romberg. Regularized gradient descent ascent for two-player zero-sum markov games. *Advances in Neural Information Processing Systems*, 35:34546–34558, 2022.

Sihan Zeng, Thinh Doan, and Justin Romberg. Connected superlevel set in (deep) reinforcement learning and its application to minimax theorems. *Advances in Neural Information Processing Systems*, 36:20146–20163, 2023.

Haifeng Zhang, Weizhe Chen, Zeren Huang, Minne Li, Yaodong Yang, Weinan Zhang, and Jun Wang. Bi-level actor-critic for multi-agent coordination. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7325–7332, 2020.

Ruijia Zhang, Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Understanding inverse reinforcement learning under overparameterization: Non-asymptotic analysis and global optimality. In *International Conference on Artificial Intelligence and Statistics*, pages 2944–2952. PMLR, 2025.

Banghua Zhu, Stephen Bates, Zhuoran Yang, Yixin Wang, Jiantao Jiao, and Michael I Jordan. The sample complexity of online contract design. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 1188–1188, 2023.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa with linear function approximation. *Advances in neural information processing systems*, 32, 2019.

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]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

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]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes. Provided in appendix.]
   (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). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable. Experiments are small-scale and take no more than an hour to run on a regular laptop.]

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. [Not Applicable. No such assets used.]
   (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]

Contents

1 INTRODUCTION **1**

1.1 Related Work . . . 3

2 FORMULATION **3**

2.1 Entropy Regularization . . . 4

3 ALGORITHM DEVELOPMENT **4**

3.1 Preliminaries – Penalty Reformulation . . . 5

3.2 Single-Loop Actor-Critic Algorithm . . . 6

4 CONVERGENCE ANALYSIS **7**

5 EXPERIMENTAL RESULTS **8**

6 CONCLUSION & FUTURE WORK **9**

A Frequently Used Notations, Equations, and Inequalities **14**

B Applications of Bi-Level Reinforcement Learning **17**

B.1 Reward Shaping . . . 17

B.2 Inverse Reinforcement Learning . . . 18

C Proof of Theorems **18**

C.1 Proof of Theorem 3 (replication of Theorem 1 under i.i.d. samples) . . . 21

C.2 Proof of Theorem 4 (replication of Theorem 2 under i.i.d. samples) . . . 23

D Proof of Propositions **25**

D.1 Proof of Proposition 1 . . . 25

D.2 Proof of Proposition 2 . . . 27

D.3 Proof of Proposition 3 . . . 28

D.4 Proof of Proposition 4 . . . 32

D.5 Proof of Proposition 5 . . . 36

E Proof of Supporting Results **38**

E.1 Proof of Lemma 1 . . . 38

E.2 Proof of Lemma 2 . . . 40

E.3 Proof of Lemma 3 . . . 40

E.4 Proof of Lemma 4 . . . 43

E.5 Proof of Lemma 5 . . . 43

E.6 Proof of Lemma 6 . . . 44

E.7 Proof of Lemma 7 47
E.8 Proof of Lemma 8 49
E.9 Proof of Lemma 9 50
E.10 Proof of Lemma 10 51
E.11 Proof of Lemma 11 53
E.12 Proof of Lemma 12 55
E.13 Proof of Lemma 13 56
E.14 Proof of Lemma 14 57
E.15 Proof of Lemma 15 57

F Simulation Details – GridWorld Goal Placement **61**

A Frequently Used Notations, Equations, and Inequalities

- Besides the value function defined in (2), we also define the regularized Q function and advantage function

$$\begin{aligned} Q_\tau^{x,\pi}(s,a) &\triangleq r_x(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}(s' \mid s,a) V_\tau^{x,\pi}(s), \\ A_\tau^{x,\pi}(s,a) &\triangleq Q_\tau^{x,\pi}(s,a) - \tau \log \pi(a \mid s) - V_\tau^{x,\pi}(s). \end{aligned} \tag{20}$$

- We define the filtration $\mathcal{F}_k \triangleq \{\xi_0, \cdots, \xi_k, s_0, \cdots, s_k, a_0, \cdots, a_k, \bar{s}_0, \cdots, \bar{s}_k, \bar{a}_0, \cdots, \bar{a}_k\}$.
- The PL condition in Assumption 5 implies quadratic growth, i.e. for any x

$$\mathcal{L}_{w,\tau}(x,\pi_\theta) - \mathcal{L}_{w,\tau}(x,\pi^\star_{w,\tau}(x)) \geq \frac{C_L\tau}{4w}\|\pi_\theta - \pi^\star_{w,\tau}(x)\|^2. \tag{21}$$

  In combination with (17), the inequality implies

$$\|\nabla_\theta \mathcal{L}_{w,\tau}(x,\pi_\theta)\| \geq \frac{C_L\tau}{2w}\|\pi_\theta - \pi^\star_{w,\tau}(x)\|. \tag{22}$$

- We introduce the following shorthand notations that abstract the update operators in Algorithm 1. For any $x, \pi, \pi^{\mathcal{L}}, \theta, V, s, a, s', \bar{s}, \bar{a}, \xi$, we define

$$D_w(x,\pi,\pi^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi) = \widetilde{\nabla}_x f(x,\pi^{\mathcal{L}},\xi) + \frac{1}{w}\Big(\nabla_x r_x(s,a) - \nabla_x r_x(\bar{s},\bar{a})\Big), \tag{23}$$

$$F_{w,\tau}(x,\theta,V,s,a,s',\xi) = \big(r_x(s,a) - \tau\log\pi_\theta(a\mid s) + \gamma V(s') - V(s)\big)\nabla_\theta \log\pi_\theta(a\mid s) - w\widetilde{\nabla}_\theta f(x,\pi_\theta,\xi), \tag{24}$$

$$G_\tau(x,\theta,V,s,a,s') = e_s\big(r_x(s,a) + \tau\log\pi_\theta(a\mid s) + \gamma V(s') - V(s)\big), \tag{25}$$

  where e_s is the indicator function, i.e., the entry s has a value of one and all other entries are zero.

  With (23)-(25), we can rewrite the updates of Algorithm 1

$$\begin{aligned} x_{k+1} &= x_k - \zeta_k D_{w_k}(x_k,\pi_k,\pi_k^{\mathcal{L}},s_k,a_k,\bar{s}_k,\bar{a}_k,\xi_k), \\ \theta_{k+1} &= \theta_k + \alpha_k F_{0,\tau_k}(x_k,\theta_k,\hat{V}_k,s_k,a_k,s_k',\xi_k), \\ \theta_{k+1}^{\mathcal{L}} &= \theta_k^{\mathcal{L}} + \alpha_k F_{w_k,\tau_k}(x_k,\theta_k^{\mathcal{L}},\hat{V}_k^{\mathcal{L}},\bar{s}_k,\bar{a}_k,\bar{s}_k',\xi_k), \\ \hat{V}_{k+1} &= \hat{V}_k + \beta_k G_{\tau_k}(x_k,\theta_k,\hat{V}_k,s_k,a_k,s_k'), \\ \hat{V}_{k+1}^{\mathcal{L}} &= \hat{V}_k^{\mathcal{L}} + \beta_k G_{\tau_k}(x_k,\theta_k^{\mathcal{L}},\hat{V}_k^{\mathcal{L}},\bar{s}_k,\bar{a}_k,\bar{s}_k'). \end{aligned}$$

  We also define the expected versions $\bar{D}, \bar{F}, \bar{F}$ of the (semi-)gradient operators above. The expectation is over stochastic samples from a proper stationary distribution.

$$\bar{D}_w(x,\pi,\pi^{\mathcal{L}}) \triangleq \mathbb{E}_{s\sim d_\rho^\pi, a\sim\pi(\cdot|s), \bar{s}\sim d_\rho^{\pi^{\mathcal{L}}}, \bar{a}\sim\pi^{\mathcal{L}}(\cdot|\bar{s})), \xi\sim\mu}[D_w(x,\pi,\pi^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)], \tag{26}$$

$$\bar{F}_{w,\tau}(x,\theta,V) \triangleq \mathbb{E}_{s\sim d_\rho^{\pi_\theta},a\sim\pi_\theta(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a),\xi\sim\mu}[F_{w,\tau}(x,\theta,V,s,a,s',\xi)], \tag{27}$$

$$\bar{G}_{\tau}(x,\theta,V) \triangleq \mathbb{E}_{s\sim d_\rho^{\pi_\theta},a\sim\pi_\theta(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a)}[G_{\tau}(x,\theta,V,s,a,s')]. \tag{28}$$

- It can be seen from (13) that the following relationship holds for any x

$$\nabla_x \Phi_{w,\tau}(x) = \bar{D}_w(x,\pi_\tau^\star(x),\pi_{w,\tau}^\star(x)). \tag{29}$$

  In addition, we have for any x,θ,τ

$$\bar{F}_{0,\tau}(x,\theta,V_\tau^{x,\pi_\theta}) = \nabla_\theta J_\tau(x,\pi_\theta), \tag{30}$$

$$\bar{F}_{w,\tau}(x,\theta,V_\tau^{x,\pi_\theta}) = w\nabla_\theta L_{w,\tau}(x,\pi_\theta). \tag{31}$$

- We define for $\tau > 0$

$$\ell_\tau(x) = J_\tau(x,\pi_\tau^\star(x)). \tag{32}$$

  Note that ℓ_τ is not to be confused with Φ_τ defined in (7), which is the regularized upper-level objective.

- While our goal is to quantify the convergence to a stationary point of the bi-level objective by the metric $\|\nabla_x\Phi(x_t)\|^2$ (or $\|\nabla_x\Phi_{\tau_0}(x_t)\|^2$ under a fixed regularization weight τ_0), the analysis relies on jointly bounding the convergence of all variables of Algorithm 1 through a coupled Lyapunov function, which combines the residuals shown below. To measure the convergence of θ_k and $\theta_k^{\mathcal{L}}$, we consider their distance to the optimal policy in the function space. The value function estimates $\hat{V}_k, \hat{V}_k^{\mathcal{L}}$ are measured by their ℓ_2 distance to the value functions under the latest upper-level decision variable and policy.

$$\begin{aligned}
\varepsilon_k^{\theta,\mathcal{L}} &= w_k\Big(\mathcal{L}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}}) - \mathcal{L}_{w_k,\tau_k}(x_k,\pi_{w_k,\tau_k}^\star(x_k))\Big),\\
\varepsilon_k^{\theta} &= J_{\tau_k}(x_k,\pi_{\tau_k}^\star(x_k)) - J_{\tau_k}(x_k,\pi_{\theta_k}),\\
\varepsilon_k^{V} &= \|\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}}\|^2, \quad \varepsilon_k^{V,\mathcal{L}} = \|\hat{V}_k^{\mathcal{L}} - V_{\tau_k}^{x_k,\pi_{\theta_k^{\mathcal{L}}}}\|^2.
\end{aligned} \tag{33}$$

We also introduce a number of technical lemmas, which will be used in the proofs of the propositions and theorems. We defer the proofs of the lemmas to Appendix E.

Lemma 2 *For any $k \geq 0$, we have*

$$\tau_k - \tau_{k+1} \leq \frac{8\tau_k}{3(k+1)}.$$

Lemma 2 derives a simple bound on the rate of change of the regularization weight τ_k.

Lemma 3 *Define $L_V = \max\{\frac{2L_r|\mathcal{S}||\mathcal{A}|}{1-\gamma}, \frac{(12+8\log|\mathcal{A}|)\sqrt{|\mathcal{S}|}}{(1-\gamma)^3}\}$. We have for all $w,\tau \leq 1$ and x,x',θ,θ'*

$$|J_\tau(x,\pi_\theta) - J_\tau(x',\pi_{\theta'})| \leq L_V(\|x-x'\| + \|\theta-\theta'\|), \tag{34}$$

$$\|\nabla_\theta J_\tau(x,\pi_\theta) - \nabla_\theta J_\tau(x',\pi_{\theta'})\| \leq L_V(\|x-x'\| + \|\theta-\theta'\|), \tag{35}$$

$$\|\nabla_x J_\tau(x,\pi) - \nabla_x J_\tau(x',\pi')\| \leq L_V(\|x-x'\| + \|\pi-\pi'\|), \tag{36}$$

$$\|V_\tau^{x,\pi_\theta} - V_\tau^{x',\pi_{\theta'}}\| \leq L_V(\|x-x'\| + \|\theta-\theta'\|), \tag{37}$$

$$\|\nabla_\theta V_\tau^{x,\pi_\theta} - \nabla_\theta V_\tau^{x',\pi_{\theta'}}\| \leq L_V(\|x-x'\| + \|\theta-\theta'\|), \tag{38}$$

$$\|\nabla_{\theta,\theta}\mathbb{E}_{s\sim d_\rho^{\pi_\theta}}[E(\pi_\theta,s)]\| \leq L_V. \tag{39}$$

In addition, there exists a bounded constant $L_{V,2}$[3] such that for all $\tau \leq 1$ and x,x',θ,θ'

$$\|\nabla_{x,\theta}^2 J_\tau(x,\pi_\theta) - \nabla_{x,\theta}^2 J_\tau(x',\pi_{\theta'})\| \leq L_{V,2}(\|x-x'\| + \|\theta-\theta'\|),$$

$$\|\nabla_{\theta,\theta}^2 J_\tau(x,\pi_\theta) - \nabla_{\theta,\theta}^2 J_\tau(x',\pi_{\theta'})\| \leq L_{V,2}(\|x-x'\| + \|\theta-\theta'\|).$$

[3] We skip showing the exact constant here, but note that it depends polynomially on the structural parameters of the problem.

Lemma 3 shows that the value functions/cumulative returns are Lipschitz, and have Lipschitz continuous gradients and Hessians.

Lemma 4 *For any π, π', we have*

$$\|d_\rho^\pi - d_\rho^{\pi'}\| \leq \frac{\gamma}{1-\gamma}\|\pi - \pi'\|.$$

Lemma 4 shows that the occupancy measure is a Lipschitz function of the policy, a well-known result in the literature. We include the proof in Section E.4 for completeness.

Lemma 5 *Recall that $B_V = \frac{1+\log|\mathcal{A}|}{1-\gamma}$ is the entry-wise upper bound on the magnitude of the value function introduced in the paragraph after* (3)*. We define the constants*

$$B_D = 3L_r, \quad B_F = 2(1+\gamma)B_V + 2|\log \underline{\pi}| + 2 + L_r, \quad B_G = (1+\gamma)B_V + \log|\log \underline{\pi}| + 1. \tag{40}$$

Suppose that the regularization parameters satisfy $w \leq \frac{L_r}{L_f}, \tau \leq 1$. For all $x, \theta, \theta^{\mathcal{L}}, s, a, s', \bar{s}, \bar{a}, \xi$ and $V \in \mathbb{R}^{|\mathcal{S}|}$ satisfying $V(s) \leq B_V$, we have

$$\|D_w(x, \pi_\theta, \pi_{\theta^{\mathcal{L}}}, s, a, \bar{s}, \bar{a}, \xi)\| \leq \frac{B_D}{w},$$
$$\|F_{w,\tau}(x, \theta, V, s, a, s', \xi)\| \leq B_F,$$
$$\|G_\tau(x, \theta, V, s, a, s')\| \leq B_G.$$

Lemma 6 *We define the constants*

$$L_D = 3L_r + \frac{2B_D}{1-\gamma}, \quad L_F = 3L_r + \frac{2}{\underline{\pi}} + |\log \underline{\pi}| + \frac{2+B_F}{1-\gamma} + 4, \quad L_G = L_r + \frac{B_G}{1-\gamma} + \frac{1}{\underline{\pi}} + 2. \tag{41}$$

Suppose that the regularization parameters satisfy $w \leq \min\{\frac{L_r}{L_f}, \frac{B_D}{(1-\gamma)L_f}\}, \tau \leq 1$. We have for all $x_1, x_2, \pi_1, \pi_1^{\mathcal{L}}, \pi_2, \pi_2^{\mathcal{L}}$

$$\|\bar{D}_w(x_1, \pi_1, \pi_1^{\mathcal{L}}) - \bar{D}_w(x_2, \pi_2, \pi_2^{\mathcal{L}})\| \leq \frac{L_D}{w}\Big(\|x_1 - x_2\| + \|\pi_1 - \pi_2\| + \|\pi_1^{\mathcal{L}} - \pi_2^{\mathcal{L}}\|\Big)$$
$$\|\bar{F}_{w,\tau}(x_1, \theta_1, V_1) - \bar{F}_{w,\tau}(x_2, \theta_2, V_2)\| \leq L_F(\|x_1 - x_2\| + \|\theta_1 - \theta_2\| + \|V_1 - V_2\|),$$
$$\|\bar{G}_\tau(x_1, \theta_1, V_1) - \bar{G}_\tau(x_2, \theta_2, V_2)\| \leq L_G(\|x_1 - x_2\| + \|\theta_1 - \theta_2\| + \|V_1 - V_2\|).$$

Lemma 7 *Recall the definition of $\pi^\star_{w,\tau}$ in Section 3.1. For any $w_1, w_2, \tau_1, \tau_2, x_1, x_2$, we have*

$$\|\pi^\star_{w_1,\tau_1}(x_1) - \pi^\star_{w_2,\tau_2}(x_2)\| \leq (\frac{2L_f w_2}{C_L\tau_1} + \frac{2L_V}{C_L\tau_1})\|x_1 - x_2\| + \frac{2L_f|w_1 - w_2|}{C_L\tau_1} + \frac{6|\tau_1 - \tau_2||\mathcal{S}|\log|\mathcal{A}|}{(1-\gamma)C_L\tau_1}.$$

In addition, for any $w, \tau > 0$, we have

$$\|\pi^\star_\tau(x) - \pi^\star_{w,\tau}(x)\| \leq \frac{2L_f w}{C_L\tau}.$$

Lemmas 6 and 7 show that the update operators introduced in (23)-(28) are (approximately) bounded and Lipschitz.

Lemma 8 *Define $L_\star = \frac{4+8\log|\mathcal{A}|}{\underline{\sigma}(1-\gamma)^4}$. For any x and $\tau \geq 0$, we have*

$$\|\pi^\star_\tau(x) - \pi^\star(x)\| \leq L_\star\tau.$$

Lemma 8 bounds the distance between the regularized best response $\pi^\star_\tau(x)$ to $\pi^\star(x)$ defined in (5) by a linear function of τ.

Lemma 9 *Define $L_L = L_V + L_f + \frac{L_V(C_L+2L_V)}{C_L}$. We have for all $w, \tau \leq 1$ and x, x', θ, θ'*

$$\|\nabla_\theta\mathcal{L}_{w,\tau}(x, \pi_\theta) - \nabla_\theta\mathcal{L}_{w,\tau}(x', \pi_{\theta'})\| \leq \frac{L_L}{w}\|x - x'\| + \frac{L_L}{w}\|\theta - \theta'\|, \tag{42}$$

$$\|\nabla_x\mathcal{L}_{w,\tau}(x, \pi_\theta) - \nabla_x\mathcal{L}_{w,\tau}(x', \pi_{\theta'})\| \leq \frac{L_L}{w\tau}\|x - x'\| + \frac{L_L}{w}\|\theta - \theta'\|. \tag{43}$$

Lemma 9 establishes the Lipschitz continuity of the gradients of $\mathcal{L}_{w,\tau}$.

Lemma 10 *Recall the definition of ℓ_τ in* (32)*. We have for all x_1, x_2*

$$\|\nabla \ell_\tau(x_1) - \nabla \ell_\tau(x_2)\| \leq \Big(L_V + \frac{2L_V^2}{C_L\tau}\Big)\|x_1 - x_2\|,$$
$$\|\nabla_x \Phi_{w,\tau}(x_1) - \nabla_x \Phi_{w,\tau}(x_2)\| \leq \Big(L_f + \frac{4L_f L_V}{C_L\tau} + \frac{4L_V^2}{C_L w\tau}\Big)\|x_1 - x_2\|,$$
$$\|\nabla_x \Phi_\tau(x_1) - \nabla_x \Phi_\tau(x_2)\| \leq (1 + \frac{2L_V}{C_L\tau})\Big(\frac{2L_f L_V}{C_L\tau} + \frac{2L_f L_V L_{V,2}}{\underline{\sigma} C_L\tau} + \frac{2L_f L_V^2 L_{V,2}}{\underline{\sigma}^2 C_L\tau} + \frac{2L_f L_V^2}{\underline{\sigma} C_L\tau}\Big)\|x_1 - x_2\|.$$

Particularly, if $w, \tau \leq 1$, we have

$$\|\nabla \ell_\tau(x_1) - \nabla \ell_\tau(x_2)\| \leq \frac{L_\Phi}{\tau}\|x_1 - x_2\|,$$
$$\|\nabla_x \Phi_{w,\tau}(x_1) - \nabla_x \Phi_{w,\tau}(x_2)\| \leq \frac{L_\Phi}{w\tau}\|x_1 - x_2\|,$$

where $L_\Phi = \max\{L_V + \frac{2L_V^2}{C_L}, L_f + \frac{4L_f L_V}{C_L} + \frac{4L_V^2}{C_L}, \frac{4L_f L_V}{C_L\tau} + \frac{4L_f L_V L_{V,2}}{\underline{\sigma} C_L\tau} + \frac{4L_f L_V^2 L_{V,2}}{\underline{\sigma}^2 C_L\tau} + \frac{4L_f L_V^2}{\underline{\sigma} C_L\tau}\}$.
In addition, if $\tau \leq \frac{2L_V}{C_L}$, we have

$$\|\nabla_x \Phi_\tau(x_1) - \nabla_x \Phi_\tau(x_2)\| \leq \frac{L_\Phi}{\tau}\|x_1 - x_2\|.$$

Lemma 10 establishes the Lipschitz continuity of the gradients of ℓ_τ, $\Phi_{w,\tau}$, and Φ_τ.

Lemma 11 *We have for any $w, \tau > 0$*

$$\|\nabla_x \Phi_\tau(x) - \nabla_x \Phi_{w,\tau}(x)\| \leq \frac{4L_f L_V w}{C_L\underline{\sigma}\tau}(L_f + \frac{2L_f L_{V,2}}{C_L\tau}),$$
$$\|\nabla_x \Phi(x) - \nabla_x \Phi_{w,\tau}(x)\| \leq \frac{4L_f L_V w}{C_L\underline{\sigma}\tau}(L_f + \frac{2L_f L_{V,2}}{C_L\tau}) + \frac{L_\star L_f(L_V + 1) + L_\star L_f L_{V,2} + L_\star L_V L_{V,2} + L_f L_{V,2}(4 + 8\log|\mathcal{A}|)}{(1-\gamma)^4\underline{\sigma}^2}\tau.$$

The lemma demonstrate how the magnitude of the difference between $\nabla_x \Phi_\tau(x)$ and $\nabla_x \Phi_{w,\tau}(x)$ and that between $\nabla_x \Phi(x)$ and $\nabla_x \Phi_{w,\tau}(x)$ scale with w and τ.

Lemma 12 *For any $\tau > 0$, we have*

$$\|\nabla_x \Phi_\tau(x) - \nabla_x \Phi(x)\| \leq \frac{L_\star L_f L_V^2 L_{V,2}\tau}{\underline{\sigma}^2}.$$

This lemma shows that distance between $\nabla_x \Phi_\tau(x)$ and $\nabla_x \Phi(x)$ is bounded by a function linear in τ.

B Applications of Bi-Level Reinforcement Learning

Bi-level RL provides a unifying framework for problems in which a high-level objective depends on the optimal solution of a lower-level policy optimization problem. In this section, we describe several representative applications that naturally fit into the bi-level RL formulation studied in this paper, besides RLHF discussed in Section 5.

B.1 Reward Shaping

In many reinforcement learning environments, the reward signal is sparse or delayed, making policy optimization difficult in practice. A common approach to address this challenge is reward shaping, where an auxiliary reward is added to the original environment reward to provide denser learning signals.

Let $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denote the original reward function (which we have no control over), and $\hat{r}_x$ an auxiliary reward parameterized by a variable x. The lower-level agent optimizes the shaped reward $r + \hat{r}_x$, while the upper-level optimization adjusts the shaping reward so that the induced optimal policy performs well with respect to the true task objective. This leads to the following formulation

$$\max_x J_r\left(\pi_{r+\hat{r}_x}\right) \quad \text{s.t.} \quad \pi_{r+\hat{r}_x} \in \operatorname*{argmax}_\pi J_{r+\hat{r}_x}(\pi), \tag{44}$$

where $J_r(\pi) = \mathbb{E}_\pi\left[\sum_{k=0}^\infty \gamma^k r\left(s_k, a_k\right)\right]$ for any policy π and reward r. For (44) to be well-defined, we need either $\pi^\star(x)$ to be unique or all lower-level optimal policies to evaluate to the same upper-level objective. When the problem structure does not guarantee that this is always true, we can slightly revise the formulation to include a selection mapping on the set of lower-level optimal policies as in (1) and (6).

B.2 Inverse Reinforcement Learning

Inverse reinforcement learning (IRL) aims to recover a reward function that explains observed expert behavior. Given a collection of expert demonstrations $\mathcal{D}$, assumed to be generated by an optimal policy under some unknown reward, the goal is to infer a reward such that the induced optimal policy matches the demonstrations.

Let $L(\pi, \mathcal{D})$ be a loss measuring the discrepancy between a policy π and the expert data, for example via maximum likelihood or a divergence between trajectory distributions. IRL can be expressed as the bi-level problem

$$\min_r L\left(\pi_r, \mathcal{D}\right) \quad \text{s.t.} \quad \pi_r \in \operatorname*{argmax}_\pi J_r(\pi).$$

Here, the reward function plays the role of the upper-level decision variable, while the lower-level problem corresponds to policy optimization under that reward. Again, we may need to include a selection mapping on the set of lower-level optimal policies as in (1) and (6), to ensure that the problem is well-defined.

C Proof of Theorems

We study a slightly simplified variant of Algorithm 1, which is presented as Algorithm 2. The sole distinction between the two is that Algorithm 2 uses i.i.d. samples drawn from the stationary distribution, instead of continuously generated Markovian samples. Stochastic approximation and RL algorithms have been extensively analyzed under Markovian sampling [Zou et al., 2019, Wu et al., 2020], and it is well-established that Markovian samples affect convergence rates only by a logarithmic factor. This simplification enables us to concentrate on the novel aspects we introduce to bi-level RL, without being distracted by standard technical considerations related to Markovian samples.

Theorem 3 (replication of Theorem 1 under i.i.d. samples) *Consider the iterates of Algorithm 2 under the step sizes and weights*

$$\zeta_k = \frac{\zeta_0}{(k+1)^{c_\zeta}}, \quad \alpha_k = \frac{\alpha_0}{(k+1)^{c_\alpha}}, \quad \beta_k = \frac{\beta_0}{(k+1)^{c_\beta}}, \quad w_k = \frac{w_0}{(k+1)^{c_w}}, \quad \tau_k = \frac{\tau_0}{(k+1)^{c_\tau}},$$

Algorithm 2 Actor Critic Algorithm for Bi-Level RL

1: **Initialize:** control variable x_0, policy parameters θ_0 and $\theta_0^{\mathcal{L}}$, value function estimates $\hat{V}_0, \hat{V}_0^{\mathcal{L}} \in \mathbb{R}^{|\mathcal{S}|}$
2: **for** iteration $k = 0, 1, 2, ...$ **do**
3: Trajectory 1:
Get samples $s_k \sim d_\rho^{\pi_{\theta_k}}, a_k \sim \pi_{\theta_k}(\cdot \mid s_k), s_k' \sim \mathcal{P}(\cdot \mid s_k, a_k)$. Receive reward $r_{x_k}(s_k, a_k)$.
4: Trajectory 2:
Get samples $\bar{s}_k \sim d_\rho^{\pi_{\theta_k^{\mathcal{L}}}}, \bar{a}_k \sim \pi_{\theta_k^{\mathcal{L}}}(\cdot \mid s_k), \bar{s}_k' \sim \mathcal{P}(\cdot \mid \bar{s}_k, a_k)$. Receive reward $r_{x_k}(\bar{s}_k, \bar{a}_k)$.
5: Observe/Obtain $\xi_k \sim \mu$
6: Control variable update:

$$x_{k+1} = x_k - \zeta_k \left(\widetilde{\nabla}_x f(x_k, \pi_{\theta_k}, \xi_k) + \frac{1}{w_k} \Big(\nabla_x r_{x_k}(s_k, a_k) - \nabla_x r_{x_k}(\bar{s}_k, \bar{a}_k) \Big) \right). \tag{45}$$

7: Policy update:

$$\theta_{k+1} = \theta_k + \alpha_k \Big(r_{x_k}(s_k, a_k) + \tau_k E(\pi_{\theta_k}, s_k) + \gamma \hat{V}_k(s_{k+1}) \Big) \nabla_\theta \log \pi_{\theta_k}(a_k \mid s_k), \tag{46}$$

$$\theta_{k+1}^{\mathcal{L}} = \theta_k^{\mathcal{L}} + \alpha_k \Big(\big(r_{x_k}(\bar{s}_k, \bar{a}_k) + \tau_k E(\pi_{\theta_k^{\mathcal{L}}}, \bar{s}_k) + \gamma \hat{V}_k^{\mathcal{L}}(\bar{s}_{k+1}) \big) \nabla_\theta \log \pi_{\theta_k^{\mathcal{L}}}(\bar{a}_k \mid \bar{s}_k) - w_k \widetilde{\nabla}_\theta f(x_k, \pi_{\theta_k^{\mathcal{L}}}, \xi_k) \Big), \tag{47}$$

$$\pi_k = \mathrm{softmax}(\theta_k), \quad \pi_k^{\mathcal{L}} = \mathrm{softmax}(\theta_k^{\mathcal{L}}). \tag{48}$$

8: Value function update:

$$\begin{aligned} \hat{V}_{k+1} &= \Pi_{B_V} \left(\hat{V}_k + \beta_k e_{s_k} \Big(r_{x_k}(s_k, a_k) + \tau_k E(\pi_{\theta_k}, s_k) + \gamma \hat{V}_k(s_{k+1}) - \hat{V}_k(s_k) \Big) \right), \\ \hat{V}_{k+1}^{\mathcal{L}} &= \Pi_{B_V} \left(\hat{V}_k^{\mathcal{L}} + \beta_k e_{\bar{s}_k} \Big(r_{x_k}(\bar{s}_k, \bar{a}_k) + \tau_k E(\pi_{\theta_k}^{\mathcal{L}}, \bar{s}_k) + \gamma \hat{V}_k^{\mathcal{L}}(\bar{s}_{k+1}) - \hat{V}_k^{\mathcal{L}}(\bar{s}_k) \Big) \right). \end{aligned} \tag{49}$$

9: **end for**

with $c_\zeta = \frac{9}{10}, c_\alpha = \frac{1}{2}, c_\beta = \frac{1}{2}, c_w = \frac{3}{20}, c_\tau = \frac{1}{20}$ *and* $\zeta_0, \alpha_0, \beta_0, w_0, \tau_0$ *selected such that*[4]

$$\begin{gathered} \zeta_0 \le \alpha_0 \le \beta_0 \le w_0 \le \tau_0 \le 1, \\ \alpha_0 \le \min\{\frac{4L_\Phi}{3L_V^2}, \frac{4L_\Phi}{3L_L^2}, \frac{3B_F}{8}\}, \quad \beta_0 \le \min\{\frac{1-\gamma}{L_G^2}, \frac{(1-\gamma)L_V^2}{8|\mathcal{S}|\log^2|\mathcal{A}|}\}, \\ w_0 \le \min\{\bar{w}, \frac{L_r}{L_f}, \frac{B_D}{(1-\gamma)L_f}\}, \quad \tau_0 \le \min\{\frac{L_f|\mathcal{S}|}{C_L}, \frac{2L_V}{C_L}, \frac{2L_{V,2}}{C_L}\}, \\ \frac{\zeta_0}{\alpha_0} \le \min\left\{ \frac{C_L^2\tau_0^2}{1024(L_L^2+L_V^2)}, \frac{C_L^2 w_0^2 \tau_0^2}{512 L_D^2}, \frac{C_L^2 w_0 \tau_0^2}{128 L_D L_L}, \frac{(1-\gamma)C_L^2\tau_0^2}{6144 L_V^2 L_D^2}, \frac{B_F}{B_D} \right\}, \\ \frac{\alpha_0}{\beta_0} \le \left\{ \frac{1-\gamma}{2\sqrt{6}L_V L_F}, \frac{1-\gamma}{48L_V^2}, \frac{1-\gamma}{8L_F^2}, \frac{8(1-\gamma)(L_L^2+L_V^2)}{3L_V^2 C_L^2}, \sqrt{\frac{32B_G}{B_F^2(L_L+L_V)}}, \frac{1-\gamma}{36B_F^2 L_V \tau_0} \right\}, \\ \frac{\alpha_0}{\beta_0^2} \le \frac{4B_G}{11B_F^2 L_V}, \quad \frac{\alpha_0}{\tau_0} \le \frac{1}{L_V}. \end{gathered} \tag{50}$$

Then, under Assumptions 1-5, we have for all $k \ge 0$,

$$\min_{t<k} \mathbb{E}[\|\nabla_x \Phi(x_t)\|^2] \le \frac{40}{3\zeta_0(k+1)^{1/10}} \left(\Phi_{\tau_0}(x_0) + \varepsilon_0^\theta + \varepsilon_0^{\theta,\mathcal{L}} + \varepsilon_0^V + \varepsilon_0^{V,\mathcal{L}} \right) + \mathcal{O}\left(\frac{\log(k+1)}{(k+1)^{1/10}} \right).$$

[4]Note that the step sizes satisfying the conditions always exist and can be found in the order of $\tau_0, w_0, \beta_0, \alpha_0, \zeta_0$ – we first select w_0, τ_0 small enough to satisfy their upper bounds; then we select β_0; then we select α_0 with respect to β_0, w_0, τ_0; and finally we select ζ_0 with respect to $\alpha_0, \beta_0, w_0, \tau_0$.

Theorem 4 (replication of Theorem 2 under i.i.d. samples) *Given any fixed regularization weight* $\tau_0 \leq \min\{1, \frac{L_f|\mathcal{S}|}{C_L}, \frac{2L_V}{C_L}, \frac{2L_{V,2}}{C_L}\}$[5]*, i.e.,* $\tau_k = \tau_0$ *for all* $k \geq 0$*, consider the iterates of Algorithm 2 under the step sizes and penalty weight*

$$\zeta_k = \frac{\zeta_0}{(k+1)^{c_\zeta}}, \quad \alpha_k = \frac{\alpha_0}{(k+1)^{c_\alpha}}, \quad \beta_k = \frac{\beta_0}{(k+1)^{c_\beta}}, \quad w_k = \frac{w_0}{(k+1)^{c_w}}$$

with $c_\zeta = \frac{2}{3}, c_\alpha = \frac{1}{2}, c_\beta = \frac{1}{2}, c_w = \frac{1}{6}$ *and* $\zeta_0, \alpha_0, \beta_0, w_0$ *selected such that* (50) *holds. Then, under Assumptions 1-5, we have for all* $k \geq 0$*,*

$$\min_{t<k} \mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_t)\|^2] \leq \frac{20}{3\zeta_0(k+1)^{1/3}} \left(\Phi_{\tau_0}(x_0) + \varepsilon_0^\theta + \varepsilon_0^{\theta,\mathcal{L}} + \varepsilon_0^V + \varepsilon_0^{V,\mathcal{L}}\right) + \mathcal{O}\left(\frac{\log(k+1)}{(k+1)^{1/3}}\right).$$

We break down the proofs of the theorems into two parts. First, in the following propositions, we individually establish the iteration-wise convergence of the upper-level decision variable (in Proposition 1 under decaying regularization and Proposition 2 under fixed regularization), policy iterates (in Propositions 3-4), and value function estimates (in Proposition 5). Then, in Sections C.1-C.2, we combine the convergence of these variables and bound their joint convergence through a coupled Lyapunov function. The proofs of the propositions are deferred to Section D.

Proposition 1 *Under Assumptions 1-5 and step sizes satisfying* (50)*, the iterates of Algorithm 2 satisfy for all* $k \geq 0$

$$\begin{aligned}\frac{\zeta_k}{4}\mathbb{E}[\|\nabla_x \Phi(x_k)\|^2] \leq{}& \mathbb{E}[\Phi_{\tau_k}(x_k) - \Phi_{\tau_{k+1}}(x_{k+1})] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] \\ &+ \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2] + \frac{L_\star^2 L_f^2 L_V^4 L_{V,2}^2 \zeta_k \tau_k^2}{2\underline{\sigma}^4} \\ &+ \frac{256 L_f^4 L_V^2 L_{V,2}^2 \zeta_k w_k^2}{C_L^4 \underline{\sigma}^2 \tau_k^4} + \frac{B_D^2 L_\Phi \zeta_k^2}{2\tau_k w_k^2} + \frac{16 L_f |\mathcal{S}| \log|\mathcal{A}|}{(1-\gamma)C_L(k+1)}.\end{aligned}$$

Proposition 2 *Under Assumptions 1-5 and step sizes satisfying* (50)*, the iterates of Algorithm 2 satisfy for all* $k \geq 0$

$$\begin{aligned}\frac{\zeta_k}{2}\mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_k)\|^2] \leq{}& \mathbb{E}[\Phi_{\tau_0}(x_k) - \Phi_{\tau_0}(x_{k+1})] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_0}(x_k)\|^2] \\ &+ \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_0}(x_k)\|^2] + C_{2,\tau_0}\zeta_k w_k^2 + \frac{B_D^2 L_{\Phi,\tau_0}\zeta_k^2}{2w_k^2},\end{aligned}$$

where $C_{2,\tau_0} = \left(\frac{4L_f L_V}{C_L \underline{\sigma}\tau_0}(L_f + \frac{2L_f L_{V,2}}{C_L \tau_0})\right)^2$ *and* $L_{\Phi,\tau_0} = (1 + \frac{2L_V}{C_L\tau_0})\left(\frac{2L_f L_V}{C_L\tau_0} + \frac{2L_f L_V L_{V,2}}{\underline{\sigma} C_L \tau_0} + \frac{2L_f L_V^2 L_{V,2}}{\underline{\sigma}^2 C_L \tau_0} + \frac{2L_f L_V^2}{\underline{\sigma} C_L \tau}\right)$.

Proposition 3 *Under Assumptions 1-5 and step sizes satisfying* (50)*, the iterates of Algorithm 2 satisfy for all* $k \geq 0$

$$\begin{aligned}\mathbb{E}[\varepsilon_{k+1}^\theta - \varepsilon_k^\theta] \leq{}& -\frac{\alpha_k}{8}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2] + 2L_F^2\alpha_k\mathbb{E}[\varepsilon_k^V] + \frac{32L_V^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2] \\ &- \frac{C_L^2\alpha_k\tau_k^2}{64}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] + \frac{64L_D^2L_V^2\zeta_k^2}{C_L^2\alpha_k w_k^2\tau_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2] \\ &+ \frac{2B_D^2 L_\Phi \zeta_k^2}{w_k^2\tau_k} + \frac{2B_D B_F L_V \zeta_k \alpha_k}{w_k} + \frac{B_F^2 L_V \alpha_k^2}{2} + \frac{16\log|\mathcal{A}|\tau_k}{3(1-\gamma)(k+1)}.\end{aligned}$$

Proposition 4 *Under Assumptions 1-5 and step sizes satisfying* (50)*, the iterates of Algorithm 2 satisfy for all* $k \geq 0$

$$\begin{aligned}&\mathbb{E}[\varepsilon_{k+1}^{\theta,\mathcal{L}} - \varepsilon_k^{\theta,\mathcal{L}}] \\ &\leq -\frac{w_k^2\alpha_k}{8}\mathbb{E}[\|\nabla_\theta \mathcal{L}_{w_k,\tau_k}(x_k, \pi_{\theta_k^{\mathcal{L}}})\|^2] + 2L_F^2\alpha_k\mathbb{E}[\varepsilon_k^{V,\mathcal{L}}] + \frac{32L_L^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2]\end{aligned}$$

[5]Note that Propositions 3-5 use the Lipschitz continuity conditions established on operator/functions such as $L_{w,\tau}$ and ℓ_τ under $\tau \leq \min\{1, \frac{L_f|\mathcal{S}|}{C_L}, \frac{2L_V}{C_L}, \frac{2L_{V,2}}{C_L}\}$, a condition imposed so that we can present the associated Lipschitz constants in a more concise form. As the proof of Theorem 4 is based on Propositions 3-5, we state that the result holds for this range of τ_0. However, the same proof technique applies verbatim for any arbitrary $\tau_0 > 0$; only the values of the Lipschitz constants would change accordingly.

$$- \frac{C_L^2\alpha_k\tau_k^2}{64}\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2] + \frac{64L_D^2L_L^2\zeta_k^2}{C_L^2\alpha_k w_k^2\tau_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{2B_D^2L_\Phi\zeta_k^2}{w_k^2\tau_k} + \frac{2B_DB_FL_L\zeta_k\alpha_k}{w_k} + \frac{B_F^2L_L\alpha_k^2}{2} + \frac{32\log|\mathcal{A}|\tau_k}{3(1-\gamma)(k+1)}.$$

Proposition 5 (Value Function Convergence) *Under Assumptions 1-5 and step sizes satisfying* (50)*, the iterates of Algorithm 2 satisfy for all $k \geq 0$*

$$\mathbb{E}[\varepsilon_{k+1}^V] \leq \Big(1 - \frac{(1-\gamma)\beta_k}{4}\Big)\mathbb{E}[\varepsilon_k^V] + \frac{12L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{6L_V^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2] + \frac{6L_V^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{22B_F^2L_V\tau_0\alpha_k^2}{\alpha_0} + \frac{64L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2} + 8B_G\beta_k^2,$$
$$\mathbb{E}[\varepsilon_{k+1}^{V,\mathcal{L}}] \leq \Big(1 - \frac{(1-\gamma)\beta_k}{4}\Big)\mathbb{E}[\varepsilon_k^{V,\mathcal{L}}] + \frac{12L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{6L_V^2w_k^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_\theta\mathcal{L}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}})\|^2] + \frac{6L_V^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{22B_F^2L_V\tau_0\alpha_k^2}{\alpha_0} + \frac{64L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2} + 8B_G\beta_k^2.$$

C.1 Proof of Theorem 3 (replication of Theorem 1 under i.i.d. samples)

Combining the bounds in Propositions 1-5, we have for any $k \geq 0$

$$\frac{\zeta_k}{4}\mathbb{E}[\|\nabla_x\Phi(x_k)\|^2]$$
$$\leq \mathbb{E}[\Phi_{\tau_k}(x_k) - \Phi_{\tau_{k+1}}(x_{k+1})] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{L_\star^2L_f^2L_V^4L_{V,2}^2\zeta_k\tau_k^2}{2\underline{\sigma}^4} + \frac{256L_f^4L_V^2L_{V,2}^2\zeta_kw_k^2}{C_L^4\underline{\sigma}^2\tau_k^4} + \frac{B_D^2L_\Phi\zeta_k^2}{2\tau_kw_k^2} + \frac{16L_f|\mathcal{S}|\log|\mathcal{A}|}{(1-\gamma)C_L(k+1)}$$
$$+ \mathbb{E}[\varepsilon_k^{\theta,\mathcal{L}} - \varepsilon_{k+1}^{\theta,\mathcal{L}}] - \frac{w_k^2\alpha_k}{8}\mathbb{E}[\|\nabla_\theta\mathcal{L}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}})\|^2] + 2L_F^2\alpha_k\mathbb{E}[\varepsilon_k^{V,\mathcal{L}}] + \frac{32L_L^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]$$
$$- \frac{C_L^2\alpha_k\tau_k^2}{64}\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2] + \frac{64L_D^2L_L^2\zeta_k^2}{C_L^2\alpha_kw_k^2\tau_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{2B_D^2L_\Phi\zeta_k^2}{w_k^2\tau_k} + \frac{2B_DB_FL_L\zeta_k\alpha_k}{w_k} + \frac{B_F^2L_L\alpha_k^2}{2} + \frac{32\log|\mathcal{A}|\tau_k}{3(1-\gamma)(k+1)}$$
$$+ \mathbb{E}[\varepsilon_k^{\theta} - \varepsilon_{k+1}^{\theta}] - \frac{\alpha_k}{8}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2] + 2L_F^2\alpha_k\mathbb{E}[\varepsilon_k^V] + \frac{32L_V^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]$$
$$- \frac{C_L^2\alpha_k\tau_k^2}{64}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] + \frac{64L_D^2L_V^2\zeta_k^2}{C_L^2\alpha_kw_k^2\tau_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{2B_D^2L_\Phi\zeta_k^2}{w_k^2\tau_k} + \frac{2B_DB_FL_V\zeta_k\alpha_k}{w_k} + \frac{B_F^2L_V\alpha_k^2}{2} + \frac{16\log|\mathcal{A}|\tau_k}{3(1-\gamma)(k+1)}$$
$$- \mathbb{E}[\varepsilon_{k+1}^V] + \Big(1 - \frac{(1-\gamma)\beta_k}{4}\Big)\mathbb{E}[\varepsilon_k^V] + \frac{12L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{6L_V^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2] + \frac{6L_V^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{22B_F^2L_V\tau_0\alpha_k^2}{\alpha_0} + \frac{64L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2} + 8B_G\beta_k^2$$
$$- \mathbb{E}[\varepsilon_{k+1}^{V,\mathcal{L}}] + \Big(1 - \frac{(1-\gamma)\beta_k}{4}\Big)\mathbb{E}[\varepsilon_k^{V,\mathcal{L}}] + \frac{12L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]$$

$$
\begin{aligned}
&+ \frac{6L_V^2 w_k^2 \alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_\theta \mathcal{L}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}})\|^2] + \frac{6L_V^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2] \\
&+ \frac{22B_F^2 L_V \tau_0 \alpha_k^2}{\alpha_0} + \frac{64L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2} + 8B_G\beta_k^2 \\
\le\ & \mathbb{E}[\Phi_{\tau_k}(x_k) - \Phi_{\tau_{k+1}}(x_{k+1}) + \varepsilon_k^{\theta,\mathcal{L}} - \varepsilon_{k+1}^{\theta,\mathcal{L}} + \varepsilon_k^{\theta} - \varepsilon_{k+1}^{\theta} + \varepsilon_k^{V} - \varepsilon_{k+1}^{V} + \varepsilon_k^{V,\mathcal{L}} - \varepsilon_{k+1}^{V,\mathcal{L}}] \\
&+ \Big(\frac{32(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_k^2} + \frac{12L_V^2\zeta_k^2}{(1-\gamma)\beta_k}\Big)\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2] + \Big(-\frac{\alpha_k}{8} + \frac{6L_V^2\alpha_k^2}{(1-\gamma)\beta_k}\Big)\mathbb{E}[\|\|\nabla_\theta\mathcal{L}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}})\|^2] \\
&+ \Big(-\frac{w_k^2\alpha_k}{8} + \frac{6L_V^2 w_k^2\alpha_k^2}{(1-\gamma)\beta_k}\Big)\mathbb{E}[\|\|\nabla_\theta\mathcal{L}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}})\|^2] \\
&+ \Big(-\frac{C_L^2\alpha_k\tau_k^2}{64} + \frac{2L_D^2\zeta_k}{w_k^2} + \frac{128L_D^2L_L^2\zeta_k^2}{C_L^2\alpha_k w_k^2\tau_k^2} + \frac{24L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\Big)\mathbb{E}[\|\pi_k - \pi_{\tau_k}^\star(x_k)\|^2] \\
&+ \Big(-\frac{C_L^2\alpha_k\tau_k^2}{64} + \frac{2L_D^2\zeta_k}{w_k^2} + \frac{128L_D^2L_L^2\zeta_k^2}{C_L^2\alpha_k w_k^2\tau_k^2} + \frac{24L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\Big)\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi_{w_k,\tau_k}^\star(x_k)\|^2] \\
&+ \Big(-\frac{(1-\gamma)\beta_k}{4} + 2L_F^2\alpha_k\Big)\mathbb{E}[\varepsilon_k^V] + \Big(-\frac{(1-\gamma)\beta_k}{4} + 2L_F^2\alpha_k\Big)\mathbb{E}[\varepsilon_k^{V,\mathcal{L}}] \\
&+ \frac{256L_f^4L_V^2L_{V,2}^2\zeta_k w_k^2}{C_L^4\underline{\sigma}^2\tau_k^4} + \frac{L_\star^2L_f^2L_V^4L_{V,2}^2\zeta_k\tau_k^2}{2\underline{\sigma}^4} + \frac{9B_D^2L_\Phi\zeta_k^2}{2w_k^2\tau_k} + \frac{4B_DB_FL_L\zeta_k\alpha_k}{w_k} + 48B_G\beta_k^2 \\
&+ \frac{128L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2} + \frac{16\log|\mathcal{A}|\tau_k}{(1-\gamma)(k+1)} + \frac{16L_f|\mathcal{S}|\log|\mathcal{A}|}{(1-\gamma)C_L(k+1)},
\end{aligned}
$$

where to get the second inequality we combine the terms $\frac{22B_F^2L_V\tau_0\alpha_k^2}{\alpha_0}$ and $8B_G\beta_k^2$ under the condition $\alpha_0 \le \frac{4B_G\beta_0^2}{11B_F^2L_V}$, and the terms $\frac{B_F^2L_L\alpha_k^2}{2} + \frac{B_F^2L_V\alpha_k^2}{2}$ and $16B_G\beta_k^2$ under the condition $\frac{\alpha_k}{\beta_k} \le \sqrt{\frac{32B_G}{B_F^2(L_L+L_V)}}$.

Note that the highlighted red coefficients in the inequality above are non-positive and the blue coefficients can be combined under the step size conditions $\zeta_0 \le \beta_0$, $\tau_k \le \frac{L_f|\mathcal{S}|}{C_L}$, and

$$
\frac{\alpha_0}{\beta_0} \le \left\{\frac{1-\gamma}{48L_V^2}, \frac{1-\gamma}{8L_F^2}, \frac{8(1-\gamma)(L_L^2+L_V^2)}{3L_V^2C_L^2}\right\}, \quad \frac{\zeta_0}{\alpha_0} \le \min\left\{\frac{C_L^2w_0^2\tau_0^2}{512L_D^2}, \frac{C_L^2w_0\tau_0^2}{128L_DL_L}, \frac{(1-\gamma)C_L^2\tau_0^2}{6144L_V^2L_D^2}\right\}.
$$

This allows us to simplify the inequality and obtain

$$
\begin{aligned}
&\frac{\zeta_k}{4}\mathbb{E}[\|\nabla_x\Phi(x_k)\|^2] \\
&\le \mathbb{E}[\Phi_{\tau_k}(x_k) - \Phi_{\tau_{k+1}}(x_{k+1}) + \varepsilon_k^{\theta,\mathcal{L}} - \varepsilon_{k+1}^{\theta,\mathcal{L}} + \varepsilon_k^{\theta} - \varepsilon_{k+1}^{\theta} + \varepsilon_k^{V} - \varepsilon_{k+1}^{V} + \varepsilon_k^{V,\mathcal{L}} - \varepsilon_{k+1}^{V,\mathcal{L}}] \\
&\quad + \frac{64(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2] + \frac{256L_f^4L_V^2L_{V,2}^2\zeta_k w_k^2}{C_L^4\underline{\sigma}^2\tau_k^4} + \frac{L_\star^2L_f^2L_V^4L_{V,2}^2\zeta_k\tau_k^2}{2\underline{\sigma}^4} + \frac{9B_D^2L_\Phi\zeta_k^2}{2w_k^2\tau_k} \\
&\quad + \frac{4B_DB_FL_L\zeta_k\alpha_k}{w_k} + 48B_G\beta_k^2 + \frac{128L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2} + \frac{32L_f|\mathcal{S}|\log|\mathcal{A}|}{(1-\gamma)C_L(k+1)}.
\end{aligned} \tag{51}
$$

We can relate the second term on the right hand side of (51) to $\|\nabla_x\Phi(x_k)\|^2$ using Lemma 11

$$
\begin{aligned}
&\frac{64(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2 \\
&\le \frac{128(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\|\nabla_x\Phi(x_k)\|^2 + \frac{128(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\|\nabla_x\Phi(x_k) - \nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2 \\
&\le \frac{128(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\|\nabla_x\Phi(x_k)\|^2 + \frac{128(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\Big(\frac{4L_fL_Vw_k}{C_L\underline{\sigma}\tau_k}(L_f + \frac{2L_fL_{V,2}}{C_L\tau_k}) \\
&\qquad + \frac{L_\star L_f(L_V+1) + L_\star L_fL_{V,2} + L_\star L_VL_{V,2} + L_fL_{V,2}(4+8\log|\mathcal{A}|)}{(1-\gamma)^4\underline{\sigma}^2}\tau_k\Big)^2 \\
&\le \frac{\zeta_k}{8}\|\nabla_x\Phi(x_k)\|^2 + \frac{128(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\Big(\frac{4L_fL_Vw_k}{C_L\underline{\sigma}\tau_k}(L_f + \frac{2L_fL_{V,2}}{C_L\tau_k})
\end{aligned}
$$

$$+ \frac{L_\star L_f(L_V+1) + L_\star L_f L_{V,2} + L_\star L_V L_{V,2} + L_f L_{V,2}(4+8\log|\mathcal{A}|)}{(1-\gamma)^4\underline{\sigma}^2}\tau_k\Big)^2$$
$$= \frac{\zeta_k}{8}\|\nabla_x\Phi(x_k)\|^2 + \frac{128(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\Big(\frac{4L_fL_Vw_k}{C_L\underline{\sigma}\tau_k}(L_f+\frac{2L_fL_{V,2}}{C_L\tau_k}) + C_1\tau_k\Big)^2, \tag{52}$$

where we derive the third inequality from the step size condition $\frac{\zeta_0}{\alpha_0} \leq \frac{C_L^2\tau_0^2}{1024(L_L^2+L_V^2)}$, and define in the last equation $C_1 = \frac{L_\star L_f(L_V+1)+L_\star L_f L_{V,2}+L_\star L_V L_{V,2}+L_f L_{V,2}(4+8\log|\mathcal{A}|)}{(1-\gamma)^4\underline{\sigma}^2}$.

Combining (51) and (52) and substituting in the step size decay rates,

$$\begin{aligned}
\frac{\zeta_k}{8}\mathbb{E}[\|\nabla_x\Phi(x_k)\|^2] &\leq \mathbb{E}[\Phi_{\tau_k}(x_k) - \Phi_{\tau_{k+1}}(x_{k+1}) + \varepsilon_k^{\theta,\mathcal{L}} - \varepsilon_{k+1}^{\theta,\mathcal{L}} + \varepsilon_k^{\theta} - \varepsilon_{k+1}^{\theta} + \varepsilon_k^{V} - \varepsilon_{k+1}^{V} + \varepsilon_k^{V,\mathcal{L}} - \varepsilon_{k+1}^{V,\mathcal{L}}] \\
&\quad + \frac{128(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\Big(\frac{4L_fL_Vw_k}{C_L\underline{\sigma}\tau_k}(L_f+\frac{2L_fL_{V,2}}{C_L\tau_k}) + C_1\tau_k\Big)^2 \\
&\quad + \frac{256L_f^4L_V^2L_{V,2}^2\zeta_kw_k^2}{C_L^4\underline{\sigma}^2\tau_k^4} + \frac{L_\star^2L_f^2L_V^4L_{V,2}^2\zeta_k\tau_k^2}{2\underline{\sigma}^4} + \frac{9B_D^2L_\Phi\zeta_k^2}{2w_k^2\tau_k} \\
&\quad + \frac{4B_DB_FL_L\zeta_k\alpha_k}{w_k} + 48B_G\beta_k^2 + \frac{128L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2} + \frac{32L_f|\mathcal{S}|\log|\mathcal{A}|}{(1-\gamma)C_L(k+1)} \\
&\leq \mathbb{E}[\Phi_{\tau_k}(x_k) - \Phi_{\tau_{k+1}}(x_{k+1}) + \varepsilon_k^{\theta,\mathcal{L}} - \varepsilon_{k+1}^{\theta,\mathcal{L}} + \varepsilon_k^{\theta} - \varepsilon_{k+1}^{\theta} + \varepsilon_k^{V} - \varepsilon_{k+1}^{V} + \varepsilon_k^{V,\mathcal{L}} - \varepsilon_{k+1}^{V,\mathcal{L}}] \\
&\quad + \mathcal{O}\left(\frac{1}{(k+1)}\right).
\end{aligned} \tag{53}$$

Re-arranging the terms and summing over iterations,

$$\sum_{t=0}^{k-1}\frac{\zeta_0}{8(t+1)^{9/10}}\mathbb{E}[\|\nabla_x\Phi(x_t)\|^2] \leq \Phi_{\tau_0}(x_0) + \varepsilon_0^{\theta} + \varepsilon_0^{\theta,\mathcal{L}} + \varepsilon_0^{V} + \varepsilon_0^{V,\mathcal{L}} + \mathcal{O}\left(\sum_{t=0}^{k-1}\frac{1}{(t+1)}\right). \tag{54}$$

The following inequalities on the summation of step sizes are standard results in the literature and easy to verify.

$$\sum_{t=0}^{k-1}\frac{1}{t+1} \leqslant \frac{\log(k+1)}{\log(2)},$$
$$\sum_{t=0}^{k}\frac{1}{(t+1)^u} \geq \frac{(1-\frac{1}{2^{1-u}})(k+1)^{1-u}}{1-u}, \quad \forall u \in (0,1)$$

and with $u = 9/10$,

$$\sum_{t=0}^{k}\frac{1}{(t+1)^{9/10}} \geq \frac{0.06(k+1)^{1/10}}{1/10} = \frac{3(k+1)^{1/10}}{5}.$$

This allows us to further simplify (54)

$$\begin{aligned}
\min_{t<k}\mathbb{E}[\|\nabla_x\Phi(x_t)\|^2] &\leq \frac{1}{\sum_{t=0}^{k-1}\frac{\zeta_0}{8(t+1)^{9/10}}}\sum_{t=0}^{k-1}\frac{\zeta_0}{8(t+1)^{9/10}}\mathbb{E}[\|\nabla_x\Phi(x_t)\|^2] \\
&\leq \frac{40}{3\zeta_0(k+1)^{1/10}}\left(\Phi_{\tau_0}(x_0) + \varepsilon_0^{\theta} + \varepsilon_0^{\theta,\mathcal{L}} + \varepsilon_0^{V} + \varepsilon_0^{V,\mathcal{L}} + \mathcal{O}(\log(k+1))\right).
\end{aligned}$$

■

C.2 Proof of Theorem 4 (replication of Theorem 2 under i.i.d. samples)

We combine Propositions 2 and 3. Note that $\tau_k = \tau_0$. We can follow a line of analysis identical to what leads to (51) in the proof of Theorem 3 and show the following inequality

$$\frac{\zeta_k}{2}\mathbb{E}[\|\nabla_x\Phi_{\tau_0}(x_k)\|^2] \leq \mathbb{E}[\Phi_{\tau_0}(x_k) - \Phi_{\tau_0}(x_{k+1}) + \varepsilon_k^{\theta,\mathcal{L}} - \varepsilon_{k+1}^{\theta,\mathcal{L}} + \varepsilon_k^{\theta} - \varepsilon_{k+1}^{\theta} + \varepsilon_k^{V} - \varepsilon_{k+1}^{V} + \varepsilon_k^{V,\mathcal{L}} - \varepsilon_{k+1}^{V,\mathcal{L}}]$$

$$+ \frac{64(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_0^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2] + C_{2,\tau_0}\zeta_k w_k^2 + \frac{4B_D^2 L_\Phi \zeta_k^2}{w_k^2\tau_0} + \frac{B_D^2 L_{\Phi,\tau_0}\zeta_k^2}{2w_k^2}$$
$$+ \frac{4B_D B_F L_L \zeta_k \alpha_k}{w_k} + 48B_G\beta_k^2 + \frac{128L_V^2\tau_0^2}{3(1-\gamma)\beta_k(k+1)^2} + \frac{48\log|\mathcal{A}|\tau_0}{3(1-\gamma)(k+1)}. \tag{55}$$

The step size conditions required to show (55) are

$$\frac{\alpha_0}{\beta_0} \le \left\{\frac{1-\gamma}{48L_V^2}, \frac{1-\gamma}{8L_F^2}, \frac{8(1-\gamma)(L_L^2+L_V^2)}{3L_V^2C_L^2}, \sqrt{\frac{32B_G}{B_F^2(L_L+L_V)}}\right\}, \quad \frac{\alpha_0}{\beta_0^2} \le \frac{4B_G}{11B_F^2L_V},$$
$$\zeta_0 \le \beta_0, \quad \frac{\zeta_0}{\alpha_0} \le \min\left\{\frac{C_L^2w_0^2\tau_0^2}{512L_D^2}, \frac{C_L^2w_0\tau_0^2}{128L_DL_L}, \frac{(1-\gamma)C_L^2\tau_0^2}{6144L_V^2L_D^2}\right\}.$$

We can relate the second term on the right hand side of (55) to $\|\nabla_x\Phi_{\tau_0}(x_k)\|^2$ using Lemma 11

$$\begin{aligned}
&\frac{64(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_0^2}\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2 \\
&\le \frac{128(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_0^2}\|\nabla_x\Phi_{\tau_0}(x_k)\|^2 + \frac{128(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_0^2}\|\nabla_x\Phi_{\tau_0}(x_k) - \nabla_x\Phi_{w_k,\tau_0}(x_k)\|^2 \\
&\le \frac{128(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_0^2}\|\nabla_x\Phi_{\tau_0}(x_k)\|^2 + \frac{128(L_L^2+L_V^2)\zeta_k^2}{C_L^2\alpha_k\tau_0^2}\Big(\frac{4L_fL_Vw_k}{C_L\underline{\sigma}\tau_0}(L_f + \frac{2L_fL_{V,2}}{C_L\tau_0})\Big)^2 \\
&\le \frac{\zeta_k}{4}\|\nabla_x\Phi_{\tau_0}(x_k)\|^2 + \frac{128(L_L^2+L_V^2)\zeta_k^2w_k^2}{C_L^2\alpha_k\tau_0^2}\Big(\frac{4L_fL_V}{C_L\underline{\sigma}\tau_0}(L_f + \frac{2L_fL_{V,2}}{C_L\tau_0})\Big)^2 \\
&= \frac{\zeta_k}{4}\|\nabla_x\Phi_{\tau_0}(x_k)\|^2 + \frac{128(L_L^2+L_V^2)C_{2,\tau_0}\zeta_k^2w_k^2}{C_L^2\alpha_k\tau_0^2},
\end{aligned} \tag{56}$$

where the third inequality is due to the step size condition $\frac{\zeta_0}{\alpha_0} \le \frac{C_L^2\tau_0^2}{512(L_L^2+L_V^2)}$.

Combining (55) and (56) and plugging in the step size decay rates,

$$\begin{aligned}
\frac{\zeta_k}{4}\mathbb{E}[\|\nabla_x\Phi_{\tau_0}(x_k)\|^2] &\le \mathbb{E}[\Phi_{\tau_0}(x_k) - \Phi_{\tau_0}(x_{k+1}) + \varepsilon_k^{\theta,\mathcal{L}} - \varepsilon_{k+1}^{\theta,\mathcal{L}} + \varepsilon_k^{\theta} - \varepsilon_{k+1}^{\theta} + \varepsilon_k^{V} - \varepsilon_{k+1}^{V} + \varepsilon_k^{V,\mathcal{L}} - \varepsilon_{k+1}^{V,\mathcal{L}}] \\
&\quad + \frac{128(L_L^2+L_V^2)C_{2,\tau_0}\zeta_k^2w_k^2}{C_L^2\alpha_k\tau_0^2} + 2C_{2,\tau_0}\zeta_kw_k^2 + \frac{4B_D^2L_\Phi\zeta_k^2}{w_k^2\tau_0} + \frac{B_D^2L_{\Phi,\tau_0}\zeta_k^2}{2w_k^2} \\
&\quad + \frac{4B_DB_FL_L\zeta_k\alpha_k}{w_k} + 48B_G\beta_k^2 + \frac{128L_V^2\tau_0^2}{3(1-\gamma)\beta_k(k+1)^2} + \frac{48\log|\mathcal{A}|\tau_0}{3(1-\gamma)(k+1)} \\
&\le \mathbb{E}[\Phi_{\tau_0}(x_k) - \Phi_{\tau_0}(x_{k+1}) + \varepsilon_k^{\theta,\mathcal{L}} - \varepsilon_{k+1}^{\theta,\mathcal{L}} + \varepsilon_k^{\theta} - \varepsilon_{k+1}^{\theta} + \varepsilon_k^{V} - \varepsilon_{k+1}^{V} + \varepsilon_k^{V,\mathcal{L}} - \varepsilon_{k+1}^{V,\mathcal{L}}] \\
&\quad + \mathcal{O}\left(\frac{\zeta_k^2w_k^2}{\alpha_k} + \zeta_kw_k^2 + \frac{\zeta_k^2}{w_k^2} + \frac{\zeta_k\alpha_k}{w_k} + \beta_k^2 + \frac{1}{\beta_k(k+1)^2} + \frac{1}{k+1}\right) \\
&\le \mathbb{E}[\Phi_{\tau_0}(x_k) - \Phi_{\tau_0}(x_{k+1}) + \varepsilon_k^{\theta,\mathcal{L}} - \varepsilon_{k+1}^{\theta,\mathcal{L}} + \varepsilon_k^{\theta} - \varepsilon_{k+1}^{\theta} + \varepsilon_k^{V} - \varepsilon_{k+1}^{V} + \varepsilon_k^{V,\mathcal{L}} - \varepsilon_{k+1}^{V,\mathcal{L}}] \\
&\quad + \mathcal{O}\left(\frac{1}{k+1}\right).
\end{aligned}$$

Re-arranging the terms and summing over iterations,

$$\sum_{t=0}^{k-1}\frac{\zeta_0}{4(t+1)^{2/3}}\mathbb{E}[\|\nabla_x\Phi_{\tau_0}(x_t)\|^2] \le \Phi_{\tau_0}(x_0) + \varepsilon_0^{\theta} + \varepsilon_0^{\theta,\mathcal{L}} + \varepsilon_0^{V} + \varepsilon_0^{V,\mathcal{L}} + \mathcal{O}\left(\sum_{t=0}^{k-1}\frac{1}{(t+1)}\right). \tag{57}$$

Again, the following inequalities on the summation of step sizes are standard results in the literature.

$$\sum_{t=0}^{k-1}\frac{1}{t+1} \leqslant \frac{\log(k+1)}{\log(2)},$$

$$\sum_{t=0}^{k} \frac{1}{(t+1)^u} \geq \frac{(1-\frac{1}{2^{1-u}})(k+1)^{1-u}}{1-u}, \quad \forall u \in (0,1)$$

and with $u = 2/3$,

$$\sum_{t=0}^{k} \frac{1}{(t+1)^{2/3}} \geq \frac{0.2(k+1)^{1/3}}{1/3} = \frac{3(k+1)^{1/3}}{5}.$$

This allows us to further simplify (57)

$$\begin{aligned}
\min_{t<k} \mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_t)\|^2] &\leq \frac{1}{\sum_{t=0}^{k-1} \frac{\zeta_0}{4(t+1)^{2/3}}} \sum_{t=0}^{k-1} \frac{\zeta_0}{4(t+1)^{2/3}} \mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_t)\|^2] \\
&\leq \frac{20}{3\zeta_0 (k+1)^{1/3}} \left(\Phi_{\tau_0}(x_0) + \varepsilon_0^{\theta} + \varepsilon_0^{\theta,\mathcal{L}} + \varepsilon_0^{V} + \varepsilon_0^{V,\mathcal{L}} + \mathcal{O}(\log(k+1)) \right).
\end{aligned}$$

■

D Proof of Propositions

D.1 Proof of Proposition 1

We know from Lemma 10 that under the step size condition $\tau_k \leq \frac{2L_V}{C_L}$, the objective Φ_{τ_k} has $\frac{L_\Phi}{\tau_k}$-Lipschitz gradients. This implies

$$\begin{aligned}
&\Phi_{\tau_k}(x_{k+1}) - \Phi_{\tau_k}(x_k) \\
&\leq \langle \nabla_x \Phi_{\tau_k}(x_k), x_{k+1} - x_k \rangle + \frac{L_\Phi}{2\tau_k} \|x_{k+1} - x_k\|^2 \\
&= -\zeta_k \langle \nabla_x \Phi_{\tau_k}(x_k), D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k) \rangle + \frac{L_\Phi \zeta_k^2}{2\tau_k} \|D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k)\|^2 \\
&= -\zeta_k \|\nabla_x \Phi_{\tau_k}(x_k)\|^2 + \zeta_k \langle \nabla_x \Phi_{\tau_k}(x_k), \nabla_x \Phi_{\tau_k}(x_k) - D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k) \rangle \\
&\quad + \frac{L_\Phi \zeta_k^2}{2\tau_k} \|D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k)\|^2.
\end{aligned}$$

By the law of total expectation,

$$\begin{aligned}
&\mathbb{E}[\Phi_{\tau_k}(x_{k+1}) - \Phi_{\tau_k}(x_k)] \\
&\leq -\zeta_k \mathbb{E}[\|\nabla_x \Phi_{\tau_k}(x_k)\|^2] + \frac{L_\Phi \zeta_k^2}{2\tau_k} \mathbb{E}[\|D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k)\|^2] \\
&\quad + \zeta_k \mathbb{E}[\langle \nabla_x \Phi_{\tau_k}(x_k), \nabla_x \Phi_{\tau_k}(x_k) - \mathbb{E}[D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k) \mid \mathcal{F}_{k-1}] \rangle] \\
&= -\zeta_k \mathbb{E}[\|\nabla_x \Phi_{\tau_k}(x_k)\|^2] + \frac{L_\Phi \zeta_k^2}{2\tau_k} \mathbb{E}[\|D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k)\|^2] \\
&\quad + \zeta_k \mathbb{E}[\langle \nabla_x \Phi_{\tau_k}(x_k), \nabla_x \Phi_{\tau_k}(x_k) - \bar{D}_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}) \rangle] \\
&\leq -\zeta_k \mathbb{E}[\|\nabla_x \Phi_{\tau_k}(x_k)\|^2] + \frac{B_D^2 L_\Phi \zeta_k^2}{2\tau_k w_k^2} + \frac{\zeta_k}{2} \mathbb{E}[\|\nabla_x \Phi_{\tau_k}(x_k)\|^2] + \frac{\zeta_k}{2} \mathbb{E}[\|\nabla_x \Phi_{\tau_k}(x_k) - \bar{D}_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}})\|^2] \\
&= -\frac{\zeta_k}{2} \mathbb{E}[\|\nabla_x \Phi_{\tau_k}(x_k)\|^2] + \frac{B_D^2 L_\Phi \zeta_k^2}{2\tau_k w_k^2} \\
&\quad + \frac{\zeta_k}{2} \mathbb{E}\left[\left\| \left(\nabla_x \Phi_{\tau_k}(x_k) - \nabla_x \Phi_{w_k,\tau_k}(x_k) \right) + \left(\bar{D}_{w_k}(x_k, \pi_{\tau_k}^\star(x_k), \pi_{w_k,\tau_k}^\star(x_k)) - \bar{D}_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}) \right) \right\|^2 \right] \\
&\leq -\frac{\zeta_k}{2} \mathbb{E}[\|\nabla_x \Phi_{\tau_k}(x_k)\|^2] + \zeta_k \mathbb{E}[\|\bar{D}_{w_k}(x_k, \pi_{\tau_k}^\star(x_k), \pi_{w_k,\tau_k}^\star(x_k)) - \bar{D}_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}})\|^2] \\
&\quad + \zeta_k \mathbb{E}[\|\nabla_x \Phi_{\tau_k}(x_k) - \nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2] + \frac{B_D^2 L_\Phi \zeta_k^2}{2\tau_k w_k^2},
\end{aligned} \tag{58}$$

where the second inequality applies Lemma 5 and the last equation follows from (29).

To bound the second term on the right hand side of (58), we apply Lemma 6

$$\begin{aligned}
&\|\bar{D}_{w_k}(x_k, \pi^\star_{\tau_k}(x_k), \pi^\star_{w_k,\tau_k}(x_k)) - \bar{D}_{w_k}(x_k, \pi_k, \pi^{\mathcal{L}}_k)\|^2 \\
&\leq \frac{L_D^2}{w_k^2}\Big(\|\pi_k - \pi^\star_{\tau_k}(x_k)\| + \|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|\Big)^2 \\
&\leq \frac{2L_D^2}{w_k^2}\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \frac{2L_D^2}{w_k^2}\|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|^2.
\end{aligned} \tag{59}$$

For the third term of (58), we have from Lemma 11

$$\|\nabla_x \Phi_{\tau_k}(x_k) - \nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2 \leq \Big(\frac{4L_f L_V w_k}{C_L \underline{\sigma}\tau_k}(L_f + \frac{2L_f L_{V,2}}{C_L \tau_k})\Big)^2. \tag{60}$$

Substituting (59) and (60) into (58),

$$\begin{aligned}
&\mathbb{E}[\Phi_{\tau_k}(x_{k+1}) - \Phi_{\tau_k}(x_k)] \\
&\leq -\frac{\zeta_k}{2}\mathbb{E}[\|\nabla_x \Phi_{\tau_k}(x_k)\|^2] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|^2] \\
&\quad + \zeta_k(\frac{4L_f L_V w_k}{C_L \underline{\sigma}\tau_k}(L_f + \frac{2L_f L_{V,2}}{C_L \tau_k})\Big)^2 + \frac{B_D^2 L_\Phi \zeta_k^2}{2\tau_k w_k^2}.
\end{aligned} \tag{61}$$

Under the choice of step size $\tau_k \leq \frac{2L_{V,2}}{C_L}$, we can further simplify (61)

$$\begin{aligned}
&\mathbb{E}[\Phi_{\tau_k}(x_{k+1}) - \Phi_{\tau_k}(x_k)] \\
&\leq -\frac{\zeta_k}{2}\mathbb{E}[\|\nabla_x \Phi_{\tau_k}(x_k)\|^2] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|^2] \\
&\quad + \zeta_k(\frac{4L_f L_V w_k}{C_L \underline{\sigma}\tau_k} \cdot \frac{4L_f L_{V,2}}{C_L \tau_k}\Big)^2 + \frac{B_D^2 L_\Phi \zeta_k^2}{2\tau_k w_k^2} \\
&= -\frac{\zeta_k}{2}\mathbb{E}[\|\nabla_x \Phi_{\tau_k}(x_k)\|^2] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|^2] \\
&\quad + \frac{256 L_f^4 L_V^2 L_{V,2}^2 \zeta_k w_k^2}{C_L^4 \underline{\sigma}^2 \tau_k^4} + \frac{B_D^2 L_\Phi \zeta_k^2}{2\tau_k w_k^2}.
\end{aligned} \tag{62}$$

The next step is to bridge the gap between $\Phi_{\tau_k}(x_{k+1})$ and $\Phi_{\tau_{k+1}}(x_{k+1})$. By the definition of Φ_τ in (7)

$$\begin{aligned}
\Phi_{\tau_{k+1}}(x_{k+1}) - \Phi_{\tau_k}(x_{k+1}) &= f(x_{k+1}, \pi^\star_{\tau_{k+1}}(x_{k+1})) - f(x_{k+1}, \pi^\star_{\tau_k}(x_{k+1})) \\
&\leq L_f \|\pi^\star_{\tau_{k+1}}(x_{k+1}) - \pi^\star_{\tau_k}(x_{k+1})\| \\
&\leq L_f \cdot \frac{6(\tau_k - \tau_{k+1})|\mathcal{S}|\log|\mathcal{A}|}{(1-\gamma)C_L \tau_k} \\
&\leq \frac{16 L_f |\mathcal{S}|\log|\mathcal{A}|}{(1-\gamma)C_L(k+1)},
\end{aligned} \tag{63}$$

where the first inequality is due to Assumption 3, and the second inequality is due to Lemma 7, and the last inequality applies Lemma 2.

Finally, we bridge the gap between $\|\nabla_x \Phi_{\tau_k}(x_k)\|^2$ and $\|\nabla_x \Phi(x_k)\|^2$ by invoking Lemma 12

$$\|\nabla_x \Phi_{\tau_k}(x) - \nabla_x \Phi(x)\| \leq \frac{L_\star L_f L_V^2 L_{V,2}\tau_k}{\underline{\sigma}^2}. \tag{64}$$

Combining (62)-(64),

$$\mathbb{E}[\Phi_{\tau_{k+1}}(x_{k+1}) - \Phi_{\tau_k}(x_k)]$$

$$
\begin{aligned}
&= \mathbb{E}[\Phi_{\tau_k}(x_{k+1}) - \Phi_{\tau_k}(x_k)] + \mathbb{E}[\Phi_{\tau_{k+1}}(x_{k+1}) - \Phi_{\tau_k}(x_{k+1})] \\
&\leq -\frac{\zeta_k}{2}\mathbb{E}[\|\nabla_x \Phi_{\tau_k}(x_k)\|^2] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2] \\
&\quad + \frac{256L_f^4L_V^2L_{V,2}^2\zeta_k w_k^2}{C_L^4\underline{\sigma}^2\tau_k^4} + \frac{B_D^2L_\Phi\zeta_k^2}{2\tau_k w_k^2} + \frac{16L_f|\mathcal{S}|\log|\mathcal{A}|}{(1-\gamma)C_L(k+1)} \\
&\leq -\frac{\zeta_k}{4}\mathbb{E}[\|\nabla_x \Phi(x_k)\|^2] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2] \\
&\quad + \frac{L_\star^2L_f^2L_V^4L_{V,2}^2\zeta_k\tau_k^2}{2\underline{\sigma}^4} + \frac{256L_f^4L_V^2L_{V,2}^2\zeta_k w_k^2}{C_L^4\underline{\sigma}^2\tau_k^4} + \frac{B_D^2L_\Phi\zeta_k^2}{2\tau_k w_k^2} + \frac{16L_f|\mathcal{S}|\log|\mathcal{A}|}{(1-\gamma)C_L(k+1)},
\end{aligned}
$$

where the last inequality follows from the simple fact that $-\frac{a^2}{2} \leq -\frac{b^2}{4} + \frac{(a-b)^2}{2}$ for any scalar a, b.

■

D.2 Proof of Proposition 2

The proof is almost identical to that of Proposition 1. We include the proof here for completeness.

From Lemma 10, we know that under a fixed regularization weight τ_0, the objective Φ_{τ_0} has L_{Φ,τ_0} Lipschitz gradients, where we define

$$
L_{\Phi,\tau_0} \triangleq (1 + \frac{2L_V}{C_L\tau_0})\Big(\frac{2L_fL_V}{C_L\tau_0} + \frac{2L_fL_VL_{V,2}}{\underline{\sigma}C_L\tau_0} + \frac{2L_fL_V^2L_{V,2}}{\underline{\sigma}^2C_L\tau_0} + \frac{2L_fL_V^2}{\underline{\sigma}C_L\tau}\Big).
$$

This implies

$$
\begin{aligned}
&\Phi_{\tau_0}(x_{k+1}) - \Phi_{\tau_0}(x_k) \\
&\leq \langle \nabla_x \Phi_{\tau_0}(x_k), x_{k+1} - x_k\rangle + \frac{L_{\Phi,\tau_0}}{2}\|x_{k+1} - x_k\|^2 \\
&= -\zeta_k\langle \nabla_x \Phi_{\tau_0}(x_k), D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k)\rangle + \frac{L_{\Phi,\tau_0}\zeta_k^2}{2}\|D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k)\|^2 \\
&= -\zeta_k\|\nabla_x \Phi_{\tau_0}(x_k)\|^2 + \zeta_k\langle \nabla_x \Phi_{\tau_0}(x_k), \nabla_x \Phi_{\tau_0}(x_k) - D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k)\rangle \\
&\quad + \frac{L_{\Phi,\tau_0}\zeta_k^2}{2}\|D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k)\|^2.
\end{aligned} \tag{65}
$$

By the law of total expectation,

$$
\begin{aligned}
&\mathbb{E}[\Phi_{\tau_0}(x_{k+1}) - \Phi_{\tau_0}(x_k)] \\
&\leq -\zeta_k\mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_k)\|^2] + \frac{L_{\Phi,\tau_0}\zeta_k^2}{2}\mathbb{E}[\|D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k)\|^2] \\
&\quad + \zeta_k\mathbb{E}[\langle \nabla_x \Phi_{\tau_0}(x_k), \nabla_x \Phi_{\tau_0}(x_k) - \mathbb{E}[D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k) \mid \mathcal{F}_{k-1}]\rangle] \\
&= -\zeta_k\mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_k)\|^2] + \frac{L_{\Phi,\tau_0}\zeta_k^2}{2}\mathbb{E}[\|D_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k)\|^2] \\
&\quad + \zeta_k\mathbb{E}[\langle \nabla_x \Phi_{\tau_0}(x_k), \nabla_x \Phi_{\tau_0}(x_k) - \bar{D}_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}})\rangle] \\
&\leq -\zeta_k\mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_k)\|^2] + \frac{B_D^2L_{\Phi,\tau_0}\zeta_k^2}{2w_k^2} + \frac{\zeta_k}{2}\mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_k)\|^2] + \frac{\zeta_k}{2}\mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_k) - \bar{D}_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}})\|^2] \\
&= -\frac{\zeta_k}{2}\mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_k)\|^2] + \frac{B_D^2L_{\Phi,\tau_0}\zeta_k^2}{2w_k^2} \\
&\quad + \frac{\zeta_k}{2}\mathbb{E}\left[\left\|\Big(\nabla_x \Phi_{\tau_0}(x_k) - \nabla_x \Phi_{w_k,\tau_0}(x_k)\Big) + \Big(\bar{D}_{w_k}(x_k, \pi^\star_{\tau_0}(x_k), \pi^\star_{w_k,\tau_0}(x_k)) - \bar{D}_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}})\Big)\right\|^2\right] \\
&\leq -\frac{\zeta_k}{2}\mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_k)\|^2] + \zeta_k\mathbb{E}[\|\bar{D}_{w_k}(x_k, \pi^\star_{\tau_0}(x_k), \pi^\star_{w_k,\tau_0}(x_k)) - \bar{D}_{w_k}(x_k, \pi_k, \pi_k^{\mathcal{L}})\|^2] \\
&\quad + \zeta_k\mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_k) - \nabla_x \Phi_{w_k,\tau_0}(x_k)\|^2] + \frac{B_D^2L_{\Phi,\tau_0}\zeta_k^2}{2w_k^2},
\end{aligned} \tag{66}
$$

where the second inequality applies Lemma 5 and the last equation follows from (29).

To bound the second term on the right hand side of (66), we apply Lemma 6

$$\begin{aligned}
&\|\bar{D}_{w_k}(x_k, \pi^\star_{\tau_0}(x_k), \pi^\star_{w_k,\tau_0}(x_k)) - \bar{D}_{w_k}(x_k, \pi_k, \pi^{\mathcal{L}}_k)\|^2 \\
&\leq \frac{L_D^2}{w_k^2}\Big(\|\pi_k - \pi^\star_{\tau_0}(x_k)\| + \|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_0}(x_k)\|\Big)^2 \\
&\leq \frac{2L_D^2}{w_k^2}\|\pi_k - \pi^\star_{\tau_0}(x_k)\|^2 + \frac{2L_D^2}{w_k^2}\|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_0}(x_k)\|^2. \qquad (67)
\end{aligned}$$

For the third term of (66), we have from Lemma 11

$$\|\nabla_x \Phi_{\tau_0}(x_k) - \nabla_x \Phi_{w_k,\tau_0}(x_k)\|^2 \leq \Big(\frac{4L_f L_V w_k}{C_L \underline{\sigma}\tau_0}(L_f + \frac{2L_f L_{V,2}}{C_L \tau_0})\Big)^2 = C_{2,\tau_0} w_k^2, \qquad (68)$$

where we define $C_{2,\tau_0} = \Big(\frac{4L_f L_V}{C_L \underline{\sigma}\tau_0}(L_f + \frac{2L_f L_{V,2}}{C_L \tau_0})\Big)^2$.

Substituting (67) and (68) into (66),

$$\begin{aligned}
&\mathbb{E}[\Phi_{\tau_0}(x_{k+1}) - \Phi_{\tau_0}(x_k)] \\
&\leq -\frac{\zeta_k}{2}\mathbb{E}[\|\nabla_x \Phi_{\tau_0}(x_k)\|^2] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_0}(x_k)\|^2] + \frac{2L_D^2\zeta_k}{w_k^2}\mathbb{E}[\|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_0}(x_k)\|^2] \\
&\quad + C_{2,\tau_0}\zeta_k w_k^2 + \frac{B_D^2 L_{\Phi,\tau_0}\zeta_k^2}{2w_k^2}.
\end{aligned}$$

■

D.3 Proof of Proposition 3

The proof depends on an intermediate result that bounds an important cross term. We state it in the lemma below and defer its proof to Section E.13.

Lemma 13 *Under the assumptions and step sizes of Proposition 3, we have for all $k \geq 0$*

$$\begin{aligned}
&\mathbb{E}[-\langle \nabla_x J_{\tau_k}(x_k, \pi_{\theta_k}) - \nabla_x J_{\tau_k}(x_k, \pi_{\theta^\star_{\tau_k}(x_k)}), x_{k+1} - x_k\rangle] \\
&\leq \frac{C_L^2 \alpha_k \tau_k^2}{64}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] \\
&\quad + \frac{64L_D^2 L_V^2 \zeta_k^2}{C_L^2 \alpha_k w_k^2 \tau_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|^2] + \frac{32L_V^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2].
\end{aligned}$$

We now proceed to the proof of Proposition 3. We consider the following decomposition and bound each term on the right hand side individually.

$$\begin{aligned}
&- J_{\tau_{k+1}}(x_{k+1}, \pi_{\theta_{k+1}}) + J_{\tau_k}(x_k, \pi_{\theta_k}) \\
&= \Big(- J_{\tau_k}(x_{k+1}, \pi_{\theta_{k+1}}) + J_{\tau_k}(x_{k+1}, \pi_{\theta_k})\Big) + \Big(- J_{\tau_k}(x_{k+1}, \pi_{\theta_k}) + J_{\tau_k}(x_k, \pi_{\theta_k})\Big) \\
&\quad + \Big(- J_{\tau_{k+1}}(x_{k+1}, \pi_{\theta_{k+1}}) + J_{\tau_k}(x_{k+1}, \pi_{\theta_{k+1}})\Big). \qquad (69)
\end{aligned}$$

Bound the First Term of (69)**.** As J_τ has L_V-Lipschitz gradients,

$$\begin{aligned}
&- J_{\tau_k}(x_{k+1}, \pi_{\theta_{k+1}}) + J_{\tau_k}(x_{k+1}, \pi_{\theta_k}) \\
&\leq \langle -\nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k}), \theta_{k+1} - \theta_k\rangle + \frac{L_V}{2}\|\theta_{k+1} - \theta_k\|^2
\end{aligned}$$

$$
\begin{aligned}
&= -\alpha_k \langle \nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k}), F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s'_k, \xi_k)\rangle + \frac{L_V \alpha_k^2}{2} \|F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s'_k, \xi_k)\|^2 \\
&= -\alpha_k \langle \nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k}), F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s'_k, \xi_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\rangle \\
&\quad - \alpha_k \langle \nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k}), \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, V_{\tau_k}^{x_k, \pi_{\theta_k}})\rangle \\
&\quad - \alpha_k \langle \nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k}), \nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\rangle + \frac{L_V \alpha_k^2}{2} \|F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s'_k, \xi_k)\|^2, \qquad (70)
\end{aligned}
$$

where the final equation follows from $\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k}) = \bar{F}_{0,\tau_k}(x_k, \theta_k, V_{\tau_k}^{x_k, \pi_{\theta_k}})$.

To bound the first term of (70),

$$
\begin{aligned}
&- \alpha_k \mathbb{E}[\langle \nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k}), F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s'_k, \xi_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\rangle] \\
&= -\alpha_k \mathbb{E}[\langle \nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k}), \mathbb{E}[F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s'_k, \xi_k) \mid \mathcal{F}_{k-1}] - \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\rangle] \\
&\quad + \alpha_k \mathbb{E}[\langle \nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k}) - \nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k}), F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s'_k, \xi_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\rangle] \\
&= \alpha_k \mathbb{E}[\langle \nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k}) - \nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k}), F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s'_k, \xi_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\rangle] \\
&\le \alpha_k \cdot L_V \mathbb{E}[\|x_{k+1} - x_k\|] \cdot 2B_F \\
&\le 2B_F L_V \alpha_k \cdot \frac{B_D \zeta_k}{w_k} \\
&= \frac{2B_D B_F L_V \zeta_k \alpha_k}{w_k}, \qquad (71)
\end{aligned}
$$

where the second inequality follows from Lemma 5.

For the second term of (70),

$$
\begin{aligned}
&- \alpha_k \langle \nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k}), \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, V_{\tau_k}^{x_k, \pi_{\theta_k}})\rangle \\
&\le \frac{\alpha_k}{8} \|\nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k})\|^2 + 2\alpha_k \|\bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, V_{\tau_k}^{x_k, \pi_{\theta_k}})\|^2 \\
&\le \frac{\alpha_k}{4} \|\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2 + \frac{\alpha_k}{4} \|\nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k}) - \nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2 \\
&\quad + 2\alpha_k \|\bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, V_{\tau_k}^{x_k, \pi_{\theta_k}})\|^2 \\
&\le \frac{\alpha_k}{4} \|\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2 + \frac{L_V^2 \alpha_k}{4} \Big(\|x_{k+1} - x_k\|^2\Big) + 2L_F^2 \alpha_k \|V_{\tau_k}^{x_k, \pi_{\theta_k}} - \hat{V}_k\|^2 \\
&\le \frac{\alpha_k}{4} \|\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2 + \frac{L_V^2 \alpha_k}{4} \cdot \Big(\frac{B_D \zeta_k}{w_k}\Big)^2 + 2L_F^2 \alpha_k \varepsilon_k^V \\
&\le \frac{\alpha_k}{4} \|\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2 + \frac{B_D^2 L_V^2 \zeta_k^2 \alpha_k}{4w_k^2} + 2L_F^2 \alpha_k \varepsilon_k^V, \qquad (72)
\end{aligned}
$$

where the third inequality is a result of the Lipschitz continuity of J_{τ_k} and $\bar{F}_{0,\tau_k}$.

For the third term of (70),

$$
\begin{aligned}
&- \alpha_k \langle \nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k}), \nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\rangle \\
&= -\alpha_k \|\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2 + \alpha_k \langle \nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k}) - \nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k}), \nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\rangle \\
&\le -\frac{\alpha_k}{2} \|\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2 + \frac{\alpha_k}{2} \|\nabla_\theta J_{\tau_k}(x_{k+1}, \pi_{\theta_k}) - \nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2 \\
&\le -\frac{\alpha_k}{2} \|\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2 + \frac{L_V^2 \alpha_k}{2} \Big(\|x_{k+1} - x_k\|^2\Big) \\
&\le -\frac{\alpha_k}{2} \|\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2 + \frac{L_V^2 \alpha_k}{2} \cdot \Big(\frac{B_D \zeta_k}{w_k}\Big)^2 \\
&\le -\frac{\alpha_k}{2} \|\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2 + \frac{B_D^2 L_V^2 \zeta_k^2 \alpha_k}{2w_k^2}, \qquad (73)
\end{aligned}
$$

where the second inequality follows from the L_V-smoothness continuity of J_τ from (35) of Lemma 3.

Substituting (71)-(73) into (70),

$$\begin{aligned}
&\mathbb{E}[-J_{\tau_{k+1}}(x_{k+1},\pi_{\theta_{k+1}})+J_{\tau_k}(x_k,\pi_{\theta_k})]\\
&\leq \frac{2B_DB_FL_V\zeta_k\alpha_k}{w_k}+\frac{\alpha_k}{4}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2]+\frac{B_D^2L_V^2\zeta_k^2\alpha_k}{4w_k^2}+2L_F^2\alpha_k\mathbb{E}[\varepsilon_k^V]\\
&\quad-\frac{\alpha_k}{2}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2]+\frac{B_D^2L_V^2\zeta_k^2\alpha_k}{2w_k^2}+\frac{L_V\alpha_k^2}{2}\mathbb{E}[\|F_{0,\tau_k}(x_k,\theta_k,\hat{V}_k,s_k,a_k,s_k',\xi_k)\|^2]\\
&\leq -\frac{\alpha_k}{4}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2]+2L_F^2\alpha_k\mathbb{E}[\varepsilon_k^V]+\frac{3B_D^2L_V^2\zeta_k^2\alpha_k}{4w_k^2}+\frac{2B_DB_FL_V\zeta_k\alpha_k}{w_k}+\frac{B_F^2L_V\alpha_k^2}{2}\\
&\leq -\frac{\alpha_k}{8}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2]+2L_F^2\alpha_k\mathbb{E}[\varepsilon_k^V]+\frac{3B_D^2L_V^2\zeta_k^2\alpha_k}{4w_k^2}+\frac{2B_DB_FL_V\zeta_k\alpha_k}{w_k}+\frac{B_F^2L_V\alpha_k^2}{2}\\
&\quad-\frac{C_L^2\alpha_k\tau_k^2}{32}\mathbb{E}[\|\pi_{\theta_k}-\pi^\star_{\tau_k}(x_k)\|^2],
\end{aligned}\tag{74}$$

where the last inequality plugs in the relationship

$$\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|\geq\frac{C_L\tau_k}{2}\|\pi_{\theta_k}-\pi^\star_{\tau_k}(x_k)\|.\tag{75}$$

Note that $J_\tau(x,\pi)=\lim_{w\to 0}w\mathcal{L}_{w,\tau}(x,\pi)$, which implies that (75) follows from (22).

Bound the Second Term of (69)**.** We use $\theta^\star_\tau(x)$ to denote a softmax parameter that encodes the policy $\pi^\star_\tau(x)$. Again, as J_τ has L_V-Lipschitz gradients,

$$\begin{aligned}
&-J_{\tau_k}(x_{k+1},\pi_{\theta_k})+J_{\tau_k}(x_k,\pi_{\theta_k})\\
&\leq -\langle\nabla_xJ_{\tau_k}(x_k,\pi_{\theta_k}),x_{k+1}-x_k\rangle+\frac{L_V}{2}\|x_{k+1}-x_k\|^2\\
&=-\langle\nabla_xJ_{\tau_k}(x_k,\pi_{\theta_k})-\nabla_xJ_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)}),x_{k+1}-x_k\rangle+\frac{L_V}{2}\|x_{k+1}-x_k\|^2\\
&\quad-\langle\nabla_xJ_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)}),x_{k+1}-x_k\rangle\\
&=-\langle\nabla_xJ_{\tau_k}(x_k,\pi_{\theta_k})-\nabla_xJ_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)}),x_{k+1}-x_k\rangle+\frac{L_V}{2}\|x_{k+1}-x_k\|^2\\
&\quad-\langle\nabla_x\ell_{\tau_k}(x_k),x_{k+1}-x_k\rangle\\
&\leq -\langle\nabla_xJ_{\tau_k}(x_k,\pi_{\theta_k})-\nabla_xJ_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)}),x_{k+1}-x_k\rangle+\frac{L_V}{2}\|x_{k+1}-x_k\|^2\\
&\quad+\Big(-\ell_{\tau_k}(x_{k+1})+\ell_{\tau_k}(x_k)\Big)+\frac{L_\Phi}{2\tau_k}\|x_{k+1}-x_k\|^2\\
&\leq -\langle\nabla_xJ_{\tau_k}(x_k,\pi_{\theta_k})-\nabla_xJ_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)}),x_{k+1}-x_k\rangle\\
&\quad+\Big(-J_{\tau_k}(x_{k+1},\pi^\star_{\tau_k}(x_{k+1}))+J_{\tau_k}(x_k,\pi^\star_{\tau_k}(x_k))\Big)+\frac{L_\Phi}{\tau_k}\|x_{k+1}-x_k\|^2,
\end{aligned}$$

where the last inequality is due to $L_V\leq L_\Phi$ and the step size condition $\tau_k\leq 1$, and the second equation uses the relationship $\nabla_xJ_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)})=\nabla_x\ell_{\tau_k}(x_k)$, which is due to $\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)})=0$ by the first-order optimality condition. The second inequality is due to the fact that ℓ_τ is $\frac{L_\Phi}{\tau}$-smooth when $\tau\leq 1$ (established in Lemma 10) and that for an L-smooth function f, we have

$$f(y)-f(x)\leq\langle\nabla f(x),y-x\rangle+\frac{L}{2}\|x-y\|^2.$$

Taking the expectation and plugging in the result from Lemma 13,

$$\begin{aligned}
&\mathbb{E}[-J_{\tau_k}(x_{k+1},\pi_{\theta_k})+J_{\tau_k}(x_k,\pi_{\theta_k})]\\
&\leq\frac{C_L^2\alpha_k\tau_k^2}{64}\mathbb{E}[\|\pi_k-\pi^\star_{\tau_k}(x_k)\|^2]\\
&\quad+\frac{64L_D^2L_V^2\zeta_k^2}{C_L^2\alpha_kw_k^2\tau_k^2}\mathbb{E}[\|\pi_k-\pi^\star_{\tau_k}(x_k)\|^2+\|\pi_k^{\mathcal{L}}-\pi^\star_{w_k,\tau_k}(x_k)\|^2]+\frac{32L_V^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}[-J_{\tau_k}(x_{k+1},\pi^\star_{\tau_k}(x_{k+1}))+J_{\tau_k}(x_k,\pi^\star_{\tau_k}(x_k))]+\frac{L_\Phi}{\tau_k}\mathbb{E}[\|x_{k+1}-x_k\|^2]\\
\le\ &\frac{C_L^2\alpha_k\tau_k^2}{64}\mathbb{E}[\|\pi_k-\pi^\star_{\tau_k}(x_k)\|^2]+\frac{64L_D^2L_V^2\zeta_k^2}{C_L^2\alpha_k w_k^2\tau_k^2}\mathbb{E}[\|\pi_k-\pi^\star_{\tau_k}(x_k)\|^2+\|\pi_k^{\mathcal{L}}-\pi^\star_{w_k,\tau_k}(x_k)\|^2]\\
&+\mathbb{E}[-J_{\tau_k}(x_{k+1},\pi^\star_{\tau_k}(x_{k+1}))+J_{\tau_k}(x_k,\pi^\star_{\tau_k}(x_k))]+\frac{32L_V^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]+\frac{B_D^2L_\Phi\zeta_k^2}{w_k^2\tau_k},
\end{aligned}
\tag{76}
$$

where the last inequality follows from $\|x_{k+1}-x_k\|\le\frac{B_D\zeta_k}{w_k}$.

Bound the Third Term of (69)**.**

$$
\begin{aligned}
-J_{\tau_{k+1}}(x_{k+1},\pi_{\theta_{k+1}})+J_{\tau_k}(x_{k+1},\pi_{\theta_{k+1}})&=\frac{\tau_k-\tau_{k+1}}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_{k+1}}}[E(\pi_{k+1},s)]\\
&\le\frac{\log|\mathcal{A}|(\tau_k-\tau_{k+1})}{(1-\gamma)}\\
&\le\frac{8\log|\mathcal{A}|\tau_k}{3(1-\gamma)(k+1)},
\end{aligned}
\tag{77}
$$

where the second inequality follows from Lemma 2.

Collecting the bounds in (74)-(77) and plugging them into (69), we get

$$
\begin{aligned}
&\mathbb{E}[-J_{\tau_{k+1}}(x_{k+1},\pi_{\theta_{k+1}})+J_{\tau_k}(x_k,\pi_{\theta_k})]\\
\le\ &-\frac{\alpha_k}{8}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2]+2L_F^2\alpha_k\mathbb{E}[\varepsilon_k^V]+\frac{3B_D^2L_V^2\zeta_k^2\alpha_k}{4w_k^2}+\frac{2B_DB_FL_V\zeta_k\alpha_k}{w_k}+\frac{B_F^2L_V\alpha_k^2}{2}\\
&-\frac{C_L^2\alpha_k\tau_k^2}{32}\mathbb{E}[\|\pi_{\theta_k}-\pi^\star_{\tau_k}(x_k)\|^2]\\
&+\frac{C_L^2\alpha_k\tau_k^2}{64}\mathbb{E}[\|\pi_k-\pi^\star_{\tau_k}(x_k)\|^2]+\frac{64L_D^2L_V^2\zeta_k^2}{C_L^2\alpha_k w_k^2\tau_k^2}\mathbb{E}[\|\pi_k-\pi^\star_{\tau_k}(x_k)\|^2+\|\pi_k^{\mathcal{L}}-\pi^\star_{w_k,\tau_k}(x_k)\|^2]\\
&+\mathbb{E}[-J_{\tau_k}(x_{k+1},\pi^\star_{\tau_k}(x_{k+1}))+J_{\tau_k}(x_k,\pi^\star_{\tau_k}(x_k))]+\frac{32L_V^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]+\frac{B_D^2L_\Phi\zeta_k^2}{w_k^2\tau_k}\\
&+\frac{8\log|\mathcal{A}|\tau_k}{3(1-\gamma)(k+1)}\\
\le\ &-\frac{\alpha_k}{8}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2]+2L_F^2\alpha_k\mathbb{E}[\varepsilon_k^V]+\frac{32L_V^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]\\
&-\frac{C_L^2\alpha_k\tau_k^2}{64}\mathbb{E}[\|\pi_k-\pi^\star_{\tau_k}(x_k)\|^2]+\frac{64L_D^2L_V^2\zeta_k^2}{C_L^2\alpha_k w_k^2\tau_k^2}\mathbb{E}[\|\pi_k-\pi^\star_{\tau_k}(x_k)\|^2+\|\pi_k^{\mathcal{L}}-\pi^\star_{w_k,\tau_k}(x_k)\|^2]\\
&+\frac{2B_D^2L_\Phi\zeta_k^2}{w_k^2\tau_k}+\frac{2B_DB_FL_V\zeta_k\alpha_k}{w_k}+\frac{B_F^2L_V\alpha_k^2}{2}+\frac{8\log|\mathcal{A}|\tau_k}{3(1-\gamma)(k+1)}\\
&+\mathbb{E}[-J_{\tau_k}(x_{k+1},\pi^\star_{\tau_k}(x_{k+1}))+J_{\tau_k}(x_k,\pi^\star_{\tau_k}(x_k))],
\end{aligned}
$$

where in the last inequality we have combined the terms $\frac{3B_D^2L_V^2\zeta_k^2\alpha_k}{4w_k^2}$ and $\frac{B_D^2L_\Phi\zeta_k^2}{w_k^2\tau_k}$ under the step size conditions $\tau_k\le 1$ and $\alpha_k\le\frac{4L_\Phi}{3L_V^2}$.

Recall the definition of ε_k^θ in (33). We can re-arrange the terms in the inequality above and obtain

$$
\begin{aligned}
&\mathbb{E}[\varepsilon_{k+1}^\theta-\varepsilon_k^\theta]\\
&=\mathbb{E}[-J_{\tau_{k+1}}(x_{k+1},\pi_{\theta_{k+1}})+J_{\tau_k}(x_k,\pi_{\theta_k})]\\
&\quad+\mathbb{E}[J_{\tau_k}(x_{k+1},\pi^\star_{\tau_k}(x_{k+1}))-J_{\tau_k}(x_k,\pi^\star_{\tau_k}(x_k))]\\
&\quad+\mathbb{E}[J_{\tau_{k+1}}(x_{k+1},\pi^\star_{\tau_k}(x_{k+1}))-J_{\tau_k}(x_{k+1},\pi^\star_{\tau_{k+1}}(x_{k+1}))]\\
&\le-\frac{\alpha_k}{8}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2]+2L_F^2\alpha_k\mathbb{E}[\varepsilon_k^V]+\frac{32L_V^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]
\end{aligned}
$$

$$- \frac{C_L^2 \alpha_k \tau_k^2}{64} \mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] + \frac{64 L_D^2 L_V^2 \zeta_k^2}{C_L^2 \alpha_k w_k^2 \tau_k^2} \mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{2B_D^2 L_\Phi \zeta_k^2}{w_k^2 \tau_k} + \frac{2B_D B_F L_V \zeta_k \alpha_k}{w_k} + \frac{B_F^2 L_V \alpha_k^2}{2} + \frac{16 \log |\mathcal{A}| \tau_k}{3(1-\gamma)(k+1)},$$

where the bound on $J_{\tau_{k+1}}(x_{k+1}, \pi^\star_{\tau_k}(x_{k+1})) - J_{\tau_k}(x_{k+1}, \pi^\star_{\tau_k}(x_{k+1})) \leq \frac{8 \log |\mathcal{A}| \tau_k}{3(1-\gamma)(k+1)}$ can be obtained in a manner similar to (77).

■

D.4 Proof of Proposition 4

The proof depends on an intermediate result that bounds an important cross term. We state it in the lemma below and defer its proof to Section E.14.

Lemma 14 *Under the assumptions and step sizes of Proposition 4, we have for all $k \geq 0$*

$$\mathbb{E}[\langle \nabla_x \mathcal{L}_{w_k,\tau_k}(x_k, \pi_{\theta_k^{\mathcal{L}}}) - \nabla_x \mathcal{L}_{w_k,\tau_k}(x_k, \pi_{\theta^\star_{w_k,\tau_k}(x_k)}), x_{k+1} - x_k \rangle]$$
$$\leq \frac{C_L^2 \alpha_k \tau_k^2}{64} \mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{64 L_D^2 L_L^2 \zeta_k^2}{C_L^2 \alpha_k w_k^2 \tau_k^2} \mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2] + \frac{32 L_L^2 \zeta_k^2}{C_L^2 \alpha_k \tau_k^2} \mathbb{E}[\|\nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2].$$

We now proceed to the proof of the proposition. We define the re-weighted functions

$$\mathcal{L}^{\text{reweight}}_{w,\tau}(x, \pi) \triangleq w \mathcal{L}_{w,\tau}(x, \pi) = w f(x, \pi) + (J_\tau(x, \pi^\star_\tau(x)) - J_\tau(x, \pi)), \quad \Phi^{\text{reweight}}_{w,\tau}(x) \triangleq w \Phi_{w,\tau}(x).$$

We consider the following decomposition and bound each term on the right hand side individually.

$$\mathcal{L}^{\text{reweight}}_{w_{k+1},\tau_{k+1}}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}}) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta_k^{\mathcal{L}}})$$
$$= \Big(\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}}) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta_k^{\mathcal{L}}})\Big) + \Big(\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta_k^{\mathcal{L}}}) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta_k^{\mathcal{L}}})\Big)$$
$$+ \Big(\mathcal{L}^{\text{reweight}}_{w_{k+1},\tau_{k+1}}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}}) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}})\Big). \tag{78}$$

Bound the First Term of (78)**.** As $\mathcal{L}^{\text{reweight}}_{w,\tau}$ has L_L-Lipschitz gradients with respect to θ (shown in Lemma 9) under the condition $w, \tau \leq 1$,

$$\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}}) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta_k^{\mathcal{L}}})$$
$$\leq \langle \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta_k^{\mathcal{L}}}), \theta^{\mathcal{L}}_{k+1} - \theta_k^{\mathcal{L}} \rangle + \frac{L_L}{2} \|\theta^{\mathcal{L}}_{k+1} - \theta_k^{\mathcal{L}}\|^2$$
$$= \alpha_k \langle \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta_k^{\mathcal{L}}}), F_{w_k,\tau_k}(x_k, \theta_k^{\mathcal{L}}, \hat{V}_k^{\mathcal{L}}, \bar{s}_k, \bar{a}_k, \bar{s}'_k, \xi_k) \rangle$$
$$+ \frac{L_L \alpha_k^2}{2} \|F_{w_k,\tau_k}(x_k, \theta_k^{\mathcal{L}}, \hat{V}_k^{\mathcal{L}}, \bar{s}_k, \bar{a}_k, \bar{s}'_k, \xi_k)\|^2$$
$$= \alpha_k \langle \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta_k^{\mathcal{L}}}), F_{w_k,\tau_k}(x_k, \theta_k^{\mathcal{L}}, \hat{V}_k^{\mathcal{L}}, \bar{s}_k, \bar{a}_k, \bar{s}'_k, \xi_k) - \bar{F}_{w_k,\tau_k}(x_k, \theta_k^{\mathcal{L}}, \hat{V}_k^{\mathcal{L}}) \rangle$$
$$+ \alpha_k \langle \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta_k^{\mathcal{L}}}), \bar{F}_{w_k,\tau_k}(x_k, \theta_k^{\mathcal{L}}, \hat{V}_k^{\mathcal{L}}) - \bar{F}_{w_k,\tau_k}(x_k, \theta_k^{\mathcal{L}}, V_{\tau_k}^{x_k, \pi_{\theta_k^{\mathcal{L}}}}) \rangle$$
$$- \alpha_k \langle \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta_k^{\mathcal{L}}}), \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta_k^{\mathcal{L}}}) \rangle$$
$$+ \frac{L_L \alpha_k^2}{2} \|F_{w_k,\tau_k}(\theta_k, \omega_k, \hat{\mu}_k, \hat{V}_{f,k}, s_k, a_k, b_k, s'_k, \xi_k)\|^2, \tag{79}$$

where the final equation follows from

$$\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta_k^{\mathcal{L}}}) = w_k \nabla_\theta \mathcal{L}_{w_k,\tau_k}(x_k, \pi_{\theta_k^{\mathcal{L}}}) = -\bar{F}_{w_k,\tau_k}(x_k, \theta_k^{\mathcal{L}}, V_{\tau_k}^{x_k, \pi_{\theta_k^{\mathcal{L}}}}).$$

To bound the first term of (79),

$$
\begin{aligned}
&\alpha_k \mathbb{E}[\langle \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_k}), F_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, \hat{V}^{\mathcal{L}}_k, \bar{s}_k, \bar{a}_k, \bar{s}'_k, \xi_k) - \bar{F}_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, \hat{V}^{\mathcal{L}}_k)\rangle] \\
&= \alpha_k \mathbb{E}[\langle \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k}), \mathbb{E}[F_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, \hat{V}^{\mathcal{L}}_k, \bar{s}_k, \bar{a}_k, \bar{s}'_k, \xi_k) \mid \mathcal{F}_{k-1}] - \bar{F}_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, \hat{V}^{\mathcal{L}}_k)\rangle] \\
&\quad - \alpha_k \mathbb{E}[\langle \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k}) - \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_k}), F_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, \hat{V}^{\mathcal{L}}_k, \bar{s}_k, \bar{a}_k, \bar{s}'_k, \xi_k) - \bar{F}_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, \hat{V}^{\mathcal{L}}_k)\rangle] \\
&= -\alpha_k \mathbb{E}[\langle \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k}) - \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_k}), F_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, \hat{V}^{\mathcal{L}}_k, \bar{s}_k, \bar{a}_k, \bar{s}'_k, \xi_k) - \bar{F}_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, \hat{V}^{\mathcal{L}}_k)\rangle] \\
&\leq \alpha_k \cdot L_L \mathbb{E}[\|x_{k+1} - x_k\|] \cdot 2B_F \\
&\leq 2B_F L_L \alpha_k \cdot \frac{B_D \zeta_k}{w_k} \\
&= \frac{2B_D B_F L_L \zeta_k \alpha_k}{w_k},
\end{aligned} \tag{80}
$$

where the first inequality is due to (42) of Lemma 9, and the second inequality follows from Lemma 5.

For the second term of (79),

$$
\begin{aligned}
&\alpha_k \langle \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_k}), \bar{F}_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, \hat{V}^{\mathcal{L}}_k) - \bar{F}_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, V^{x_k,\pi_{\theta^{\mathcal{L}}_k}}_{\tau_k})\rangle \\
&\leq \frac{\alpha_k}{8} \|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_k})\|^2 + 2\alpha_k \|\bar{F}_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, \hat{V}^{\mathcal{L}}_k) - \bar{F}_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, V^{x_k,\pi_{\theta^{\mathcal{L}}_k}}_{\tau_k})\|^2 \\
&\leq \frac{\alpha_k}{4} \|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2 + \frac{\alpha_k}{4} \|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k}) - \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_k})\|^2 \\
&\quad + 2\alpha_k \|\bar{F}_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, \hat{V}^{\mathcal{L}}_k) - \bar{F}_{w_k,\tau_k}(x_k, \theta^{\mathcal{L}}_k, V^{x_k,\pi_{\theta^{\mathcal{L}}_k}}_{\tau_k})\|^2 \\
&\leq \frac{\alpha_k}{4} \|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2 + \frac{L_L^2 \alpha_k}{4} \Big(\|x_{k+1} - x_k\|^2\Big) + 2L_F^2 \alpha_k \|V^{x_k,\pi_{\theta^{\mathcal{L}}_k}}_{\tau_k} - \hat{V}^{\mathcal{L}}_k\|^2 \\
&\leq \frac{\alpha_k}{4} \|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2 + \frac{L_L^2 \alpha_k}{4} \cdot \Big(\frac{B_D \zeta_k}{w_k}\Big)^2 + 2L_F^2 \alpha_k \varepsilon^{V,\mathcal{L}}_k \\
&\leq \frac{\alpha_k}{4} \|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2 + \frac{B_D^2 L_L^2 \zeta_k^2 \alpha_k}{4w_k^2} + 2L_F^2 \alpha_k \varepsilon^{V,\mathcal{L}}_k,
\end{aligned} \tag{81}
$$

where the third inequality is again a result of the Lipschitz continuity of $\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}$ from Lemma 3 and the Lipschitz continuity of $\bar{F}_{w_k,\tau_k}$.

For the third term of (79),

$$
\begin{aligned}
&-\alpha_k \langle \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_k}), \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\rangle \\
&= -\alpha_k \|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2 + \alpha_k \langle \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k}) - \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_k}), \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\rangle \\
&\leq -\frac{\alpha_k}{2} \|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2 + \frac{\alpha_k}{2} \|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_k}) - \nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2 \\
&\leq -\frac{\alpha_k}{2} \|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2 + \frac{L_L^2 \alpha_k}{2} \Big(\|x_{k+1} - x_k\|^2\Big) \\
&\leq -\frac{\alpha_k}{2} \|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2 + \frac{L_L^2 \alpha_k}{2} \cdot \Big(\frac{B_D \zeta_k}{w_k}\Big)^2 \\
&\leq -\frac{\alpha_k}{2} \|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2 + \frac{B_D^2 L_L^2 \zeta_k^2 \alpha_k}{2w_k^2},
\end{aligned} \tag{82}
$$

where the second inequality again follows from the L_L-smoothness of $\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}$ shown in (42) of Lemma 9.

Substituting (80)-(82) into (79),

$$
\begin{aligned}
&\mathbb{E}[\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_k}) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}})] \\
&\leq \frac{2B_D B_F L_L \zeta_k \alpha_k}{w_k} + \frac{\alpha_k}{4} \mathbb{E}[\|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2] + \frac{B_D^2 L_L^2 \zeta_k^2 \alpha_k}{4w_k^2} + 2L_F^2 \alpha_k \mathbb{E}[\varepsilon^{V,\mathcal{L}}_k] \\
&\quad - \frac{\alpha_k}{2} \mathbb{E}[\|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2] + \frac{B_D^2 L_L^2 \zeta_k^2 \alpha_k}{2w_k^2} + \frac{L_L \alpha_k^2}{2} \mathbb{E}[\|F_{w_k,\tau_k}(\theta_k, \omega_k, \hat{\mu}_k, \hat{V}_{f,k}, s_k, a_k, b_k, s'_k, \xi_k)\|^2]
\end{aligned}
$$

$$
\begin{aligned}
&\leq -\frac{\alpha_k}{4}\mathbb{E}[\|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^{\mathcal{L}}_k})\|^2] + 2L_F^2\alpha_k\mathbb{E}[\varepsilon_k^{V,\mathcal{L}}] + \frac{3B_D^2L_L^2\zeta_k^2\alpha_k}{4w_k^2} + \frac{2B_DB_FL_L\zeta_k\alpha_k}{w_k} + \frac{B_F^2L_L\alpha_k^2}{2}\\
&\leq -\frac{\alpha_k}{8}\mathbb{E}[\|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^{\mathcal{L}}_k})\|^2] + 2L_F^2\alpha_k\mathbb{E}[\varepsilon_k^{V,\mathcal{L}}] + \frac{3B_D^2L_L^2\zeta_k^2\alpha_k}{4w_k^2} + \frac{2B_DB_FL_L\zeta_k\alpha_k}{w_k} + \frac{B_F^2L_L\alpha_k^2}{2}\\
&\quad - \frac{C_L^2\alpha_k\tau_k^2}{32}\mathbb{E}[\|\pi_{\theta^{\mathcal{L}}_k} - \pi^\star_{w_k,\tau_k}(x_k)\|^2],
\end{aligned}
\tag{83}
$$

where the last inequality follows from (22), which states that

$$
\|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^{\mathcal{L}}_k})\| \geq \frac{C_L\tau_k}{2}\|\pi_{\theta^{\mathcal{L}}_k} - \pi^\star_{w_k,\tau_k}(x_k)\|.
$$

Bound the Second Term of (78)**.** We use $\theta^\star_{w,\tau}(x)$ to denote a softmax parameter that encodes the policy $\pi^\star_{w,\tau}(x)$. We know from Lemma 9 that $\mathcal{L}^{\text{reweight}}_{w,\tau}$ is $\frac{L_L}{\tau}$-smooth with respect to x under $w,\tau\leq 1$,

$$
\begin{aligned}
&\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1},\pi_{\theta^{\mathcal{L}}_k}) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^{\mathcal{L}}_k})\\
&\leq \langle \nabla_x \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^{\mathcal{L}}_k}), x_{k+1}-x_k\rangle + \frac{L_L}{2\tau_k}\|x_{k+1}-x_k\|^2\\
&= \langle \nabla_x \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^{\mathcal{L}}_k}) - \nabla_x \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)}), x_{k+1}-x_k\rangle + \frac{L_L}{2\tau_k}\|x_{k+1}-x_k\|^2\\
&\quad + \langle \nabla_x \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)}), x_{k+1}-x_k\rangle\\
&= \langle \nabla_x \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^{\mathcal{L}}_k}) - \nabla_x \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)}), x_{k+1}-x_k\rangle + \frac{L_L}{2\tau_k}\|x_{k+1}-x_k\|^2\\
&\quad + \langle \nabla_x \Phi^{\text{reweight}}_{w_k,\tau_k}(x_k), x_{k+1}-x_k\rangle\\
&\leq \langle \nabla_x \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^{\mathcal{L}}_k}) - \nabla_x \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)}), x_{k+1}-x_k\rangle + \frac{L_L}{2\tau_k}\|x_{k+1}-x_k\|^2\\
&\quad + \left(\Phi^{\text{reweight}}_{w_k,\tau_k}(x_{k+1},) - \Phi^{\text{reweight}}_{w_k,\tau_k}(x_k)\right) + \frac{L_\Phi}{2\tau_k}\|x_{k+1}-x_k\|^2\\
&\leq \langle \nabla_x \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^{\mathcal{L}}_k}) - \nabla_x \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)}), x_{k+1}-x_k\rangle\\
&\quad + \left(\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1},\pi^\star_{w_k,\tau_k}(x_{k+1})) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi^\star_{w_k,\tau_k}(x_k))\right) + \frac{L_\Phi}{\tau_k}\|x_{k+1}-x_k\|^2,
\end{aligned}
$$

where the last inequality is due to $L_L \leq L_\Phi$, and the second equation uses the relationship $\nabla_x \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)}) = \nabla_x \Phi^{\text{reweight}}_{w_k,\tau_k}(x_k)$, which is due to $\nabla_\theta \mathcal{L}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)}) = 0$ by the first-order optimality condition. The second inequality is due to the fact that $\Phi_{w,\tau}$ is $\frac{L_\Phi}{w\tau}$-smooth when $w,\tau\leq 1$ (established in Lemma 10) and that for an L-smooth function f, we have

$$
-f(y) + f(x) \leq \langle -\nabla f(x), y-x\rangle + \frac{L}{2}\|x-y\|^2.
$$

Taking the expectation and plugging in the result from Lemma 14,

$$
\begin{aligned}
&\mathbb{E}[\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1},\pi_{\theta^{\mathcal{L}}_k}) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^{\mathcal{L}}_k})]\\
&\leq \frac{C_L^2\alpha_k\tau_k^2}{64}\mathbb{E}[\|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|^2]\\
&\quad + \frac{64L_D^2L_L^2\zeta_k^2}{C_L^2\alpha_k w_k^2\tau_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|^2] + \frac{32L_L^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]\\
&\quad + \mathbb{E}[\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1},\pi^\star_{w_k,\tau_k}(x_{k+1})) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi^\star_{w_k,\tau_k}(x_k))] + \frac{L_\Phi}{\tau_k}\mathbb{E}[\|x_{k+1}-x_k\|^2]\\
&\leq \frac{C_L^2\alpha_k\tau_k^2}{64}\mathbb{E}[\|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|^2] + \frac{64L_D^2L_L^2\zeta_k^2}{C_L^2\alpha_k w_k^2\tau_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|^2]\\
&\quad + \mathbb{E}[\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1},\pi^\star_{w_k,\tau_k}(x_{k+1})) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi^\star_{w_k,\tau_k}(x_k))] + \frac{32L_L^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2] + \frac{B_D^2L_\Phi\zeta_k^2}{w_k^2\tau_k},
\end{aligned}
\tag{84}
$$

where the last inequality follows from $\|x_{k+1} - x_k\| \leq \frac{B_D \zeta_k}{w_k}$.

Bound the Third Term of (78)**.** By the definition of $\mathcal{L}_{w,\tau}$ in (11),

$$
\begin{aligned}
&\mathcal{L}^{\text{reweight}}_{w_{k+1},\tau_{k+1}}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}}) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}}) \\
&= (w_{k+1} - w_k) f(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}}) \\
&\qquad + \Big(J_{\tau_{k+1}}(x_{k+1}, \pi^\star_{\tau_{k+1}}(x_{k+1})) - J_{\tau_k}(x_{k+1}, \pi^\star_{\tau_k}(x_{k+1})) \Big) - \Big(J_{\tau_{k+1}}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}}) - J_{\tau_k}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}}) \Big) \\
&\leq (w_{k+1} - w_k) f(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}}) + \frac{\log|\mathcal{A}|(\tau_k - \tau_{k+1})}{1-\gamma} + \frac{\tau_k - \tau_{k+1}}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{\pi^{\mathcal{L}}_{k+1}}}[E(\pi^{\mathcal{L}}_{k+1}, s)] \\
&\leq 0 + \frac{\log|\mathcal{A}|(\tau_k - \tau_{k+1})}{1-\gamma} + \frac{\tau_k - \tau_{k+1}}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{\pi^{\mathcal{L}}_{k+1}}}[E(\pi^{\mathcal{L}}_{k+1}, s)] \\
&\leq \frac{16 \log|\mathcal{A}| \tau_k}{3(1-\gamma)(k+1)},
\end{aligned} \tag{85}
$$

where the first inequality follows from Zeng et al. [2022][Lemma 3], the second inequality is due to the fact that f is non-negative from Assumption 3, and the third inequality follows from Lemma 2 and the relationship $E(\pi, s) \leq \log|\mathcal{A}|$ for any policy π.

Collecting the bounds in (83)-(85) and plugging them into (78), we get

$$
\begin{aligned}
&\mathbb{E}[\mathcal{L}^{\text{reweight}}_{w_{k+1},\tau_{k+1}}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}}) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})] \\
&\leq -\frac{\alpha_k}{8} \mathbb{E}[\|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2] + 2L_F^2 \alpha_k \mathbb{E}[\varepsilon_k^{V,\mathcal{L}}] + \frac{3B_D^2 L_L^2 \zeta_k^2 \alpha_k}{4w_k^2} + \frac{2B_D B_F L_L \zeta_k \alpha_k}{w_k} + \frac{B_F^2 L_L \alpha_k^2}{2} \\
&\quad - \frac{C_L^2 \alpha_k \tau_k^2}{32} \mathbb{E}[\|\pi_{\theta^{\mathcal{L}}_k} - \pi^\star_{w_k,\tau_k}(x_k)\|^2] \\
&\quad + \frac{C_L^2 \alpha_k \tau_k^2}{64} \mathbb{E}[\|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|^2] + \frac{64 L_D^2 L_L^2 \zeta_k^2}{C_L^2 \alpha_k w_k^2 \tau_k^2} \mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|^2] \\
&\quad + \mathbb{E}[\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi^\star_{w_k,\tau_k}(x_{k+1})) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi^\star_{w_k,\tau_k}(x_k))] + \frac{32 L_L^2 \zeta_k^2}{C_L^2 \alpha_k \tau_k^2} \mathbb{E}[\|\nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2] + \frac{B_D^2 L_\Phi \zeta_k^2}{w_k^2 \tau_k} \\
&\quad + \frac{16 \log|\mathcal{A}| \tau_k}{3(1-\gamma)(k+1)} \\
&\leq -\frac{\alpha_k}{8} \mathbb{E}[\|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2] + 2L_F^2 \alpha_k \mathbb{E}[\varepsilon_k^{V,\mathcal{L}}] + \frac{32 L_L^2 \zeta_k^2}{C_L^2 \alpha_k \tau_k^2} \mathbb{E}[\|\nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2] \\
&\quad - \frac{C_L^2 \alpha_k \tau_k^2}{64} \mathbb{E}[\|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|^2] + \frac{64 L_D^2 L_L^2 \zeta_k^2}{C_L^2 \alpha_k w_k^2 \tau_k^2} \mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi^{\mathcal{L}}_k - \pi^\star_{w_k,\tau_k}(x_k)\|^2] \\
&\quad + \frac{2B_D^2 L_\Phi \zeta_k^2}{w_k^2 \tau_k} + \frac{2B_D B_F L_L \zeta_k \alpha_k}{w_k} + \frac{B_F^2 L_L \alpha_k^2}{2} + \frac{16 \log|\mathcal{A}| \tau_k}{3(1-\gamma)(k+1)} \\
&\quad + \mathbb{E}[\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi^\star_{w_k,\tau_k}(x_{k+1})) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi^\star_{w_k,\tau_k}(x_k))],
\end{aligned}
$$

where in the last inequality we have combined the terms $\frac{3B_D^2 L_L^2 \zeta_k^2 \alpha_k}{4w_k^2}$ and $\frac{B_D^2 L_\Phi \zeta_k^2}{w_k^2 \tau_k}$ under the step size conditions $\tau_k \leq 1$ and $\alpha_k \leq \frac{4L_\Phi}{3L_L^2}$.

Recall the definition of $\varepsilon_k^{\theta,\mathcal{L}}$ in (33). We can re-arrange the terms in the inequality above and obtain

$$
\begin{aligned}
&\mathbb{E}[\varepsilon_{k+1}^{\theta,\mathcal{L}} - \varepsilon_k^{\theta,\mathcal{L}}] \\
&= \mathbb{E}[\mathcal{L}^{\text{reweight}}_{w_{k+1},\tau_{k+1}}(x_{k+1}, \pi_{\theta^{\mathcal{L}}_{k+1}}) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})] \\
&\qquad - \mathbb{E}[\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi^\star_{w_k,\tau_k}(x_{k+1})) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi^\star_{w_k,\tau_k}(x_k))] \\
&\qquad - \mathbb{E}[\mathcal{L}^{\text{reweight}}_{w_{k+1},\tau_{k+1}}(x_{k+1}, \pi^\star_{w_k,\tau_k}(x_{k+1})) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi^\star_{w_k,\tau_k}(x_{k+1}))] \\
&\leq -\frac{\alpha_k}{8} \mathbb{E}[\|\nabla_\theta \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k, \pi_{\theta^{\mathcal{L}}_k})\|^2] + 2L_F^2 \alpha_k \mathbb{E}[\varepsilon_k^{V,\mathcal{L}}] + \frac{32 L_L^2 \zeta_k^2}{C_L^2 \alpha_k \tau_k^2} \mathbb{E}[\|\nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2]
\end{aligned}
$$

$$- \frac{C_L^2 \alpha_k \tau_k^2}{64} \mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2] + \frac{64 L_D^2 L_L^2 \zeta_k^2}{C_L^2 \alpha_k w_k^2 \tau_k^2} \mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{2B_D^2 L_\Phi \zeta_k^2}{w_k^2 \tau_k} + \frac{2B_D B_F L_L \zeta_k \alpha_k}{w_k} + \frac{B_F^2 L_L \alpha_k^2}{2} + \frac{32 \log |\mathcal{A}| \tau_k}{3(1-\gamma)(k+1)},$$

where the bound on $\mathcal{L}^{\text{reweight}}_{w_{k+1},\tau_{k+1}}(x_{k+1}, \pi^\star_{w_k,\tau_k}(x_{k+1})) - \mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_{k+1}, \pi^\star_{w_k,\tau_k}(x_{k+1})) \leq \frac{16 \log |\mathcal{A}| \tau_k}{3(1-\gamma)(k+1)}$ can be obtained in a manner similar to (85).

■

D.5 Proof of Proposition 5

We first establish the convergence of ε_k^V. To this end, we introduce the following technical lemma, which bounds an important cross term.

Lemma 15 *Under the assumptions and step sizes of Proposition 5, we have for all $k \geq 0$*

$$\mathbb{E}[\langle \hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}} + \beta_k \bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k), V_{\tau_k}^{x_k,\pi_{\theta_k}} - V_{\tau_{k+1}}^{x_{k+1},\pi_{\theta_{k+1}}} \rangle]$$
$$\leq \frac{(1-\gamma)\beta_k}{2} \mathbb{E}[\|\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}} + \beta_k \bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k)\|^2] + \frac{6L_V^2 \zeta_k^2}{(1-\gamma)\beta_k} \mathbb{E}[\|\nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{12 L_V^2 L_D^2 \zeta_k^2}{(1-\gamma)\beta_k} \mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] + \frac{12 L_V^2 L_D^2 \zeta_k^2}{(1-\gamma)\beta_k} \mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]$$
$$+ \frac{6L_V^2 \alpha_k^2}{(1-\gamma)\beta_k} \mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2] + \frac{6L_V^2 L_F^2 \alpha_k^2}{(1-\gamma)\beta_k} \mathbb{E}[\varepsilon_k^V] + \frac{6B_F^2 L_V \tau_0 \alpha_k^2}{\alpha_0} + \frac{32 L_V^2 \tau_k^2}{3(1-\gamma)\beta_k (k+1)^2}.$$

We defer the proof of the lemma to Appendix E.15.

By the update rule in (49), we have

$$\begin{aligned}
\varepsilon_{k+1}^V &= \|\hat{V}_{k+1} - V_{\tau_{k+1}}^{x_{k+1},\pi_{\theta_{k+1}}}\|^2 \\
&= \|\Pi_{B_V}\Big(\hat{V}_k + \beta_k G_{\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s_k')\Big) - V_{\tau_{k+1}}^{x_{k+1},\pi_{\theta_{k+1}}}\|^2 \\
&\leq \Big\|\hat{V}_k + \beta_k G_{\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s_k') - V_{\tau_{k+1}}^{x_{k+1},\pi_{\theta_{k+1}}}\Big\|^2 \\
&= \Big\|\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}} + \beta_k \bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k) + \beta_k\Big(G_f(x_k, \theta_k, \hat{V}_k, s_k, a_k, s_k') - \bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k)\Big) \\
&\quad + V_{\tau_k}^{x_k,\pi_{\theta_k}} - V_{\tau_{k+1}}^{x_{k+1},\pi_{\theta_{k+1}}}\Big\|^2 \\
&\leq \Big\|\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}} + \beta_k \bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k)\Big\|^2 \\
&\quad + 2\beta_k^2 \|G_{\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s_k') - \bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k)\|^2 \\
&\quad + 2\Big\|V_{\tau_k}^{x_k,\pi_{\theta_k}} - V_{\tau_{k+1}}^{x_{k+1},\pi_{\theta_{k+1}}}\Big\|^2 \\
&\quad + \beta_k \langle \hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}} + \beta_k \bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k), G_{\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s_k') - \bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k)\rangle \\
&\quad + \langle \hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}} + \beta_k \bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k), V_{\tau_k}^{x_k,\pi_{\theta_k}} - V_{\tau_{k+1}}^{x_{k+1},\pi_{\theta_{k+1}}}\rangle,
\end{aligned} \tag{86}$$

where the first inequality follows from the fact that the projection to a convex set is non-expansive.

To bound the first term of (86),

$$\begin{aligned}
&\Big\|\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}} + \beta_k \bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k)\Big\|^2 \\
&= \|\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}}\|^2 + \beta_k^2 \|\bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k)\|^2 + 2\beta_k \langle \hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}}, \bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k)\rangle \\
&= \|\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}}\|^2 + \beta_k^2 \|\bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k) - \bar{G}_{\tau_k}(x_k, \theta_k, V_{\tau_k}^{x_k,\pi_{\theta_k}})\|^2 \\
&\quad + 2\beta_k \langle \hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}}, \bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k) - \bar{G}_{\tau_k}(x_k, \theta_k, V_{\tau_k}^{x_k,\pi_{\theta_k}})\rangle
\end{aligned}$$

$$
\begin{aligned}
&= \|\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}}\|^2 + \beta_k^2\|\bar{G}_{\tau_k}(x_k,\theta_k,\hat{V}_k) - \bar{G}_{\tau_k}(x_k,\theta_k,V_{\tau_k}^{x_k,\pi_{\theta_k}})\|^2 \\
&\quad + 2\beta_k\Big(\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}}\Big)^\top(\gamma P^{\pi_{\theta_k}} - I)\Big(\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}}\Big) \\
&\le \|\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}}\|^2 + L_G^2\beta_k^2\|\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}}\|^2 + 2(\gamma-1)\beta_k\|\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}}\|^2 \\
&\le (1-(1-\gamma)\beta_k)\varepsilon_k^V,
\end{aligned} \tag{87}
$$

where the second equation uses the relationship $\bar{G}_\tau(x,\theta,V_\tau^{x,\pi_\theta}) = 0$ for any τ, x, θ, and the last inequality follows from the step size condition $\beta_k \le \frac{1-\gamma}{L_G^2}$.

For the third term of (86),

$$
\begin{aligned}
&2\|V_{\tau_k}^{x_k,\pi_{\theta_k}} - V_{\tau_{k+1}}^{x_{k+1},\pi_{\theta_{k+1}}}\|^2 \\
&\le 4\|V_{\tau_k}^{x_{k+1},\pi_{\theta_{k+1}}} - V_{\tau_{k+1}}^{x_{k+1},\pi_{\theta_{k+1}}}\|^2 + 4\|V_{\tau_k}^{x_k,\pi_{\theta_k}} - V_{\tau_k}^{x_{k+1},\pi_{\theta_{k+1}}}\|^2 \\
&\le 4|\mathcal{S}|\|V_{\tau_k}^{x_{k+1},\pi_{\theta_{k+1}}} - V_{\tau_{k+1}}^{x_{k+1},\pi_{\theta_{k+1}}}\|_\infty^2 + 8L_V^2\Big(\|x_k - x_{k+1}\|^2 + \|\pi_{\theta_{k+1}} - \pi_{\theta_k}\|^2\Big) \\
&\le 4|\mathcal{S}|\cdot\Big(\frac{\tau_k - \tau_{k+1}}{1-\gamma}\log|\mathcal{A}|\Big)^2 + 8L_V^2\Big(\frac{B_D^2\zeta_k^2}{w_k^2} + B_F^2\alpha_k^2\Big),
\end{aligned}
$$

where the second inequality follows from the Lipschitz continuity of the value function established in Lemma 3, and the third inequality applies Zeng et al. [2022][Lemma 3]. Plugging in the bound on $\tau_k - \tau_{k+1}$ from Lemma 2, we get

$$
\begin{aligned}
2\|V_{\tau_k}^{x_k,\pi_{\theta_k}} - V_{\tau_{k+1}}^{x_{k+1},\pi_{\theta_{k+1}}}\|^2 &\le 4|\mathcal{S}|\cdot\Big(\frac{8\tau_k}{3(1-\gamma)(k+1)}\log|\mathcal{A}|\Big)^2 + 8L_V^2\Big(\frac{B_D^2\zeta_k^2}{w_k^2} + B_F^2\alpha_k^2\Big) \\
&\le \frac{256|\mathcal{S}|\log^2|\mathcal{A}|\tau_k^2}{3(1-\gamma)^2(k+1)^2} + 16B_F^2L_V^2\alpha_k^2,
\end{aligned} \tag{88}
$$

where we use the step size condition $\zeta_k \le \frac{B_F\alpha_k w_k}{B_D}$.

For the fourth term of (86), we have in expectation

$$
\begin{aligned}
&\mathbb{E}[\langle \hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}} + \beta_k\bar{G}_{\tau_k}(x_k,\theta_k,\hat{V}_k), G_{\tau_k}(x_k,\theta_k,\hat{V}_k,s_k,a_k,s_k') - \bar{G}_{\tau_k}(x_k,\theta_k,\hat{V}_k)\rangle] \\
&= \mathbb{E}[\langle \hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}} + \beta_k\bar{G}_{\tau_k}(x_k,\theta_k,\hat{V}_k), \mathbb{E}[G_{\tau_k}(x_k,\theta_k,\hat{V}_k,s_k,a_k,s_k') - \bar{G}_{\tau_k}(x_k,\theta_k,\hat{V}_k) \mid \mathcal{F}_{k-1}]\rangle] \\
&= 0.
\end{aligned} \tag{89}
$$

Collecting the bounds from (87)-(89) and Lemma 15,

$$
\begin{aligned}
&\mathbb{E}[\varepsilon_{k+1}^V] \\
&= (1-(1-\gamma)\beta_k)\mathbb{E}[\varepsilon_k^V] + 2\beta_k^2\mathbb{E}[\|G_{\tau_k}(x_k,\theta_k,\hat{V}_k,s_k,a_k,s_k') - \bar{G}_{\tau_k}(x_k,\theta_k,\hat{V}_k)\|^2] \\
&\quad + \frac{256|\mathcal{S}|\log^2|\mathcal{A}|\tau_k^2}{3(1-\gamma)^2(k+1)^2} + 16B_F^2L_V^2\alpha_k^2 \\
&\quad + \frac{(1-\gamma)\beta_k}{2}\mathbb{E}[\|\hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}} + \beta_k\bar{G}_{\tau_k}(x_k,\theta_k,\hat{V}_k)\|^2] + \frac{6L_V^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2] \\
&\quad + \frac{12L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k - \pi_{\tau_k}^\star(x_k)\|^2] + \frac{12L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi_{w_k,\tau_k}^\star(x_k)\|^2] \\
&\quad + \frac{6L_V^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2] + \frac{6L_V^2L_F^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\varepsilon_k^V] + \frac{6B_F^2L_V\tau_0\alpha_k^2}{\alpha_0} + \frac{32L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2} \\
&\le \Big(1-(1-\gamma)\beta_k\Big)\mathbb{E}[\varepsilon_k^V] + \Big(1-(1-\gamma)\beta_k\Big)\frac{(1-\gamma)\beta_k}{2}\mathbb{E}[\varepsilon_k^V] + \frac{6L_V^2L_F^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\varepsilon_k^V] + 8B_G\beta_k^2 \\
&\quad + \frac{256|\mathcal{S}|\log^2|\mathcal{A}|\tau_k^2}{3(1-\gamma)^2(k+1)^2} + 16B_F^2L_V^2\alpha_k^2 + \frac{6L_V^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2] \\
&\quad + \frac{12L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k - \pi_{\tau_k}^\star(x_k)\|^2] + \frac{12L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi_{w_k,\tau_k}^\star(x_k)\|^2]
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{6L_V^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2] + \frac{6B_F^2 L_V\tau_0\alpha_k^2}{\alpha_0} + \frac{32L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2} \\
\leq &\Big(1-\frac{(1-\gamma)\beta_k}{4}\Big)\mathbb{E}[\varepsilon_k^V] + \frac{12L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k-\pi^\star_{\tau_k}(x_k)\|^2] + \frac{12L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k^{\mathcal{L}}-\pi^\star_{w_k,\tau_k}(x_k)\|^2] \\
&+ \frac{6L_V^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2] + \frac{6L_V^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2] \\
&+ \frac{22B_F^2L_V\tau_0\alpha_k^2}{\alpha_0} + \frac{64L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2} + 8B_G\beta_k^2,
\end{aligned}
$$

where the second inequality simplifies and combines terms based on the step size conditions $\frac{\alpha_k}{\beta_k} \leq \frac{1-\gamma}{2\sqrt{6}L_VL_F}$, $\beta_k \leq \frac{(1-\gamma)L_V^2}{8|\mathcal{S}|\log^2|\mathcal{A}|}$, and $\frac{\alpha_0}{\tau_0} \leq \frac{1}{L_V}$.

The bound on $\mathbb{E}[\varepsilon_{k+1}^{V,\mathcal{L}}]$ can be derived using an identical argument.

■

E Proof of Supporting Results

E.1 Proof of Lemma 1

Uniqueness of $\pi^\star(x)$. We first take the approach of proof by contradiction to show that $\pi^\star(x)$ is unique for any x.

Given x, suppose that there exist two distinct optimal solutions of (5), which we denote by π_1, π_2. From the definition of $\Pi^\star(x)$, we know that π_1, π_2 satisfy

$$
J(x,\pi_1) \geq J(x,\pi), \quad J(x,\pi_2) \geq J(x,\pi), \quad \forall \pi. \tag{90}
$$

We construct another policy π' as follows, inspired by the proof of Theorem 1 in Zeng et al. [2023]. For all state s and action a, $\pi'(a \mid s)$ is expressed as

$$
\pi'(a \mid s) = \frac{d_\rho^{\pi_1}(s)\pi_1(a \mid s) + d_\rho^{\pi_2}(s)\pi_2(a \mid s)}{d_\rho^{\pi_1}(s) + d_\rho^{\pi_2}(s)}. \tag{91}
$$

Note that Assumption 1 guarantees $d_\rho^\pi(s) \geq (1-\gamma)\rho_{\min} > 0$ for all s, ensuring π' is well-defined.

Our first step is to show that π' is also an optimal policy, i.e., $\pi' \in \Pi^\star(x)$. To see this, we define a modified transition kernel $\mathcal{P}_\gamma$ such that $\mathcal{P}_\gamma(s' \mid s,a) = (1-\gamma)\rho(s') + \gamma\mathcal{P}(s' \mid s,a)$ for any π, and P_γ^π such that $P_\gamma^\pi(s' \mid s) = (1-\gamma)\rho(s') + \gamma P^\pi(s' \mid s)$ for any π.

We also define a vector $B \in \mathbb{R}^{|\mathcal{S}|}$

$$
B = \mathcal{P}_\gamma^{\pi'} \cdot \left(\frac{1}{2}d_\rho^{\pi_1} + \frac{1}{2}d_\rho^{\pi_2}\right).
$$

We can express each entry of B in the following way

$$
\begin{aligned}
B(s') &= \sum_{s,a}\mathcal{P}_\gamma(s' \mid s,a)\pi'(a \mid s)\left(\frac{1}{2}d_\rho^{\pi_1}(s) + \frac{1}{2}d_\rho^{\pi_2}(s)\right) \\
&= \sum_{s,a}\mathcal{P}_\gamma(s' \mid s,a)\frac{d_\rho^{\pi_1}(s)\pi_1(a \mid s) + d_\rho^{\pi_2}(s)\pi_2(a \mid s)}{d_\rho^{\pi_1}(s) + d_\rho^{\pi_2}(s)}\left(\frac{1}{2}d_\rho^{\pi_1}(s) + \frac{1}{2}d_\rho^{\pi_2}(s)\right) \\
&= \frac{1}{2}\sum_{s,a}\mathcal{P}_\gamma(s' \mid s,a)d_\rho^{\pi_1}(s)\pi_1(a \mid s) + \frac{1}{2}\sum_{s,a}\mathcal{P}(s' \mid s,a)d_\rho^{\pi_2}(s)\pi_2(a \mid s) \\
&= \frac{1}{2}\sum_{s,a}P_\gamma^{\pi_1}(s' \mid s)d_\rho^{\pi_1}(s) + \frac{1}{2}\sum_{s,a}P_\gamma^{\pi_2}(s' \mid s)d_\rho^{\pi_2}(s)
\end{aligned}
$$

$$= \frac{1}{2} d_\rho^{\pi_1}(s') + \frac{1}{2} d_\rho^{\pi_2}(s'),$$

which leads to

$$P_\gamma^{\pi'} \cdot \left(\frac{1}{2} d_\rho^{\pi_1} + \frac{1}{2} d_\rho^{\pi_2} \right) = \frac{1}{2} d_\rho^{\pi_1} + \frac{1}{2} d_\rho^{\pi_2}. \tag{92}$$

The Markov chain induced by P_γ^π is always ergodic, assumed in Assumption 1. Under ergodicity, it is known that d_ρ^π (properly normalized) is the unique eigenvector of P_γ^π associated with eigenvalue 1. Therefore, (92) implies that $\frac{1}{2} d_\rho^{\pi_1} + \frac{1}{2} d_\rho^{\pi_2}$ is the discounted visitation distribution induced by policy π', i.e.,

$$d_\rho^{\pi'} = \frac{1}{2} d_\rho^{\pi_1} + \frac{1}{2} d_\rho^{\pi_2}. \tag{93}$$

We use $\hat{d}_\rho^\pi$ to denote the extend discounted visitation distribution over state and action such that

$$\hat{d}_\rho^\pi(s, a) = d_\rho^\pi(s) \pi(a \mid s).$$

We have from (93)

$$\begin{aligned}
\hat{d}_\rho^{\pi'}(s, a) &= d_\rho^{\pi'}(s) \pi'(a \mid s) \\
&= \left(\frac{1}{2} d_\rho^{\pi_1}(s) + \frac{1}{2} d_\rho^{\pi_2}(s) \right) \frac{d_\rho^{\pi_1}(s) \pi_1(a \mid s) + d_\rho^{\pi_2}(s) \pi_2(a \mid s)}{d_\rho^{\pi_1}(s) + d_\rho^{\pi_2} \mu_{\pi_2}(s)} \\
&= \frac{1}{2} d_\rho^{\pi_1}(s) \pi_1(a \mid s) + \frac{1}{2} d_\rho^{\pi_1}(s) \pi_2(a \mid s) \\
&= \frac{1}{2} \hat{d}_\rho^{\pi_1}(s, a) + \frac{1}{2} \hat{d}_\rho^{\pi_2}(s, a).
\end{aligned}$$

Note that the cumulative return J is linear in the space of extended discounted visitation distribution. We have

$$J(x, \pi') = \langle r_x, \hat{d}_\rho^{\pi'} \rangle = \frac{1}{2} \langle r_x, \frac{1}{2} \hat{d}_\rho^{\pi_1}(s, a) + \frac{1}{2} \hat{d}_\rho^{\pi_2}(s, a) \rangle = \frac{1}{2} J(x, \pi_1) + \frac{1}{2} J(x, \pi_2).$$

In view of (90), this implies that π' is an optimal policy, i.e., $\pi' \in \Pi^\star(x)$.

Since π' is in the constraint set for the optimization problem in (5), we can create a contradiction that π_1, π_2 are the two distinct maximizers of (5) if π' has a larger weighted entropy. The entropy function $E(;s)$ is strictly concave for all state s for policies in the interior of the simplex (note that π_1, π_2 must in the interior of the simplex, as they cannot be the optimal solution of (5) otherwise). Recall from (91) that $\pi'(\cdot \mid s)$ is a convex combination of $\pi_1(\cdot \mid s), \pi_2(\cdot \mid s)$, and by the property of strictly concave functions,

$$E(\pi', s) > \frac{d_\rho^{\pi_1}(s) E(\pi_1, s) + d_\rho^{\pi_2}(s) E(\pi_2, s)}{d_\rho^{\pi_1}(s) + d_\rho^{\pi_2}(s)}.$$

We denote $\mathcal{E}_\pi = \mathbb{E}_{s \sim d_\rho^\pi}[E(\pi, s)]$.

$$\begin{aligned}
\mathcal{E}_{\pi'} &= \langle d_\rho^{\pi'}, E(\pi', \cdot) \rangle \\
&= \sum_s \big(\frac{1}{2} d_\rho^{\pi_1}(s) + \frac{1}{2} d_\rho^{\pi_2}(s) \big) E(\pi', s) \\
&> \sum_s \big(\frac{1}{2} d_\rho^{\pi_1}(s) + \frac{1}{2} d_\rho^{\pi_2}(s) \big) \frac{d_\rho^{\pi_1}(s) E(\pi_1, s) + d_\rho^{\pi_2}(s) E(\pi_2, s)}{d_\rho^{\pi_1}(s) + d_\rho^{\pi_2}(s)} \\
&= \frac{1}{2} \langle d_\rho^{\pi_1}, E(\pi_1, s) \rangle + \frac{1}{2} \langle d_\rho^{\pi_2}, E(\pi_2, s) \rangle \\
&= \frac{1}{2} \mathcal{E}_{\pi_1} + \frac{1}{2} \mathcal{E}_{\pi_2}.
\end{aligned}$$

This contradicts the condition that π_1, π_2 maximize the weighted entropy within the constraint set $\Pi^\star(x)$ and concludes our proof on the uniqueness of $\pi^\star(x)$.

Limit Point of $\{\pi_\tau^\star(x)\}_\tau$. We then show the limit point of $\{\pi_\tau^\star(x)\}_\tau$ is $\pi^\star(x)$ as $\tau \to 0$. As x is fixed here, we simply denote $\pi_\tau = \pi_\tau^\star(x)$. We define for simplicity $\bar{E}(\pi) \triangleq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^\pi}[E(\pi, s)]$. By the Bolzano–Weierstrass theorem, as the sequence $\{\pi_\tau\}$ is bounded, it has a convergent subsequence. Let $\tau_n \to 0$ and $\pi_{\tau_n} \to \bar{\pi}$. We first need to show $\bar{\pi} \in \Pi^\star(x)$. By the definition of $\pi_\tau^\star(x)$,

$$J(x, \pi_\tau) + \tau\bar{E}(\pi_\tau) \geq J(x, \pi^\star(x)) + \tau\bar{E}(\pi^\star(x)), \tag{94}$$

leading to

$$J(x, \pi_\tau) \geq J(x, \pi^\star(x)) + \tau\Big(\bar{E}(\pi^\star(x)) - \bar{E}(\pi_\tau)\Big).$$

As J is continuous, we take $n \to \infty$

$$\limsup J(x, \pi_\tau) \geq J(x, \pi^\star(x)),$$

implying $J(x, \bar{\pi})$. This means $\bar{\pi} \in \Pi^\star$.

Then, to show $\bar{\pi}$ is the maximum entropy solution, we re-arrange the terms in (94)

$$\bar{E}(\pi_\tau) - \bar{E}(\pi^\star(x)) \geq \frac{J(x, \pi^\star(x)) - J(x, \pi_\tau)}{\tau} \geq 0,$$

where the second inequality is due to the definition of $\pi^\star(x)$.

Taking the limit,

$$\limsup \bar{E}(\pi_\tau) \geq \bar{E}(\pi^\star(x)).$$

As the limit point $\bar{\pi}$ is in $\Pi^\star$ and we have $\pi^\star(x) = \operatorname{argmax}_{\pi\in\Pi^\star} \bar{E}(\pi)$, then it holds that $\limsup \bar{E}(\pi_\tau) = \bar{E}(\pi^\star(x)) = \bar{E}(\bar{\pi})$. This allows us to conclude that $\pi_\tau^\star(x) \to \pi^\star(x)$ as $\tau \to 0$.

■

E.2 Proof of Lemma 2

We apply Zeng et al. [2022][Lemma 7]

$$\tau_k - \tau_{k+1} = \frac{\tau_0}{(k+1)^{c_\tau}} - \frac{\tau_0}{(k+2)^{c_\tau}} \leq \frac{8\tau_0}{3(k+1)^{c_\tau+1}} = \frac{8\tau_k}{3(k+1)}.$$

■

E.3 Proof of Lemma 3

We derive the inequalities on the value function V_τ^{x,π_θ} and note that the ones on the cumulative return $J_\tau(x, \pi_\theta)$ immediately follows as the cumulative return is simply an weighted average of the value function

$$J_\tau(x, \pi_\theta) = \mathbb{E}_{s\sim\rho}[V_\tau^{x,\pi_\theta}(s)].$$

Fixing x, we know that $V_\tau^{x,\pi}$ is the standard policy optimization objective (in a standard MDP) as a function of policy π. It is well-known that the following inequalities hold (for example, see Lemma B.5 of Zeng et al. [2021] and Lemma 5 of Zeng et al. [2022])

$$\|V^{x,\pi_\theta} - V^{x,\pi_{\theta'}}\| \leq \frac{2}{(1-\gamma)^2}\|\theta - \theta'\|, \tag{95}$$

$$\|\nabla_\theta V^{x,\pi_\theta} - \nabla_\theta V^{x,\pi_{\theta'}}\| \leq \frac{8}{(1-\gamma)^3}\|\theta - \theta'\|. \tag{96}$$

We define the shorthand notation

$$\mathcal{H}(\theta, s) \triangleq \mathbb{E}_{a_k \sim \pi(\cdot|s_k), s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)} \left[\sum_{k=0}^{\infty} -\gamma^k \log \pi_\theta(a_k \mid s_k) \mid s_0 = s \right].$$

This implies $V_\tau^{x,\pi_\theta}(s) = V^{x,\pi_\theta}(s) + \tau \mathcal{H}(\theta, s)$. We also define the aggregate notation

$$\mathcal{H}(\theta) = [\mathcal{H}(\theta, s_1), \mathcal{H}(\theta, s_2), \cdots] \in \mathbb{R}^{|\mathcal{S}|}.$$

Adapting Lemma 6 of Zeng et al. [2022], we have for all $s \in \mathcal{S}$

$$\|\mathcal{H}(\theta, s) - \mathcal{H}(\theta', s)\| \leq \frac{4 + 8 \log |\mathcal{A}|}{(1-\gamma)^3} \|\theta - \theta'\|, \tag{97}$$

$$\|\nabla_\theta \mathcal{H}(\theta, s) - \nabla_\theta \mathcal{H}(\theta', s)\| \leq \frac{4 + 8 \log |\mathcal{A}|}{(1-\gamma)^3} \|\theta - \theta'\|. \tag{98}$$

We obviously have the following inequalities from (97) and (98)

$$\|\mathcal{H}(\theta) - \mathcal{H}(\theta')\| \leq \frac{(4 + 8 \log |\mathcal{A}|)\sqrt{|\mathcal{S}|}}{(1-\gamma)^3} \|\theta - \theta'\|, \tag{99}$$

$$\|\nabla_\theta \mathcal{H}(\theta) - \nabla_\theta \mathcal{H}(\theta')\| \leq \frac{(4 + 8 \log |\mathcal{A}|)\sqrt{|\mathcal{S}|}}{(1-\gamma)^3} \|\theta - \theta'\|. \tag{100}$$

Note that (98) also implies

$$\|\nabla_{\theta,\theta} \mathbb{E}_{s \sim d_\rho^{\pi_\theta}} [E(\pi_\theta, s)]\| \leq \frac{4 + 8 \log |\mathcal{A}|}{(1-\gamma)^3}$$

hence leading to (39).

In addition, we have from (2)

$$\begin{aligned} |V_\tau^{x,\pi_\theta} - V_\tau^{x',\pi_\theta}| &= |\mathbb{E}_{s' \sim d_s^\pi,\, a' \sim \pi(\cdot|s')}[r_x(s,a) - r_{x'}(s,a)]| \\ &\leq \mathbb{E}_{s' \sim d_s^\pi,\, a' \sim \pi(\cdot|s')}[|r_x(s,a) - r_{x'}(s,a)|] \\ &\leq \mathbb{E}_{s' \sim d_s^\pi,\, a' \sim \pi(\cdot|s')}[L_r \|x - x'\|] \\ &\leq L_r \|x - x'\|, \end{aligned} \tag{101}$$

where the first inequality is a result of Jensen's inequality and the second inequality is due to Assumption 4.

We can express $\nabla_\theta V^{x,\pi_\theta}$ as follows [Agarwal et al., 2021]

$$\nabla_{\theta_{s',a'}} V^{x,\pi_\theta}(s) = \frac{1}{1-\gamma} d_s^{\pi_\theta}(s') \pi_\theta(a' \mid s') A^{x,\pi_\theta}(s', a').$$

This implies

$$\begin{aligned} |\nabla_{\theta_{s',a'}} V^{x,\pi_\theta}(s) - \nabla_{\theta_{s',a'}} V^{x',\pi_\theta}(s)| &= \frac{1}{1-\gamma} d_s^{\pi_\theta}(s') \pi_\theta(a' \mid s') |A^{x,\pi_\theta}(s', a') - A^{x',\pi_\theta}(s', a')| \\ &\leq \frac{1}{1-\gamma} |d_s^{\pi_\theta}(s')| |\pi_\theta(a' \mid s')| \cdot 2L_r \|x - x'\| \\ &\leq \frac{2}{1-\gamma} \|x - x'\|, \end{aligned}$$

where the first inequality is due to the fact that the advantage function is $2L_r$-Lipschitz with respect to x, since it is the difference of the value function and Q function, which are themselves L_r-Lipschitz with respect to x as can be seen from (101). Then, as the entropy regularizer is not a function of x, we have

$$\begin{aligned} \|\nabla_\theta V_\tau^{x,\pi_\theta} - \nabla_\theta V_\tau^{x,\pi_{\theta'}}\| &= \|\nabla_\theta V^{x,\pi_\theta} - \nabla_\theta V^{x,\pi_{\theta'}}\| \\ &\leq \frac{2L_r |\mathcal{S}| |\mathcal{A}|}{1-\gamma} \|x - x'\|. \end{aligned} \tag{102}$$

Combining (95), (99), and (101), we have for all $\tau \leq 1$

$$\begin{aligned}\|V_\tau^{x,\pi_\theta} - V_\tau^{x',\pi_{\theta'}}\| &\leq \|V_\tau^{x,\pi_\theta} - V_\tau^{x',\pi_\theta}\| + \|V_\tau^{x',\pi_\theta} - V_\tau^{x',\pi_{\theta'}}\| \\ &\leq \|V_\tau^{x,\pi_\theta} - V_\tau^{x',\pi_\theta}\| + \tau\|\mathcal{H}(\theta) - \mathcal{H}(\theta')\| + \|V^{x',\pi_\theta} - V^{x',\pi_{\theta'}}\| \\ &\leq L_r\|x - x'\| + \frac{(4 + 8\log|\mathcal{A}|)\sqrt{|\mathcal{S}|}}{(1-\gamma)^3}\|\theta - \theta'\| + \frac{2}{(1-\gamma)^2}\|\theta - \theta'\| \\ &\leq L_r\|x - x'\| + \frac{(6 + 8\log|\mathcal{A}|)\sqrt{|\mathcal{S}|}}{(1-\gamma)^3}\|\theta - \theta'\|.\end{aligned}$$

This shows the Lipschitz continuity of the value function.

To show smoothness with respect to θ, we combine (96), (100), and (102)

$$\begin{aligned}\|\nabla_\theta V_\tau^{x,\pi_\theta} - \nabla_\theta V_\tau^{x',\pi_{\theta'}}\| &\leq \|\nabla_\theta V_\tau^{x,\pi_\theta} - \nabla_\theta V_\tau^{x',\pi_\theta}\| + \|\nabla_\theta V_\tau^{x',\pi_\theta} - \nabla_\theta V_\tau^{x',\pi_{\theta'}}\| \\ &\leq \|\nabla_\theta V_\tau^{x,\pi_\theta} - \nabla_\theta V_\tau^{x',\pi_\theta}\| + \tau\|\nabla_\theta\mathcal{H}(\theta) - \nabla_\theta\mathcal{H}(\theta')\| + \|\nabla_\theta V^{x',\pi_\theta} - \nabla_\theta V^{x',\pi_{\theta'}}\| \\ &\leq \frac{2L_r|\mathcal{S}||\mathcal{A}|}{1-\gamma}\|x - x'\| + \frac{(4 + 8\log|\mathcal{A}|)\sqrt{|\mathcal{S}|}}{(1-\gamma)^3}\|\theta - \theta'\| + \frac{8}{(1-\gamma)^3}\|\theta - \theta'\| \\ &\leq \frac{2L_r|\mathcal{S}||\mathcal{A}|}{1-\gamma}\|x - x'\| + \frac{(12 + 8\log|\mathcal{A}|)\sqrt{|\mathcal{S}|}}{(1-\gamma)^3}\|\theta - \theta'\|.\end{aligned}$$

To show (36),

$$\begin{aligned}&\|\nabla_x J_\tau(x,\pi) - \nabla_x J_\tau(x',\pi')\| \\ &\leq \|\nabla_x J_\tau(x,\pi) - \nabla_x J_\tau(x,\pi')\| + \|\nabla_x J_\tau(x,\pi') - \nabla_x J_\tau(x',\pi')\| \\ &\leq \|\sum_{s,a}(d_\rho^\pi(s,a) - d_\rho^{\pi'}(s,a))\nabla_x r_x(s,a)\| + \mathbb{E}_{s\sim d_\rho^{\pi'},\, a\sim\pi'(\cdot|s)}[\|\nabla_x r_x(s,a) - \nabla_x r_{x'}(s,a)\|] \\ &\leq L_r\|d_\rho^\pi - d_\rho^{\pi'}\| + \mathbb{E}_{s\sim d_\rho^{\pi'},\, a\sim\pi'(\cdot|s)}[L_r\|x - x'\|] \\ &= \left\|(1-\gamma)\Big((I-\gamma P^\pi)^{-1} - (I-\gamma P^{\pi'})^{-1}\Big)\rho\right\| + L_r\|x - x'\| \\ &\leq (1-\gamma)\cdot\frac{\gamma}{(1-\gamma)^2}\|\pi - \pi'\|\|\rho\| + L_r\|x - x'\| \\ &\leq L_V(\|\pi - \pi'\| + \|x - x'\|),\end{aligned}$$

where the fourth inequality is due to the fact that $(I - \gamma P^\pi)$ is $\gamma/(1-\gamma)^2$ Lipschitz in π.

Finally, we show $\nabla^2_{x,\theta}J_\tau(x,\pi_\theta)$ and $\nabla^2_{\theta,\theta}J_\tau(x,\pi_\theta)$ are Lipschitz. Since $\nabla_{\theta,\theta}\mathbb{E}_{s\sim d_\rho^{\pi_\theta}}[E(\pi_\theta,s)]$ does not depend on x and can be shown to have Lipschitz Hessians by extending the argument in Mei et al. [2020][Lemma 14] (we skip showing the exact constant here), the problem reduces to showing that $\nabla^2_{x,\theta}J(x,\pi_\theta)$ and $\nabla^2_{\theta,\theta}J(x,\pi_\theta)$ are Lipschitz.

From (15), we have

$$\nabla^2_{x,\theta}J(x,\pi_\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_\theta},a\sim\pi_\theta(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a)}\left[\Big(\nabla_x r_x(s,a) + \gamma\nabla_x V_\tau^{x,\pi_\theta}(s')\Big)\nabla_\theta\log\pi_\theta(a\mid s)\right]. \tag{103}$$

We define $\mathrm{traj} = \{s_0, a_0, s_1, a_1, \cdots\}$ and use $p(\pi, \mathrm{traj})$ to denote the probability that the trajectory traj is generated under the policy π, i.e.,

$$p(\pi,\mathrm{traj}) = \rho(s_0)\prod_{k=0}^{\infty}\pi(a_k\mid s_k)\mathcal{P}(s_{k+1}\mid s_k,a_k).$$

Adapting the result from Shen et al. [2019], we have

$$\nabla^2_{\theta,\theta}J(x,\pi_\theta) = \mathbb{E}_{\mathrm{traj}\sim p(\pi_\theta,\cdot)}\Big[\nabla_\theta\phi(x,\pi,\mathrm{traj})\nabla_\theta p(\pi_\theta,\mathrm{traj})^\top + \nabla^2_{\theta,\theta}\phi(x,\pi,\mathrm{traj})\Big], \tag{104}$$

where we define $\phi(x,\pi,\mathrm{traj}) = \sum_{t=0}^{\infty}(\sum_{k=t}^{\infty}\gamma^k(r_x(s_k,a_k))^k)\log\pi(a_t\mid s_t)$.

As the right hand side expressions in (103) and (104) are the composition of Lipschitz functions, we know that $\nabla^2_{x,\theta} J(x, \pi_\theta)$ and $\nabla^2_{x,\theta} J(\theta, \pi_\theta)$ are Lipschitz.

■

E.4 Proof of Lemma 4

We can write the discounted visitation distribution as follows

$$d_\rho^\pi = (1-\gamma)(I - \gamma P^\pi)^{-1}\rho,$$

which implies

$$\begin{aligned}
\|d_\rho^\pi - d_\rho^{\pi'}\| &\leq (1-\gamma)\|(I-\gamma P^\pi)^{-1}\rho - (I-\gamma P^{\pi'})^{-1}\|\|\rho\| \\
&= (1-\gamma)\|\rho\|\Big\|(I-\gamma P^{\pi'})^{-1}\Big((I-\gamma P^{\pi'}) - (I-\gamma P^\pi)\Big)(I-\gamma P^\pi)^{-1}\Big\| \\
&\leq (1-\gamma)\cdot 1 \cdot \|(I-\gamma P^\pi)^{-1}\|\|(I-\gamma P^{\pi'})^{-1}\| \cdot \gamma\|P^\pi - P^{\pi'}\| \\
&\leq \frac{\gamma}{1-\gamma}\|\pi - \pi'\|,
\end{aligned}$$

where the third inequality follows from the standard result $\|P^\pi - P^{\pi'}\| \leq \sqrt{|\mathcal{A}|}\|\pi - \pi'\|$.

■

E.5 Proof of Lemma 5

By the definition of D_w in (23),

$$\begin{aligned}
\|D_w(x, \pi, \pi^{\mathcal{L}}, s, a, \bar{s}, \bar{a}, \xi)\| &= \left\|\widetilde{\nabla}_x f(x, \pi^{\mathcal{L}}, \xi) + \frac{1}{w}\Big(\nabla_x r_x(s,a) + \nabla_x r_x(\bar{s}, \bar{a})\Big)\right\| \\
&\leq \|\widetilde{\nabla}_x f(x, \pi^{\mathcal{L}}, \xi)\| + \frac{1}{w}\|\nabla_x r_x(s,a) + \nabla_x r_x(\bar{s}, \bar{a})\| \\
&\leq L_f + \frac{1}{w}(L_r + L_r) \\
&\leq \frac{3L_r}{w},
\end{aligned}$$

where the second inequality follows from Assumption 3 and the condition $\|\nabla_x r_x(s,a)\| \leq L_r$, $\forall x, s, a$, which follows from Assumption 4, and the last inequality is due to $w \leq \frac{L_r}{L_f}$.

To bound F, note that $\nabla_\theta \log \pi_\theta(a \mid s)$ has the following closed-form expression entry-wise

$$\frac{\partial \log \pi_\theta(a \mid s)}{\partial \theta_{s',a'}} = \mathbf{1}\left[s = s'\right]\left(\mathbf{1}\left[a = a'\right] - \pi_\theta\left(a' \mid s\right)\right). \tag{105}$$

This implies

$$\|\nabla_\theta \log \pi_\theta(a \mid s)\|_2 \leq \|\nabla_\theta \log \pi_\theta(a \mid s)\|_1 \leq 1 + 1 = 2. \tag{106}$$

By the definition of $F_{w,\tau}$ in (24),

$$\begin{aligned}
\|F_{w,\tau}(x, \theta, V, s, a, s', \xi)\| &\leq \|\nabla_\theta \log \pi_\theta(a \mid s)\|\Big|r_x(s,a) - \tau \log \pi_\theta(a \mid s) + \gamma V(s') - V(s)\Big| + w\|\widetilde{\nabla}_\theta f(x, \pi_\theta, \xi)\| \\
&\leq 2(1 + \tau|\log \underline{\pi}| + \gamma B_V + B_V) + wL_f \\
&\leq 2(1+\gamma)B_V + 2|\log \underline{\pi}| + 2 + L_r,
\end{aligned}$$

where the second inequality applies (106), and the last inequality follows from $w \leq \frac{L_r}{L_f}, \tau \leq 1$.

Finally, by the definition of G in (25),

$$\begin{aligned}\|G_\tau(x,\theta,V,s,a,s')\| &\le \|e_s\|\Big|r_x(s,a)-\tau\log\pi_\theta(a\mid s)+\gamma V(s')-V(s)\Big|\\ &\le 1\cdot(1+\tau|\log\underline{\pi}|+\gamma B_V+B_V)\\ &\le (1+\gamma)B_V+|\log\underline{\pi}|+1.\end{aligned}$$

■

E.6 Proof of Lemma 6

By the definition of $\bar{D}_w$ in (26),

$$\begin{aligned}&\|\bar{D}_w(x_1,\pi_1,\pi_1^{\mathcal{L}})-\bar{D}_w(x_2,\pi_2,\pi_2^{\mathcal{L}})\|\\ &=\Big\|\mathbb{E}_{s\sim d_\rho^{\pi_1},a\sim\pi_1(\cdot|s),\bar{s}\sim d_\rho^{\pi_1^{\mathcal{L}}},\bar{a}\sim\pi_1^{\mathcal{L}}(\cdot|\bar{s})),\xi\sim\mu}[D_w(x_1,\pi_1,\pi_1^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)]\\ &\qquad-\mathbb{E}_{s\sim d_\rho^{\pi_2},a\sim\pi_2(\cdot|s),\bar{s}\sim d_\rho^{\pi_2^{\mathcal{L}}},\bar{a}\sim\pi_2^{\mathcal{L}}(\cdot|\bar{s})),\xi\sim\mu}[D_w(x_2,\pi_2,\pi_2^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)]\Big\|\\ &=\Big\|\sum_{s,a,\bar{s},\bar{a},\xi}\Big(d_\rho^{\pi_1}(s)\pi_1(a\mid s)d_\rho^{\pi_1^{\mathcal{L}}}(\bar{s})\pi_1^{\mathcal{L}}(\bar{a}\mid\bar{s})-d_\rho^{\pi_2}(s)\pi_2(a\mid s)d_\rho^{\pi_2^{\mathcal{L}}}(\bar{s})\pi_2^{\mathcal{L}}(\bar{a}\mid\bar{s})\Big)\mu(\xi)D_w(x_2,\pi_2,\pi_2^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)\\ &\qquad+\mathbb{E}_{s\sim d_\rho^{\pi_1},a\sim\pi_1(\cdot|s),\bar{s}\sim d_\rho^{\pi_1^{\mathcal{L}}},\bar{a}\sim\pi_1^{\mathcal{L}}(\cdot|\bar{s})),\xi\sim\mu}[D_w(x_1,\pi_1,\pi_1^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)-D_w(x_2,\pi_2,\pi_2^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)]\Big\|\\ &\le\Big\|\mathbb{E}_{s\sim d_\rho^{\pi_1},a\sim\pi_1(\cdot|s),\bar{s}\sim d_\rho^{\pi_1^{\mathcal{L}}},\bar{a}\sim\pi_1^{\mathcal{L}}(\cdot|\bar{s})),\xi\sim\mu}[D_w(x_1,\pi_1,\pi_1^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)-D_w(x_2,\pi_2,\pi_2^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)]\Big\|\\ &\qquad+\frac{B_D}{w}\Big|\sum_{s,a,\bar{s},\bar{a}}\Big(d_\rho^{\pi_1}(s)\pi_1(a\mid s)d_\rho^{\pi_1^{\mathcal{L}}}(\bar{s})\pi_1^{\mathcal{L}}(\bar{a}\mid\bar{s})-d_\rho^{\pi_2}(s)\pi_2(a\mid s)d_\rho^{\pi_2^{\mathcal{L}}}(\bar{s})\pi_2^{\mathcal{L}}(\bar{a}\mid\bar{s})\Big)\Big|\\ &\le\Big\|\mathbb{E}_{s\sim d_\rho^{\pi_1},a\sim\pi_1(\cdot|s),\bar{s}\sim d_\rho^{\pi_1^{\mathcal{L}}},\bar{a}\sim\pi_1^{\mathcal{L}}(\cdot|\bar{s})),\xi\sim\mu}[D_w(x_1,\pi_1,\pi_1^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)-D_w(x_2,\pi_2,\pi_2^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)]\Big\|\\ &\qquad+\frac{B_D}{w}\Big|\sum_{s,a,\bar{s},\bar{a}}\Big(d_\rho^{\pi_1}(s)\pi_1(a\mid s)-d_\rho^{\pi_2}(s)\pi_2(a\mid s)\Big)d_\rho^{\pi_1^{\mathcal{L}}}(\bar{s})\pi_1^{\mathcal{L}}(\bar{a}\mid\bar{s})\Big|\\ &\qquad+\frac{B_D}{w}\Big|\sum_{s,a,\bar{s},\bar{a}}\Big(d_\rho^{\pi_1^{\mathcal{L}}}(\bar{s})\pi_1^{\mathcal{L}}(\bar{a}\mid\bar{s})-d_\rho^{\pi_2^{\mathcal{L}}}(\bar{s})\pi_2^{\mathcal{L}}(\bar{a}\mid\bar{s})\Big)d_\rho^{\pi_2}(s)\pi_2(a\mid s)\Big|\\ &=\Big\|\mathbb{E}_{s\sim d_\rho^{\pi_1},a\sim\pi_1(\cdot|s),\bar{s}\sim d_\rho^{\pi_1^{\mathcal{L}}},\bar{a}\sim\pi_1^{\mathcal{L}}(\cdot|\bar{s})),\xi\sim\mu}[D_w(x_1,\pi_1,\pi_1^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)-D_w(x_2,\pi_2,\pi_2^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)]\Big\|\\ &\qquad+\frac{B_D}{w}\Big(\Big|\sum_{s,a}(d_\rho^{\pi_1}(s)\pi_1(a\mid s)-d_\rho^{\pi_2}(s)\pi_2(a\mid s))\Big|+\Big|\sum_{\bar{s},\bar{a}}(d_\rho^{\pi_1^{\mathcal{L}}}(\bar{s})\pi_1^{\mathcal{L}}(\bar{a}\mid\bar{s})-d_\rho^{\pi_2^{\mathcal{L}}}(\bar{s})\pi_2^{\mathcal{L}}(\bar{a}\mid\bar{s}))\Big|\Big).\end{aligned}\tag{107}$$

To bound the first term of (107), note that by Jensen's inequality it suffices to bound the norm of the term within the expectation

$$\begin{aligned}&\|D_w(x_1,\pi_1,\pi_1^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)-D_w(x_2,\pi_2,\pi_2^{\mathcal{L}},s,a,\bar{s},\bar{a},\xi)\|\\ &=\Big\|\widetilde{\nabla}_x f(x_1,\pi_1^{\mathcal{L}},\xi)+\frac{1}{w}\Big(\nabla_x r_{x_1}(s,a)-\nabla_x r_{x_1}(\bar{s},\bar{a})\Big)-\widetilde{\nabla}_x f(x_2,\pi_2^{\mathcal{L}},\xi)-\frac{1}{w}\Big(\nabla_x r_{x_2}(s,a)-\nabla_x r_{x_2}(\bar{s},\bar{a})\Big)\Big\|\\ &\le\|\widetilde{\nabla}_x f(x_1,\pi_1^{\mathcal{L}},\xi)-\widetilde{\nabla}_x f(x_2,\pi_2^{\mathcal{L}},\xi)\|+\frac{1}{w}\|\nabla_x r_{x_1}(s,a)-\nabla_x r_{x_2}(s,a)\|+\frac{1}{w}\|\nabla_x r_{x_1}(\bar{s},\bar{a})-\nabla_x r_{x_2}(\bar{s},\bar{a})\|\\ &\le L_f\Big(\|x_1-x_2\|+\|\pi_1^{\mathcal{L}}-\pi_2^{\mathcal{L}}\|\Big)+\frac{2L_r}{w}\|x_1-x_2\|.\end{aligned}\tag{108}$$

To bound the second term of (107),

$$\Big|\sum_{s,a}(d_\rho^{\pi_1}(s)\pi_1(a\mid s)-d_\rho^{\pi_2}(s)\pi_2(a\mid s))\Big|+\Big|\sum_{\bar{s},\bar{a}}(d_\rho^{\pi_1^{\mathcal{L}}}(\bar{s})\pi_1^{\mathcal{L}}(\bar{a}\mid\bar{s})-d_\rho^{\pi_2^{\mathcal{L}}}(\bar{s})\pi_2^{\mathcal{L}}(\bar{a}\mid\bar{s}))\Big|$$

$$
\begin{aligned}
&\leq \Big|\sum_{s,a}(d_\rho^{\pi_1}(s)-d_\rho^{\pi_2}(s))\pi_1(a\mid s)\Big| + \Big|\sum_{s,a}(\pi_1(a\mid s)-\pi_2(a\mid s))d_\rho^{\pi_2}(s)\Big| \\
&\quad + \Big|\sum_{\bar{s},\bar{a}}(d_\rho^{\pi_1^{\mathcal{L}}}(\bar{s})-d_\rho^{\pi_2^{\mathcal{L}}}(\bar{s}))\pi_1^{\mathcal{L}}(\bar{a}\mid \bar{s})\Big| + \Big|\sum_{\bar{s},\bar{a}}(\pi_1^{\mathcal{L}}(\bar{a}\mid \bar{s})-\pi_2^{\mathcal{L}}(\bar{a}\mid \bar{s}))d_\rho^{\pi_2^{\mathcal{L}}}(\bar{s})\Big| \\
&\leq \|d_\rho^{\pi_1}-d_\rho^{\pi_2}\| + \|\pi_1-\pi_2\| + \|d_\rho^{\pi_1^{\mathcal{L}}}-d_\rho^{\pi_2^{\mathcal{L}}}\| + \|\pi_1^{\mathcal{L}}-\pi_2^{\mathcal{L}}\| \\
&\leq \frac{1}{1-\gamma}\|\pi_1-\pi_2\| + \frac{1}{1-\gamma}\|\pi_1^{\mathcal{L}}-\pi_2^{\mathcal{L}}\|, \qquad (109)
\end{aligned}
$$

where the final inequality follows from Lemma 4.

Substituting (108) and (109) into (107), we have

$$
\begin{aligned}
&\|\bar{D}_w(x_1,\pi_1,\pi_1^{\mathcal{L}}) - \bar{D}_w(x_2,\pi_2,\pi_2^{\mathcal{L}})\| \\
&\leq L_f\Big(\|x_1-x_2\| + \|\pi_1^{\mathcal{L}}-\pi_2^{\mathcal{L}}\|\Big) + \frac{2L_r}{w}\|x_1-x_2\| + \frac{B_D}{(1-\gamma)w}\|\pi_1-\pi_2\| + \frac{B_D}{(1-\gamma)w}\|\pi_1^{\mathcal{L}}-\pi_2^{\mathcal{L}}\| \\
&= (L_f + \frac{2L_r}{w})\|x_1-x_2\| + (L_f + \frac{B_D}{(1-\gamma)w})\|\pi_1^{\mathcal{L}}-\pi_2^{\mathcal{L}}\| + \frac{B_D}{(1-\gamma)w}\|\pi_1-\pi_2\| \\
&\leq \frac{3L_r}{w}\|x_1-x_2\| + \frac{2B_D}{(1-\gamma)w}\|\pi_1^{\mathcal{L}}-\pi_2^{\mathcal{L}}\| + \frac{B_D}{(1-\gamma)w}\|\pi_1-\pi_2\|, \qquad (110)
\end{aligned}
$$

where the second inequality simplifies terms under the condition $w \leq \min\{\frac{L_r}{L_f}, \frac{B_D}{(1-\gamma)L_f}\}$.

Similarly, by the definition of $\bar{F}_{w,\tau}$ in (27),

$$
\begin{aligned}
&\|\bar{F}_{w,\tau}(x_1,\theta_1,V_1) - \bar{F}_{w,\tau}(x_2,\theta_2,V_2)\| \\
&= \Big\|\mathbb{E}_{s\sim d_\rho^{\pi_{\theta_1}},a\sim\pi_{\theta_1}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a),\xi\sim\mu}[F_{w,\tau}(x_1,\theta_1,V_1,s,a,s',\xi)] \\
&\qquad - \mathbb{E}_{s\sim d_\rho^{\pi_{\theta_2}},a\sim\pi_{\theta_2}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a),\xi\sim\mu}[F_{w,\tau}(x_2,\theta_2,V_2,s,a,s',\xi)]\Big\| \\
&= \Big\|\sum_{s,a,s',\xi}\Big(d_\rho^{\pi_{\theta_1}}(s)\pi_{\theta_1}(a\mid s)\mathcal{P}(s'\mid s,a) - d_\rho^{\pi_{\theta_2}}(s)\pi_{\theta_2}(a\mid s)\mathcal{P}(s'\mid s,a)\Big)\mu(\xi)F_{w,\tau}(x_2,\theta_2,V_2,s,a,s',\xi) \\
&\qquad + \mathbb{E}_{s\sim d_\rho^{\pi_{\theta_1}},a\sim\pi_{\theta_1}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a),\xi\sim\mu}[F_{w,\tau}(x_1,\theta_1,V_1,s,a,s',\xi) - F_{w,\tau}(x_2,\theta_2,V_2,s,a,s',\xi)]\Big\| \\
&\leq \Big\|\mathbb{E}_{s\sim d_\rho^{\pi_{\theta_1}},a\sim\pi_{\theta_1}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a),\xi\sim\mu}[F_{w,\tau}(x_1,\theta_1,V_1,s,a,s',\xi) - F_{w,\tau}(x_2,\theta_2,V_2,s,a,s',\xi)]\Big\| \\
&\qquad + B_F\Big|\sum_{s,a,s'}\Big(d_\rho^{\pi_{\theta_1}}(s)\pi_{\theta_1}(a\mid s)\mathcal{P}(s'\mid s,a) - d_\rho^{\pi_{\theta_2}}(s)\pi_{\theta_2}(a\mid s)\mathcal{P}(s'\mid s,a)\Big)\Big| \\
&= \Big\|\mathbb{E}_{s\sim d_\rho^{\pi_{\theta_1}},a\sim\pi_{\theta_1}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a),\xi\sim\mu}[F_{w,\tau}(x_1,\theta_1,V_1,s,a,s',\xi) - F_{w,\tau}(x_2,\theta_2,V_2,s,a,s',\xi)]\Big\| \\
&\qquad + B_F\Big|\sum_{s,a}\Big(d_\rho^{\pi_{\theta_1}}(s)\pi_{\theta_1}(a\mid s) - d_\rho^{\pi_{\theta_2}}(s)\pi_{\theta_2}(a\mid s)\Big)\Big|. \qquad (111)
\end{aligned}
$$

To bound the first term of (107), note that by Jensen's inequality it suffices to bound the norm of the term within the expectation

$$
\begin{aligned}
&\|F_{w,\tau}(x_1,\theta_1,V_1,s,a,s',\xi) - F_{w,\tau}(x_2,\theta_2,V_2,s,a,s',\xi)\| \\
&= \Big\|\big(r_{x_1}(s,a) - \tau\log\pi_{\theta_1}(a\mid s) + \gamma V_1(s') - V_1(s)\big)\nabla_\theta\log\pi_{\theta_1}(a\mid s) - w\widetilde{\nabla}_\theta f(x_1,\pi_{\theta_1},\xi) \\
&\qquad - \big(r_{x_2}(s,a) - \tau\log\pi_{\theta_2}(a\mid s) + \gamma V_2(s') - V_2(s)\big)\nabla_\theta\log\pi_{\theta_2}(a\mid s) + w\widetilde{\nabla}_\theta f(x_2,\pi_{\theta_2},\xi)\Big\| \\
&\leq w\|\widetilde{\nabla}_\theta f(x_1,\pi_{\theta_1},\xi) - \widetilde{\nabla}_\theta f(x_2,\pi_{\theta_2},\xi)\| \\
&\qquad + \|\nabla_\theta\log\pi_{\theta_1}(a\mid s) - \nabla_\theta\log\pi_{\theta_2}(a\mid s)\|\,|r_{x_1}(s,a) - \tau\log\pi_{\theta_1}(a\mid s) + \gamma V_1(s') - V_1(s)| \\
&\qquad + \|\nabla_\theta\log\pi_{\theta_2}(a\mid s)\|\,|r_{x_1}(s,a) - r_{x_2}(s,a) - \tau\log\pi_{\theta_1}(a\mid s) + \tau\log\pi_{\theta_2}(a\mid s) + \gamma V_1(s') - \gamma V_2(s') - V_1(s) + V_2(s)| \\
&\leq L_f w(\|x_1-x_2\| + \|\pi_{\theta_1}-\pi_{\theta_2}\|)
\end{aligned}
$$

$$
\begin{aligned}
&+ |r_{x_1}(s,a) - \tau \log \pi_{\theta_1}(a \mid s) + \gamma V_1(s') - V_1(s)| \|\theta_1 - \theta_2\| \\
&+ \|\nabla_\theta \log \pi_{\theta_2}(a \mid s)\| \Big(L_r \|x_1 - x_2\| + \frac{\tau}{\underline{\pi}} \|\theta_1 - \theta_2\| + (1+\gamma)\|V_1 - V_2\| \Big) \\
\leq\ & L_r(\|x_1 - x_2\| + \|\theta_1 - \theta_2\|) + (1 + |\log \underline{\pi}| + \frac{\gamma}{1-\gamma} + \frac{1}{1-\gamma})\|\theta_1 - \theta_2\| \\
&+ 2\Big(L_r \|x_1 - x_2\| + \frac{1}{\underline{\pi}} \|\theta_1 - \theta_2\| + (1+\gamma)\|V_1 - V_2\| \Big) \\
\leq\ & 3L_r \|x_1 - x_2\| + (L_r + \frac{2}{\underline{\pi}} + |\log \underline{\pi}| + \frac{2}{1-\gamma} + 1)\|\theta_1 - \theta_2\| + 4\|V_1 - V_2\|, \qquad (112)
\end{aligned}
$$

where the second inequality is due to the 1-Lipschitz continuity of $\nabla_\theta \log \pi_\theta$ and the $\log |\mathcal{A}|$-Lipschitz continuity of the entropy function with respect to the softmax parameter, and the third inequality is due to $w \leq \frac{L_r}{L_f}, \tau \leq 1$ and the fact that $\|\nabla_\theta \log \pi_\theta(a \mid s)\|_2 \leq 2$ for all θ.

To bound the second term of (111),

$$
\begin{aligned}
\Big| \sum_{s,a} (d_\rho^{\pi_{\theta_1}}(s)\pi_{\theta_1}(a \mid s) - d_\rho^{\pi_{\theta_2}}(s)\pi_{\theta_2}(a \mid s)) \Big| &\leq \Big| \sum_{s,a} (d_\rho^{\pi_{\theta_1}}(s) - d_\rho^{\pi_{\theta_2}}(s))\pi_{\theta_1}(a \mid s) \Big| \\
&\quad + \Big| \sum_{s,a} (\pi_{\theta_1}(a \mid s) - \pi_{\theta_2}(a \mid s)) d_\rho^{\pi_{\theta_2}}(s) \Big| \\
&\leq \|d_\rho^{\pi_{\theta_1}} - d_\rho^{\pi_{\theta_2}}\| + \|\pi_{\theta_1} - \pi_{\theta_2}\| \\
&\leq \frac{1}{1-\gamma} \|\pi_{\theta_1} - \pi_{\theta_2}\| \\
&\leq \frac{1}{1-\gamma} \|\theta_1 - \theta_2\|, \qquad (113)
\end{aligned}
$$

where the third inequality follows from Lemma 4.

Substituting (112) and (113) into (109), we have

$$
\begin{aligned}
&\|\bar{F}_{w,\tau}(x_1, \theta_1, V_1) - \bar{F}_{w,\tau}(x_2, \theta_2, V_2)\| \\
&\leq 3L_r \|x_1 - x_2\| + (L_r + \frac{2}{\underline{\pi}} + |\log \underline{\pi}| + \frac{2}{1-\gamma} + 1)\|\theta_1 - \theta_2\| + 4\|V_1 - V_2\| + \frac{B_F}{1-\gamma}\|\theta_1 - \theta_2\| \\
&= 3L_r \|x_1 - x_2\| + (L_r + \frac{2}{\underline{\pi}} + |\log \underline{\pi}| + \frac{2 + B_F}{1-\gamma} + 1)\|\theta_1 - \theta_2\| + 4\|V_1 - V_2\|.
\end{aligned}
$$

By the definition of $\bar{G}_\tau$ in (28),

$$
\begin{aligned}
&\|\bar{G}_\tau(x_1, \theta_1, V_1) - \bar{G}_\tau(x_2, \theta_2, V_2)\| \\
&= \Big\| \mathbb{E}_{s \sim d_\rho^{\pi_{\theta_1}}, a \sim \pi_{\theta_1}(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a)} [G_\tau(x_1, \theta_1, V_1, s, a, s')] \\
&\qquad - \mathbb{E}_{s \sim d_\rho^{\pi_{\theta_2}}, a \sim \pi_{\theta_2}(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a)} [G_\tau(x_2, \theta_2, V_2, s, a, s')] \Big\| \\
&= \Big\| \sum_{s,a,s'} \Big(d_\rho^{\pi_{\theta_1}}(s)\pi_{\theta_1}(a \mid s)\mathcal{P}(s' \mid s,a) - d_\rho^{\pi_{\theta_2}}(s)\pi_{\theta_2}(a \mid s)\mathcal{P}(s' \mid s,a) \Big) G_\tau(x_2, \theta_2, V_2, s, a, s', \xi) \\
&\qquad + \mathbb{E}_{s \sim d_\rho^{\pi_{\theta_1}}, a \sim \pi_{\theta_1}(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a)} [G_\tau(x_1, \theta_1, V_1, s, a, s') - G_\tau(x_2, \theta_2, V_2, s, a, s')] \Big\| \\
&\leq \Big\| \mathbb{E}_{s \sim d_\rho^{\pi_{\theta_1}}, a \sim \pi_{\theta_1}(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a)} [G_\tau(x_1, \theta_1, V_1, s, a, s') - G_\tau(x_2, \theta_2, V_2, s, a, s')] \Big\| \\
&\qquad + B_G \Big| \sum_{s,a,s'} \Big(d_\rho^{\pi_{\theta_1}}(s)\pi_{\theta_1}(a \mid s)\mathcal{P}(s' \mid s,a) - d_\rho^{\pi_{\theta_2}}(s)\pi_{\theta_2}(a \mid s)\mathcal{P}(s' \mid s,a) \Big) \Big| \\
&= \Big\| \mathbb{E}_{s \sim d_\rho^{\pi_{\theta_1}}, a \sim \pi_{\theta_1}(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a), \xi \sim \mu} [F_{w,\tau}(x_1, \theta_1, V_1, s, a, s', \xi) - F_{w,\tau}(x_2, \theta_2, V_2, s, a, s', \xi)] \Big\| \\
&\qquad + B_G \Big| \sum_{s,a} \Big(d_\rho^{\pi_{\theta_1}}(s)\pi_{\theta_1}(a \mid s) - d_\rho^{\pi_{\theta_2}}(s)\pi_{\theta_2}(a \mid s) \Big) \Big|
\end{aligned}
$$

$$\leq \left\|\mathbb{E}_{s\sim d_\rho^{\pi_{\theta_1}},a\sim\pi_{\theta_1}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a),\xi\sim\mu}[F_{w,\tau}(x_1,\theta_1,V_1,s,a,s',\xi)-F_{w,\tau}(x_2,\theta_2,V_2,s,a,s',\xi)]\right\| + \frac{B_G}{1-\gamma}\|\theta_1-\theta_2\|, \tag{114}$$

where the last inequality follows from (113).

The first of (114) can be bounded as follows. Again, by Jensen's inequality it suffices to bound the norm of the term within the expectation

$$\begin{aligned}
&\|G_\tau(x_1,\theta_1,V_1,s,a,s')-G_\tau(x_2,\theta_2,V_2,s,a,s')\| \\
&= \left\|\big(r_{x_1}(s,a)-\tau\log\pi_{\theta_1}(a\mid s)+\gamma V_1(s')-V_1(s)\big)e_s-\big(r_{x_2}(s,a)-\tau\log\pi_{\theta_2}(a\mid s)+\gamma V_2(s')-V_2(s)\big)e_s\right\| \\
&\leq L_r\|x_1-x_2\|+\frac{\tau}{\underline{\pi}}\|\theta_1-\theta_2\|+(1+\gamma)\|V_1-V_2\|,
\end{aligned} \tag{115}$$

where in the last inequality we again use the $\log|\mathcal{A}|$-Lipschitz continuity of the entropy function with respect to the softmax parameter.

Combining (114) and (115), we have under τ

$$\begin{aligned}
\|\bar{G}_\tau(x_1,\theta_1,V_1)-\bar{G}_\tau(x_2,\theta_2,V_2)\| &\leq L_r\|x_1-x_2\|+\frac{\tau}{\underline{\pi}}\|\theta_1-\theta_2\|+(1+\gamma)\|V_1-V_2\| \\
&\quad+\frac{B_G}{1-\gamma}\|\theta_1-\theta_2\| \\
&\leq L_r\|x_1-x_2\|+(\frac{B_G}{1-\gamma}+\frac{1}{\underline{\pi}})\|\theta_1-\theta_2\|+2\|V_1-V_2\|.
\end{aligned}$$

■

E.7 Proof of Lemma 7

The proof proceeds in a manner similar to Kwon et al. [2023][Lemma 3.2], with strong convexity replaced by the PL condition.

First, we consider a fixed τ. Recall the definition of $\pi^\star_{w,\tau}$ in Section 3.1. Let $\theta^\star_{w_1,\tau}(x_1)$ denote a softmax parameter that encodes $\pi^\star_{w_1,\tau}(x_1)$. The optimality condition of $\theta^\star_{w_1,\tau}(x_1)$ indicates

$$\nabla_\theta\mathcal{L}_{w_1,\tau}(x_1,\pi_{\theta^\star_{w_1,\tau}(x_1)})=\nabla_\theta f(x_1,\pi_{\theta^\star_{w_1,\tau}(x_1)})-\frac{1}{w_1}\nabla_\pi J_\tau(x_1,\pi_{\theta^\star_{w_1,\tau}(x_1)})=0, \tag{116}$$

which obviously implies

$$\|\nabla_\theta J_\tau(x_1,\pi_{\theta^\star_{w_1,\tau}(x_1)})\|=w_1\|\nabla_\theta f(x_1,\pi_{\theta^\star_{w_1,\tau}(x_1)})\|\leq L_f w_1. \tag{117}$$

Applying the relationship in (116), we have

$$\begin{aligned}
&\nabla_\theta\mathcal{L}_{w_2,\tau}(x_2,\pi_{\theta^\star_{w_1,\tau}(x_1)}) \\
&=\nabla_\theta f(x_2,\pi_{\theta^\star_{w_1,\tau}(x_1)})-\frac{1}{w_2}\nabla_\pi J_\tau(x_2,\pi_{\theta^\star_{w_1,\tau}(x_1)}) \\
&=\Big(\nabla_\theta f(x_2,\pi_{\theta^\star_{w_1,\tau}(x_1)})-\nabla_\theta f(x_1,\pi_{\theta^\star_{w_1,\tau}(x_1)})\Big)-\frac{1}{w_2}\Big(\nabla_\theta J_\tau(x_2,\pi_{\theta^\star_{w_1,\tau}(x_1)})-\nabla_\theta J_\tau(x_1,\pi_{\theta^\star_{w_1,\tau}(x_1)})\Big) \\
&\quad+\nabla_\theta f(x_1,\pi_{\theta^\star_{w_1,\tau}(x_1)})-\frac{1}{w_2}\nabla_\pi J_\tau(x_1,\pi_{\theta^\star_{w_1,\tau}(x_1)}) \\
&=\Big(\nabla_\theta f(x_2,\pi_{\theta^\star_{w_1,\tau}(x_1)})-\nabla_\theta f(x_1,\pi_{\theta^\star_{w_1,\tau}(x_1)})\Big)-\frac{1}{w_2}\Big(\nabla_\theta J_\tau(x_2,\pi_{\theta^\star_{w_1,\tau}(x_1)})-\nabla_\theta J_\tau(x_1,\pi_{\theta^\star_{w_1,\tau}(x_1)})\Big) \\
&\quad-(\frac{1}{w_2}-\frac{1}{w_1})\nabla_\theta J_\tau(x_1,\pi_{\theta^\star_{w_1,\tau}(x_1)}).
\end{aligned}$$

Taking the norm,

$$\begin{aligned}
&\|\nabla_\theta \mathcal{L}_{w_2,\tau}(x_2, \pi_{\theta^\star_{w_1,\tau}(x_1)})\| \\
&\leq \|\nabla_\theta f(x_2, \pi^\star_{w_1,\tau}(x_1)) - \nabla_\pi f(x_1, \pi^\star_{w_1,\tau}(x_1))\| + \frac{1}{w_2}\|\nabla_\pi J_\tau(x_2, \pi^\star_{w_1,\tau}(x_1)) - \nabla_\theta J_\tau(x_1, \pi_{\theta^\star_{w_1,\tau}(x_1)})\| \\
&\quad + |\frac{1}{w_2} - \frac{1}{w_1}|\|\nabla_\theta J_\tau(x_1, \pi_{\theta^\star_{w_1,\tau}(x_1)})\| \\
&\leq L_f\|x_1 - x_2\| + \frac{L_V}{w_2}\|x_1 - x_2\| + |\frac{1}{w_2} - \frac{1}{w_1}|\|\nabla_\theta J_\tau(x_1, \pi_{\theta^\star_{w_1,\tau}(x_1)})\| \\
&\leq (L_f + \frac{L_V}{w_2})\|x_1 - x_2\| + \frac{L_f|w_1 - w_2|}{w_2},
\end{aligned} \tag{118}$$

where the second inequality follows from the Lipschitz continuous gradients of f and J_τ, and the third inequality plugs in (117).

Due to (22), we have

$$\|\nabla_\theta \mathcal{L}_{w_2,\tau}(x, \pi_{\theta^\star_{w_1,\tau}(x_1)})\| \geq \frac{C_L\tau}{2w_2}\|\pi^\star_{w_1,\tau}(x_1) - \pi^\star_{w_2,\tau}(x_2)\|. \tag{119}$$

Combining (118) and (119),

$$\frac{C_L\tau}{2w_2}\|\pi^\star_{w_1,\tau}(x_1) - \pi^\star_{w_2,\tau}(x_2)\| \leq (L_f + \frac{L_V}{w_2})\|x_1 - x_2\| + \frac{L_f|w_1 - w_2|}{w_2}.$$

For any $w_2 > 0$, this simplifies to

$$\|\pi^\star_{w_1,\tau}(x_1) - \pi^\star_{w_2,\tau}(x_2)\| \leq (\frac{2L_f w_2}{C_L\tau} + \frac{2L_V}{C_L\tau})\|x_1 - x_2\| + \frac{2L_f|w_1 - w_2|}{C_L\tau}. \tag{120}$$

Recognizing $\pi^\star_\tau(x) = \lim_{w\to 0^+} \pi^\star_{w,\tau}(x)$, we have from (120)

$$\|\pi^\star_\tau(x) - \pi^\star_{w,\tau}(x)\| \leq \frac{2L_f w}{C_L\tau}.$$

Now, we fix w, x and show the bound on $\|\pi^\star_{w,\tau_1}(x) - \pi^\star_{w,\tau_2}(x)\|$. We use $\theta_{w,\tau_1}(x)^\star$ to denote a softmax parameter for $\pi_{w,\tau_1}(x)^\star$. The optimality condition of $\theta_{w,\tau_1}(x)^\star$ indicates

$$\nabla_\theta \mathcal{L}_{w,\tau_1}(x, \pi_{\theta^\star_{w,\tau_1}(x)}) = \nabla_\theta f(x, \pi_{\theta^\star_{w,\tau_1}(x)}) - \frac{1}{w}\nabla_\theta J_{\tau_1}(x, \pi_{\theta^\star_{w,\tau_1}(x)}) = 0.$$

Applying the equation, we get

$$\begin{aligned}
\nabla_\theta \mathcal{L}_{w,\tau_2}(x, \pi_{\theta^\star_{w,\tau_1}(x)}) &= \nabla_\theta f(x, \pi_{\theta^\star_{w,\tau_1}(x)}) - \frac{1}{w}\nabla_\theta J_{\tau_2}(x, \pi_{\theta^\star_{w,\tau_1}(x)}) \\
&= \frac{1}{w}\Big(\nabla_\theta J_{\tau_1}(x, \pi_{\theta^\star_{w,\tau_1}(x)}) - \nabla_\theta J_{\tau_2}(x, \pi_{\theta^\star_{w,\tau_1}(x)})\Big).
\end{aligned} \tag{121}$$

The regularized RL objective has a closed-form expression (see Mei et al. [2020][Lemma 10])

$$\frac{\partial J_\tau(x, \pi_\theta)}{\partial \theta(s,a)} = \frac{d_\rho^{\pi_\theta}(s)}{1-\gamma} \cdot \pi_\theta(a \mid s) \cdot A_\tau^{x,\pi_\theta}(s,a),$$

which in combination with (121) implies

$$\begin{aligned}
\|\nabla_\theta \mathcal{L}_{w,\tau_2}(x, \pi_{\theta^\star_{w,\tau_1}(x)})\| &\leq \|\nabla_\theta \mathcal{L}_{w,\tau_2}(x, \pi_{\theta^\star_{w,\tau_1}(x)})\|_1 \\
&\leq \frac{1}{(1-\gamma)w}\sum_{s,a} \pi_{\theta^\star_{w,\tau_1}(x)}(a \mid s)\Big|A_{\tau_1}^{x,\pi_{\theta^\star_{w,\tau_1}(x)}}(s,a) - A_{\tau_2}^{x,\pi_{\theta^\star_{w,\tau_1}(x)}}(s,a)\Big|.
\end{aligned} \tag{122}$$

Due to Zeng et al. [2022][Lemma 3], we have for any s

$$|V_{\tau_1}^{x,\pi}(s) - V_{\tau_2}^{x,\pi}(s)| \leq |\tau_1 - \tau_2| \log |\mathcal{A}|.$$

The definitions of the Q function and advantage function in (20) imply

$$|Q_{\tau_1}^{x,\pi}(s,a) - Q_{\tau_2}^{x,\pi}(s,a)| \leq \gamma |V_{\tau_1}^{x,\pi}(s) - V_{\tau_2}^{x,\pi}(s)| \leq \gamma |\tau_1 - \tau_2| \log |\mathcal{A}|,$$

and

$$\begin{aligned}
&\sum_a \pi(a \mid s) |A_{\tau_1}^{x,\pi}(s,a) - A_{\tau_2}^{x,\pi}(s,a)| \\
&\leq \sum_a \pi(a \mid s) |Q_{\tau_1}^{x,\pi}(s,a) - Q_{\tau_2}^{x,\pi}(s,a)| + |V_{\tau_1}^{x,\pi}(s) - V_{\tau_2}^{x,\pi}(s)| + |\tau_1 - \tau_2| E(\pi, s) \\
&\leq 3|\tau_1 - \tau_2| \log |\mathcal{A}|.
\end{aligned} \tag{123}$$

Plugging (123) into (122),

$$\|\nabla_\theta \mathcal{L}_{w,\tau_2}(x, \pi_{\theta^\star_{w,\tau_1}(x)})\| \leq \frac{3|\tau_1 - \tau_2||\mathcal{S}| \log |\mathcal{A}|}{(1-\gamma)w}.$$

Again, due to (22),

$$\frac{C_L \tau_2}{2w} \|\pi^\star_{w,\tau_1}(x) - \pi^\star_{w,\tau_2}(x)\| \leq \|\nabla_\pi \mathcal{L}_{w,\tau_2}(x, \pi^\star_{w,\tau_1}(x))\| \leq \frac{3|\tau_1 - \tau_2||\mathcal{S}| \log |\mathcal{A}|}{(1-\gamma)w}.$$

This leads to

$$\|\pi^\star_{w,\tau_1}(x) - \pi^\star_{w,\tau_2}(x)\| \leq \frac{6|\tau_1 - \tau_2||\mathcal{S}| \log |\mathcal{A}|}{(1-\gamma)C_L \tau_2}. \tag{124}$$

Putting (120) and (124) together,

$$\begin{aligned}
\|\pi^\star_{w_1,\tau_1}(x_1) - \pi^\star_{w_2,\tau_2}(x_2)\| &\leq \|\pi^\star_{w_1,\tau_1}(x_1) - \pi^\star_{w_2,\tau_1}(x_2)\| + \|\pi^\star_{w_2,\tau_2}(x_2) - \pi^\star_{w_2,\tau_1}(x_2)\| \\
&\leq (\frac{2L_f w_2}{C_L \tau_1} + \frac{2L_V}{C_L \tau_1})\|x_1 - x_2\| + \frac{2L_f |w_1 - w_2|}{C_L \tau_1} + \frac{6|\tau_1 - \tau_2||\mathcal{S}| \log |\mathcal{A}|}{(1-\gamma)C_L \tau_1}.
\end{aligned}$$

■

E.8 Proof of Lemma 8

We know from Lemma 1 that $\pi^\star(x)$, defined in (5), is the limit point of $\pi^\star_\tau(x)$ as $\tau \to 0$ Let $\theta^\star_\tau(x)$ denote a softmax parameter for $\pi^\star_\tau(x)$. By the first-order optimality condition, we have

$$\nabla_\theta J_\tau(x, \pi_{\theta^\star_\tau(x)}) = 0.$$

We further differentiate with respect to τ. Due to the differentiation chain rule,

$$\frac{d}{d\tau} \nabla_\theta J_\tau(x, \pi_{\theta^\star_\tau(x)}) = \nabla^2_{\tau,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)}) + \nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)}) \cdot \frac{d\theta^\star_\tau(x)}{d\tau} = 0.$$

As $\|\nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)})\|$ is lower bounded by $\underline{\sigma}$ due to Assumption 2, we have for any $\tau \geq 0$

$$\begin{aligned}
\left\| \frac{d\theta^\star_\tau(x)}{d\tau} \right\| &= \left\| -\left(\|\nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)})\right)^{-1} \nabla^2_{\tau,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)}) \right\| \\
&\leq \left\| \left(\|\nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)})\right)^{-1} \right\| \|\nabla^2_{\tau,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)})\|
\end{aligned}$$

$$\leq \frac{1}{\underline{\sigma}}\|\nabla^2_{\tau,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)})\|. \tag{125}$$

It is clear from (2) and (3)

$$\nabla_\tau J_\tau(x, \pi_\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}[E(\pi_\theta, s)].$$

Therefore,

$$\nabla^2_{\tau,\theta} J_\tau(x, \pi_\theta) = \frac{1}{1-\gamma}\nabla_\theta\mathbb{E}_{s\sim d_\rho^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}[E(\pi_\theta, s)]. \tag{126}$$

Zeng et al. [2022][Lemma 6] shows that $\mathbb{E}_{s\sim d_\rho^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}[E(\pi_\theta, s)]$ is Lipschitz with constant $\frac{4+8\log|\mathcal{A}|}{(1-\gamma)^3}$, which is equivalent to

$$\|\nabla_\theta\mathbb{E}_{s\sim d_\rho^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}[E(\pi_\theta, s)]\| \leq \frac{4+8\log|\mathcal{A}|}{(1-\gamma)^3}, \quad \forall\theta. \tag{127}$$

Combining (125)-(127), we have

$$\left\|\frac{d\theta^\star_\tau(x)}{d\tau}\right\| = \frac{1}{\underline{\sigma}}\cdot\frac{1}{1-\gamma}\cdot\frac{4+8\log|\mathcal{A}|}{(1-\gamma)^3} = \frac{4+8\log|\mathcal{A}|}{\underline{\sigma}(1-\gamma)^4} = L_\star.$$

This implies $\theta^\star_\tau(x)$ is $L_\star$-Lipschitz with respect to τ for $\tau \geq 0$.

■

E.9 Proof of Lemma 9

It can be seen from (12) that the gradients of the Lagrangian function have the following closed-form expressions

$$\begin{aligned}\nabla_x\mathcal{L}_{w,\tau}(x,\pi) &= \nabla_x f(x,\pi) + \frac{1}{w}\Big(\nabla_x J_\tau(x,\pi^\star_\tau(x)) + \nabla_x\pi^\star_\tau(x)\nabla_\pi J_\tau(x,\pi^\star_\tau(x)) - \nabla_x J_\tau(x,\pi)\Big)\\ &= \nabla_x f(x,\pi) + \frac{1}{w}\Big(\nabla_x J_\tau(x,\pi^\star_\tau(x)) - \nabla_x J_\tau(x,\pi)\Big), \qquad (128)\\ \nabla_\pi\mathcal{L}_{w,\tau}(x,\pi) &= \nabla_\pi f(x,\pi) - \frac{1}{w}\nabla_\pi J_\tau(x,\pi), \qquad (129)\end{aligned}$$

where the equation (128) is due to the optimality condition of J_τ at $\pi^\star_\tau(x)$.

According to (128), we have

$$\begin{aligned}&\|\nabla_x\mathcal{L}_{w,\tau}(x,\pi) - \nabla_x\mathcal{L}_{w,\tau}(x',\pi')\|\\ &= \|\nabla_x f(x,\pi) + \frac{1}{w}\Big(\nabla_x J_\tau(x,\pi^\star_\tau(x)) - \nabla_x J_\tau(x,\pi)\Big) - \nabla_x f(x',\pi') - \frac{1}{w}\Big(\nabla_x J_\tau(x',\pi^\star_\tau(x')) - \nabla_x J_\tau(x',\pi')\Big)\|\\ &\leq \|\nabla_x f(x,\pi) - \nabla_x f(x',\pi')\| + \frac{1}{w}\|\nabla_x J_\tau(x,\pi^\star_\tau(x)) - \nabla_x J_\tau(x',\pi^\star_\tau(x'))\|\\ &\quad + \frac{1}{w}\|\nabla_x J_\tau(x,\pi) - \nabla_x J_\tau(x',\pi')\|\\ &\leq L_f(\|x-x'\| + \|\pi-\pi'\|) + \frac{L_V}{w}(\|x-x'\| + \|\pi^\star_\tau(x) - \pi^\star_\tau(x')\|) + \frac{L_V}{w}(\|x-x'\| + \|\pi-\pi'\|).\end{aligned}$$

Recognizing $\pi^\star_\tau(x) = \lim_{w\to 0^+}\pi^\star_{w,\tau}(x)$, we have from Lemma 7

$$\|\pi^\star_\tau(x) - \pi^\star_\tau(x')\| \leq \frac{2L_V}{C_L\tau}\|x-x'\|.$$

Combining the two inequalities above and imposing the condition $w, \tau \leq 1$, we get

$$\begin{aligned}\|\nabla_x \mathcal{L}_{w,\tau}(x,\pi) - \nabla_x \mathcal{L}_{w,\tau}(x',\pi')\| &\leq (L_V + L_f + \frac{L_V(C_L+2L_V)}{C_L})\frac{1}{w\tau}\|x-x'\| + \frac{L_f+L_V}{w}\|\pi-\pi'\| \\ &\leq \frac{L_L}{w\tau}\|x-x'\| + \frac{L_L}{w}\|\theta-\theta'\|.\end{aligned}$$

According to (129), we have

$$\begin{aligned}&\|\nabla_\theta \mathcal{L}_{w,\tau}(x,\pi_\theta) - \nabla_\theta \mathcal{L}_{w,\tau}(x',\pi_{\theta'})\| \\ &= \|\nabla_\theta f(x,\pi_\theta) - \frac{1}{w}\nabla_\theta J_\tau(x,\pi_\theta) - \nabla_\theta f(x',\pi_{\theta'}) - \frac{1}{w}\nabla_\theta J_\tau(x',\pi_{\theta'})\| \\ &\leq \|\nabla_\theta f(x,\pi_\theta) - \nabla_\theta f(x',\pi_{\theta'})\| + \frac{1}{w}\|\nabla_\theta J_\tau(x,\pi_\theta) - \nabla_\theta J_\tau(x',\pi_{\theta'})\| \\ &\leq L_f(\|x-x'\| + \|\pi_\theta - \pi_{\theta'}\|) + \frac{L_V}{w}(\|x-x'\| + \|\pi_\theta-\pi_{\theta'}\|) \\ &\leq \frac{L_L}{w}\|x-x'\| + \frac{L_L}{w}\|\theta-\theta'\|.\end{aligned}$$

■

E.10 Proof of Lemma 10

Let $\theta_\tau^\star(x)$ denote a softmax parameter of $\pi_\tau^\star(x)$. By the definition of ℓ_τ, we have

$$\nabla \ell_\tau(x) = \nabla_x J_\tau(x, \pi_{\theta_\tau^\star(x)}),$$

due to $\nabla_\theta J_\tau(x, \pi_{\theta_\tau^\star(x)}) = 0$.

As J_τ is L_V-smooth (from Lemma 3) this implies

$$\begin{aligned}\|\nabla \ell_\tau(x_1) - \nabla \ell_\tau(x_2)\| &= \|\nabla_x J_\tau(x_1, \pi_{\theta_\tau^\star(x_1)}) - \nabla_x J_\tau(x_2, \pi_{\theta_\tau^\star(x_2)})\| \\ &\leq L_V\|x_1-x_2\| + L_V\|\pi_\tau^\star(x_1) - \pi_\tau^\star(x_2)\| \\ &\leq L_V\|x_1-x_2\| + L_V \cdot \frac{2L_V}{C_L\tau}\|x_1-x_2\| \\ &\leq \Big(L_V + \frac{2L_V^2}{C_L\tau}\Big)\|x_1-x_2\|,\end{aligned}$$

where the second inequality follows from Lemma 7 by recognizing that $\pi_\tau^\star(x) = \lim_{w\to 0}\pi_{w,\tau}^\star(x)$.

We next show the smoothness of $\Phi_{w,\tau}$. From (13) it can be seen

$$\begin{aligned}&\|\nabla_x \Phi_{w,\tau}(x_1) - \nabla_x \Phi_{w,\tau}(x_2)\| \\ &\leq \Big\|\nabla_x f(x_1,\pi_\tau^\star(x_1)) - \nabla_x f(x_2,\pi_\tau^\star(x_2))\Big\| + \frac{1}{w}\Big\|\nabla_x J_\tau(x_1,\pi_\tau^\star(x_1)) - \nabla_x J_\tau(x_2,\pi_\tau^\star(x_2))\Big\| \\ &\quad + \frac{1}{w}\Big\|\nabla_x J_\tau(x_1,\pi_{w,\tau}^\star(x_1)) - \nabla_x J_\tau(x_2,\pi_{w,\tau}^\star(x_2))\Big\| \\ &\leq L_f\|x_1-x_2\| + \Big(L_f + \frac{L_V}{w}\Big)\|\pi_\tau^\star(x_1) - \pi_\tau^\star(x_2)\| + \frac{L_V}{w}\|\pi_{w,\tau}^\star(x_1) - \pi_{w,\tau}^\star(x_2)\| \\ &\leq L_f\|x_1-x_2\| + \Big(L_f + \frac{L_V}{w}\Big)\cdot\frac{2L_V}{C_L\tau}\|x_1-x_2\| + \frac{L_V}{w}\cdot\Big(\frac{2L_f w_2}{C_L\tau} + \frac{2L_V}{C_L\tau}\Big)\|x_1-x_2\| \\ &\leq \Big(L_f + \frac{4L_fL_V}{C_L\tau} + \frac{4L_V^2}{C_L w\tau}\Big)\|x_1-x_2\|.\end{aligned}$$

Finally, we show the smoothness of Φ_τ. Let $\theta_\tau^\star(x)$ denote one of the softmax parameters that encodes $\pi_\tau^\star(x)$. We can express the hyper-gradient of Φ_τ as follows

$$\nabla_x \Phi_\tau(x) = \nabla_x f(x, \pi_{\theta_\tau^\star(x)}) - \nabla_{x,\theta}^2 J_\tau(x, \pi_{\theta_\tau^\star(x)})\nabla_{\theta,\theta}^2 J_\tau(x, \pi_{\theta_\tau^\star(x)})^{-1}\nabla_\theta f(x, \pi_{\theta_\tau^\star(x)}).$$

This implies

$$
\begin{aligned}
&\|\nabla_x \Phi_\tau(x_1) - \nabla_x \Phi_\tau(x_2)\| \\
&\leq \underbrace{\|\nabla_x f(x_1, \pi_{\theta^\star_\tau(x_1)}) - \nabla_x f(x_2, \pi_{\theta^\star_\tau(x_2)})\|}_{T_1} \\
&+\underbrace{\|\nabla^2_{x,\theta} J_\tau(x_1, \pi_{\theta^\star_\tau(x_1)}) \nabla^2_{\theta,\theta} J_\tau(x_1, \pi_{\theta^\star_\tau(x_1)})^{-1} \nabla_\theta f(x_1, \pi_{\theta^\star_\tau(x_1)}) - \nabla^2_{x,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)}) \nabla^2_{\theta,\theta} J_\tau(x_1, \pi_{\theta^\star_\tau(x_1)})^{-1} \nabla_\theta f(x_1, \pi_{\theta^\star_\tau(x_1)})\|}_{T_2} \\
&+\underbrace{\|\nabla^2_{x,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)}) \nabla^2_{\theta,\theta} J_\tau(x_1, \pi_{\theta^\star_\tau(x_1)})^{-1} \nabla_\theta f(x_1, \pi_{\theta^\star_\tau(x_1)}) - \nabla^2_{x,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)}) \nabla^2_{\theta,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)})^{-1} \nabla_\theta f(x_1, \pi_{\theta^\star_\tau(x_1)})\|}_{T_3} \\
&+\underbrace{\|\nabla^2_{x,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)}) \nabla^2_{\theta,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)})^{-1} \nabla_\theta f(x_1, \pi_{\theta^\star_\tau(x_1)}) - \nabla^2_{x,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)}) \nabla^2_{\theta,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)})^{-1} \nabla_\theta f(x_2, \pi_{\theta^\star_\tau(x_2)})\|}_{T_4}.
\end{aligned}
\tag{130}
$$

We treat each term of (130) individually. First, we bound T_1 using the smoothness property of f

$$
\begin{aligned}
T_1 &\leq L_f \Big(\|x_1 - x_2\| + \|\pi^\star_\tau(x_1) - \pi^\star_\tau(x_2)\| \Big) \\
&\leq (L_f + \frac{2L_f L_V}{C_L \tau}) \|x_1 - x_2\|,
\end{aligned}
\tag{131}
$$

where to derive the second inequality, we plug in the result from Lemma 7 to get $\|\pi^\star_\tau(x_1) - \pi^\star_\tau(x_2)\| \leq \frac{2L_V}{C_L \tau}\|x_1 - x_2\|$ (note that $\pi^\star_\tau(x) = \lim_{w \to 0} \pi^\star_{w,\tau}(x)$).

As we have $\nabla_\theta f(x, \pi_\theta) \leq L_f$ from Assumption 3 and $\|\nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)})^{-1}\| \leq \frac{1}{\underline{\sigma}}$ due to Assumption 2,

$$
\begin{aligned}
T_2 &\leq \|\nabla^2_{x,\theta} J_\tau(x_1, \pi_{\theta^\star_\tau(x_1)}) - \nabla^2_{x,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)})\| \|\nabla^2_{\theta,\theta} J_\tau(x_1, \pi_{\theta^\star_\tau(x_1)})^{-1} \nabla_\theta f(x_1, \pi_{\theta^\star_\tau(x_1)})\| \\
&\leq \frac{L_f}{\underline{\sigma}} \cdot L_{V,2} \Big(\|x_1 - x_2\| + \|\pi^\star_\tau(x_1) - \pi^\star_\tau(x_2)\| \Big) \\
&\leq (\frac{L_f L_{V,2}}{\underline{\sigma}} + \frac{2L_f L_V L_{V,2}}{\underline{\sigma} C_L \tau}) \|x_1 - x_2\|,
\end{aligned}
\tag{132}
$$

where second inequality is due to Lemma 3, and the last inequality follows from an argument similar to the one in (131).

Similarly, for T_3

$$
\begin{aligned}
T_3 &\leq \|\nabla^2_{x,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)})\| \|\nabla^2_{\theta,\theta} J_\tau(x_1, \pi_{\theta^\star_\tau(x_1)})^{-1} - \nabla^2_{\theta,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)})^{-1}\| \|\nabla_\theta f(x_1, \pi_{\theta^\star_\tau(x_1)})\| \\
&\leq L_f L_V \|\nabla^2_{\theta,\theta} J_\tau(x_1, \pi_{\theta^\star_\tau(x_1)})^{-1}\| \|\nabla^2_{\theta,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)}) - \nabla^2_{\theta,\theta} J_\tau(x_1, \pi_{\theta^\star_\tau(x_1)})^{-1}\| \|\nabla^2_{\theta,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)})^{-1}\| \\
&\leq \frac{L_f L_V}{\underline{\sigma}^2} \cdot L_{V,2} \Big(\|x_1 - x_2\| + \|\pi^\star_\tau(x_1) - \pi^\star_\tau(x_2)\| \Big) \\
&\leq (\frac{L_f L_V L_{V,2}}{\underline{\sigma}^2} + \frac{2L_f L_V^2 L_{V,2}}{\underline{\sigma}^2 C_L \tau}) \|x_1 - x_2\|.
\end{aligned}
\tag{133}
$$

For the final term, we have

$$
\begin{aligned}
T_4 &\leq \|\nabla^2_{x,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)}) \nabla^2_{\theta,\theta} J_\tau(x_2, \pi_{\theta^\star_\tau(x_2)})^{-1}\| \|\nabla_\theta f(x_1, \pi_{\theta^\star_\tau(x_1)}) - \nabla_\theta f(x_2, \pi_{\theta^\star_\tau(x_2)})\| \\
&\leq \frac{L_V}{\underline{\sigma}} \|\nabla_\theta f(x_1, \pi_{\theta^\star_\tau(x_1)}) - \nabla_\theta f(x_2, \pi_{\theta^\star_\tau(x_2)})\| \\
&\leq \frac{L_V}{\underline{\sigma}} \cdot L_f \Big(\|x_1 - x_2\| + \|\pi^\star_\tau(x_1) - \pi^\star_\tau(x_2)\| \Big) \\
&\leq (\frac{L_f L_V}{\underline{\sigma}} + \frac{2L_f L_V^2}{\underline{\sigma} C_L \tau}) \|x_1 - x_2\|.
\end{aligned}
\tag{134}
$$

We combine (131)-(134)

$$
\|\nabla_x \Phi_\tau(x_1) - \nabla_x \Phi_\tau(x_2)\| \leq (L_f + \frac{2L_f L_V}{C_L \tau}) \|x_1 - x_2\| + (\frac{L_f L_{V,2}}{\underline{\sigma}} + \frac{2L_f L_V L_{V,2}}{\underline{\sigma} C_L \tau}) \|x_1 - x_2\|
$$

$$
\begin{aligned}
&+ (\frac{L_f L_V L_{V,2}}{\underline{\sigma}^2} + \frac{2L_f L_V^2 L_{V,2}}{\underline{\sigma}^2 C_L \tau})\|x_1 - x_2\| + (\frac{L_f L_V}{\underline{\sigma}} + \frac{2L_f L_V^2}{\underline{\sigma} C_L \tau})\|x_1 - x_2\| \\
&\leq (1 + \frac{2L_V}{C_L \tau})\Big(\frac{2L_f L_V}{C_L \tau} + \frac{2L_f L_V L_{V,2}}{\underline{\sigma} C_L \tau} + \frac{2L_f L_V^2 L_{V,2}}{\underline{\sigma}^2 C_L \tau} + \frac{2L_f L_V^2}{\underline{\sigma} C_L \tau}\Big)\|x_1 - x_2\|.
\end{aligned}
$$

Imposing the step size condition $\tau \leq \frac{2L_V}{C_L}$, we get

$$
\begin{aligned}
\|\nabla_x \Phi_\tau(x_1) - \nabla_x \Phi_\tau(x_2)\| &\leq \Big(\frac{4L_f L_V}{C_L \tau} + \frac{4L_f L_V L_{V,2}}{\underline{\sigma} C_L \tau} + \frac{4L_f L_V^2 L_{V,2}}{\underline{\sigma}^2 C_L \tau} + \frac{4L_f L_V^2}{\underline{\sigma} C_L \tau}\Big)\|x_1 - x_2\| \\
&\leq \frac{L_\Phi}{\tau}\|x_1 - x_2\|.
\end{aligned}
$$

■

E.11 Proof of Lemma 11

Let $\theta_\tau^\star(x)$ denote a parameter representing $\pi_\tau^\star(x)$ through the softmax function. Define for $\tau > 0$

$$
\nabla \Phi_\tau(x) = \nabla_x f(x, \pi_\tau^\star(x)) - \nabla_{x,\theta}^2 J_\tau(x, \pi_{\theta_\tau^\star(x)}) \nabla_{\theta,\theta}^2 J_\tau(x, \pi_{\theta_\tau^\star(x)})^{-1} \nabla_\theta f(x, \pi_{\theta_\tau^\star(x)}).
$$

We consider the following decomposition

$$
\|\nabla_x \Phi(x) - \nabla_x \Phi_{w,\tau}(x)\| \leq \|\nabla_x \Phi_\tau(x) - \nabla_x \Phi_{w,\tau}(x)\| + \|\nabla_x \Phi(x) - \nabla_x \Phi_\tau(x)\|. \tag{135}
$$

We first bound the first term of (135).

To derive the bound on $\|\nabla_x \Phi_\tau(x) - \nabla_x \Phi_{w,\tau}(x)\|$, we take an argument similar to Kwon et al. [2023][Lemma A.2], which we adapt to the case of a non-convex lower level objective. Note that λ in Kwon et al. [2023] plays the same role as our $1/w$. Kwon et al. [2023][Lemma A.2] is still valid without lower level convexity, with the lower bound on $\|\nabla_{\theta,\theta}^2 J_\tau(x, \pi_{\theta_\tau^\star(x)})\|$ changed from the strong convexity coefficient to $\underline{\sigma}$. This allows us to write

$$
\begin{aligned}
&\big\|\nabla_x \Phi_\tau(x) - \nabla_x \mathcal{L}_{w,\tau}(x, \pi_\theta) + \nabla_{x,\theta}^2 J_\tau(x, \pi_{\theta_\tau^\star(x)}) \nabla_{\theta,\theta}^2 J_\tau(x, \pi_{\theta_\tau^\star(x)})^{-1} \nabla_\theta \mathcal{L}_{w,\tau}(x, \pi_\theta)\big\| \\
&\leq \frac{2L_V}{\underline{\sigma}}\|\pi_\theta - \pi_\tau^\star(x)\|\Big(L_f + \frac{L_{V,2}}{w}\|\pi_\theta - \pi_\tau^\star(x)\|\Big).
\end{aligned}
$$

Recognizing $\nabla_\theta \mathcal{L}_{w,\tau}(x, \pi_{\theta_{w,\tau}^\star(x)}) = 0$, we have

$$
\begin{aligned}
\|\nabla_x \Phi_\tau(x) - \nabla_x \Phi_{w,\tau}(x)\| &= \|\nabla_x \Phi_\tau(x) - \nabla_x \mathcal{L}_{w,\tau}(x, \pi_{w,\tau}^\star(x))\| \\
&\leq \frac{2L_V}{\underline{\sigma}}\|\pi_{w,\tau}^\star(x) - \pi_\tau^\star(x)\|\Big(L_f + \frac{L_{V,2}}{w}\|\pi_{w,\tau}^\star(x) - \pi_\tau^\star(x)\|\Big) \\
&\leq \frac{2L_V}{\underline{\sigma}} \cdot \frac{2L_f w}{C_L \tau}\Big(L_f + \frac{L_{V,2}}{w} \cdot \frac{2L_f w}{C_L \tau}\Big) \\
&= \frac{4L_f L_V w}{C_L \underline{\sigma} \tau}(L_f + \frac{2L_f L_{V,2}}{C_L \tau}),
\end{aligned} \tag{136}
$$

where the second inequality follows from Lemma 7.

Next, we bound $\|\nabla_x \Phi_\tau(x) - \nabla_x \Phi(x)\|$.

$$
\begin{aligned}
\nabla_x \Phi_\tau(x) - \nabla_x \Phi(x) &= \underbrace{\nabla_x f(x, \pi_{\theta_\tau^\star(x)}) - \nabla_x f(x, \pi_{\theta^\star(x)})}_{T_1} \\
&+ \underbrace{\nabla_{x,\theta}^2 J(x, \pi_{\theta^\star(x)}) \nabla_{\theta,\theta}^2 J(x, \pi_{\theta^\star(x)})^{-1} \nabla_\theta f(x, \pi_{\theta^\star(x)}) - \nabla_{x,\theta}^2 J(x, \pi_{\theta_\tau^\star(x)}) \nabla_{\theta,\theta}^2 J(x, \pi_{\theta_\tau^\star(x)})^{-1} \nabla_\theta f(x, \pi_{\theta_\tau^\star(x)})}_{T_2} \\
&+ \underbrace{\nabla_{x,\theta}^2 J(x, \pi_{\theta_\tau^\star(x)}) \nabla_{\theta,\theta}^2 J(x, \pi_{\theta_\tau^\star(x)})^{-1} \nabla_\theta f(x, \pi_{\theta_\tau^\star(x)}) - \nabla_{x,\theta}^2 J_\tau(x, \pi_{\theta_\tau^\star(x)}) \nabla_{\theta,\theta}^2 J_\tau(x, \pi_{\theta_\tau^\star(x)})^{-1} \nabla_\theta f(x, \pi_{\theta_\tau^\star(x)})}_{T_3}.
\end{aligned} \tag{137}
$$

To treat T_1, we have from the Lipschitz continuity of f

$$\|T_1\| \leq L_f \|\pi_\tau^\star(x) - \pi^\star(x)\|.$$

For T_2,

$$\begin{aligned}
\|T_2\| &\leq \|\nabla^2_{x,\theta} J(x, \pi_{\theta^\star(x)}) - \nabla^2_{x,\theta} J(x, \pi_{\theta_\tau^\star(x)})\| \|\nabla^2_{\theta,\theta} J(x, \pi_{\theta^\star(x)})^{-1}\| \|\nabla_\theta f(x, \pi_{\theta^\star(x)})\| \\
&\quad + \|\nabla^2_{x,\theta} J(x, \pi_{\theta_\tau^\star(x)})\| \|\nabla^2_{\theta,\theta} J(x, \pi_{\theta^\star(x)})^{-1} - \nabla^2_{\theta,\theta} J(x, \pi_{\theta_\tau^\star(x)})^{-1}\| \|\nabla_\theta f(x, \pi_{\theta^\star(x)})\| \\
&\quad + \|\nabla^2_{x,\theta} J(x, \pi_{\theta_\tau^\star(x)})\| \|\nabla^2_{\theta,\theta} J(x, \pi_{\theta_\tau^\star(x)})^{-1}\| \|\nabla_\theta f(x, \pi_{\theta^\star(x)}) - \nabla_\theta f(x, \pi_{\theta_\tau^\star(x)})\| \\
&\leq L_{V,2} \|\pi_\tau^\star(x) - \pi^\star(x)\| \cdot \frac{1}{\underline{\sigma}} \cdot L_f \\
&\quad + L_V \cdot \|\nabla^2_{\theta,\theta} J(x, \pi_{\theta_\tau^\star(x)})^{-1}\| \|\nabla^2_{\theta,\theta} J(x, \pi_{\theta_\tau^\star(x)}) - \nabla^2_{\theta,\theta} J(x, \pi_{\theta^\star(x)})\| \|\nabla^2_{\theta,\theta} J(x, \pi_{\theta^\star(x)})^{-1}\| \\
&\quad + L_V \cdot \frac{1}{\underline{\sigma}} \cdot L_f \|\pi_\tau^\star(x) - \pi^\star(x)\| \\
&\leq \frac{L_f(L_V + L_{V,2})}{\underline{\sigma}} \|\pi_\tau^\star(x) - \pi^\star(x)\| + L_V \cdot \frac{1}{\underline{\sigma}} \cdot L_{V,2} \|\pi_\tau^\star(x) - \pi^\star(x)\| \cdot \frac{1}{\underline{\sigma}} \\
&\leq \frac{L_f L_V + L_f L_{V,2} + L_V L_{V,2}}{\underline{\sigma}^2} \|\pi_\tau^\star(x) - \pi^\star(x)\|.
\end{aligned}$$

We then bound T_3. Note that

$$J_\tau(x, \pi) - J(x, \pi) = \frac{\tau}{1-\gamma} \mathbb{E}_{s \sim d_\rho^\pi}[E(\pi, s)],$$

which is independent of x. This implies $\nabla^2_{x,\theta} J(x, \pi_\theta) = \nabla^2_{x,\theta} J_\tau(x, \pi_\theta)$. Using this relationship, we have

$$\begin{aligned}
\|T_3\| &= \left\| \nabla^2_{x,\theta} J(x, \pi_\theta) \Big(\nabla^2_{\theta,\theta} J(x, \pi_{\theta_\tau^\star(x)})^{-1} \nabla_\theta f(x, \pi_{\theta_\tau^\star(x)}) - \nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta_\tau^\star(x)})^{-1} \nabla_\theta f(x, \pi_{\theta_\tau^\star(x)}) \Big) \right\| \\
&\leq L_{V,2} L_f \|\nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta_\tau^\star(x)})^{-1}\| \|\nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta_\tau^\star(x)}) - \nabla^2_{\theta,\theta} J(x, \pi_{\theta_\tau^\star(x)})\| \|\nabla^2_{\theta,\theta} J(x, \pi_{\theta_\tau^\star(x)})^{-1}\| \\
&\leq L_f L_{V,2} \cdot \frac{1}{\underline{\sigma}} \cdot \frac{\tau}{1-\gamma} \|\nabla^2_{\theta,\theta} \mathbb{E}_{s \sim d_\rho^{\pi_\tau^\star(x)}}[E(\pi_\tau^\star(x), s)]\| \cdot \frac{1}{\underline{\sigma}} \\
&\leq \frac{L_f L_{V,2}(4 + 8\log|\mathcal{A}|)\tau}{(1-\gamma)^4 \underline{\sigma}^2},
\end{aligned}$$

where the third inequality follows from the fact that $\mathbb{E}_{s \sim d_\rho^\pi}[E(\pi, s)]$ is $\frac{4+8\log|\mathcal{A}|}{(1-\gamma)^3}$-Lipschitz (see, for example, Lemma 6 of Zeng et al. [2022]).

Collecting the bounds on T_1-T_3 and substituting them into (137),

$$\begin{aligned}
\|\nabla_x \Phi_\tau(x) - \nabla_x \Phi_{w,\tau}(x)\| &\leq \|T_1\| + \|T_2\| + \|T_3\| \\
&\leq \frac{L_f(L_V + 1) + L_f L_{V,2} + L_V L_{V,2}}{\underline{\sigma}^2} \|\pi_\tau^\star(x) - \pi^\star(x)\| + \frac{L_f L_{V,2}(4 + 8\log|\mathcal{A}|)\tau}{(1-\gamma)^4 \underline{\sigma}^2} \\
&\leq \frac{L_\star L_f(L_V + 1) + L_\star L_f L_{V,2} + L_\star L_V L_{V,2}}{\underline{\sigma}^2} \tau + \frac{L_f L_{V,2}(4 + 8\log|\mathcal{A}|)\tau}{(1-\gamma)^4 \underline{\sigma}^2} \\
&\leq \frac{L_\star L_f(L_V + 1) + L_\star L_f L_{V,2} + L_\star L_V L_{V,2} + L_f L_{V,2}(4 + 8\log|\mathcal{A}|)}{(1-\gamma)^4 \underline{\sigma}^2} \tau,
\end{aligned} \tag{138}$$

where the third inequality follows from Lemma 8.

Substituting (136) and (138) into (135),

$$\begin{aligned}
\|\nabla_x \Phi(x) - \nabla_x \Phi_{w,\tau}(x)\| &\leq \|\nabla_x \Phi_\tau(x) - \nabla_x \Phi_{w,\tau}(x)\| + \|\nabla_x \Phi(x) - \nabla_x \Phi_\tau(x)\| \\
&\leq \frac{4 L_f L_V w}{C_L \underline{\sigma} \tau} \left(L_f + \frac{2 L_f L_{V,2}}{C_L \tau}\right) \\
&\quad + \frac{L_\star L_f(L_V + 1) + L_\star L_f L_{V,2} + L_\star L_V L_{V,2} + L_f L_{V,2}(4 + 8\log|\mathcal{A}|)}{(1-\gamma)^4 \underline{\sigma}^2} \tau.
\end{aligned}$$

■

E.12 Proof of Lemma 12

Let $\theta^\star(x)$ and $\theta^\star_\tau(x)$ denote one of the softmax parameters that encodes $\pi^\star(x)$ and $\pi^\star_\tau(x)$. Recall the gradient expression $\nabla_x \Phi_\tau(x)$ from (9). We can similarly write

$$\nabla_x \Phi(x) = \nabla_x f(x, \pi^\star(x)) - \nabla^2_{x,\pi} J(x, \pi^\star(x)) \nabla^2_{\pi,\pi} J(x, \pi^\star(x))^{-1} \nabla_\pi f(x, \pi^\star(x)). \tag{139}$$

Combining (9) and (139),

$$\begin{aligned}
&\|\nabla_x \Phi_\tau(x) - \nabla_x \Phi(x)\| \\
&\leq \underbrace{\|\nabla_x f(x, \pi_{\theta^\star_\tau(x)}) - \nabla_x f(x, \pi_{\theta^\star(x)})\|}_{T_1} \\
&\quad + \underbrace{\|\nabla^2_{x,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)}) \nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)})^{-1} \nabla_\theta f(x, \pi_{\theta^\star_\tau(x)}) - \nabla^2_{x,\theta} J(x, \pi_{\theta^\star(x)}) \nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)})^{-1} \nabla_\theta f(x, \pi_{\theta^\star_\tau(x)})\|}_{T_2} \\
&\quad + \underbrace{\|\nabla^2_{x,\theta} J(x, \pi_{\theta^\star(x)}) \nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)})^{-1} \nabla_\theta f(x, \pi_{\theta^\star_\tau(x)}) - \nabla^2_{x,\theta} J(x, \pi_{\theta^\star(x)}) \nabla^2_{\theta,\theta} J(x, \pi_{\theta^\star(x)})^{-1} \nabla_\theta f(x, \pi_{\theta^\star_\tau(x)})\|}_{T_3} \\
&\quad + \underbrace{\|\nabla^2_{x,\theta} J(x, \pi_{\theta^\star(x)}) \nabla^2_{\theta,\theta} J(x, \pi_{\theta^\star(x)})^{-1} \nabla_\theta f(x, \pi_{\theta^\star_\tau(x)}) - \nabla^2_{x,\theta} J(x, \pi_{\theta^\star(x)}) \nabla^2_{\theta,\theta} J(x, \pi_{\theta^\star(x)})^{-1} \nabla_\theta f(x, \pi_{\theta^\star(x)})\|}_{T_4}.
\end{aligned} \tag{140}$$

We treat each term of (140) individually. First, we bound T_1 using the smoothness property of f

$$T_1 \leq L_f \|\pi^\star_\tau(x) - \pi^\star(x)\| \leq L_\star L_f \tau, \tag{141}$$

where the second inequality follows from Lemma 8.

We have $\|\nabla_\theta f(x, \pi_\theta)\| \leq L_f$ from Assumption 3 and $\|\nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)})^{-1}\| \leq \frac{1}{\underline{\sigma}}$ due to Assumption 2. This allows us to bound T_2 as follows

$$\begin{aligned}
T_2 &\leq \|\nabla^2_{x,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)}) - \nabla^2_{x,\theta} J(x, \pi_{\theta^\star(x)})\| \|\nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)})^{-1} \nabla_\theta f(x, \pi_{\theta^\star_\tau(x)})\| \\
&\leq \frac{L_f}{\underline{\sigma}} \Big(\|\nabla^2_{x,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)}) - \nabla^2_{x,\theta} J(x, \pi_{\theta^\star_\tau(x)})\| + \|\nabla^2_{x,\theta} J(x, \pi_{\theta^\star_\tau(x)}) - \nabla^2_{x,\theta} J(x, \pi_{\theta^\star(x)})\| \Big) \\
&= \frac{L_f}{\underline{\sigma}} \|\nabla^2_{x,\theta} J(x, \pi_{\theta^\star_\tau(x)}) - \nabla^2_{x,\theta} J(x, \pi_{\theta^\star(x)})\| \\
&\leq \frac{L_f}{\underline{\sigma}} \cdot L_{V,2} \|\pi_{\theta^\star_\tau(x)} - \pi_{\theta^\star(x)}\| \\
&= \frac{L_f}{\underline{\sigma}} \cdot L_{V,2} \|\pi^\star_\tau(x) - \pi^\star(x)\| \\
&\leq \frac{L_\star L_f L_{V,2} \tau}{\underline{\sigma}},
\end{aligned} \tag{142}$$

where the first equation is due to the fact that $J_\tau(x, \pi) - J(x, \pi)$ is independent of x, so the derivative with respect to x is zero. The last inequality plugs in (141).

Similarly, for T_3

$$\begin{aligned}
T_3 &\leq \|\nabla^2_{x,\theta} J(x, \pi_{\theta^\star(x)})\| \|\nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)})^{-1} - \nabla^2_{\theta,\theta} J(x, \pi_{\theta^\star(x)})^{-1}\| \|\nabla_\theta f(x, \pi_{\theta^\star_\tau(x)})\| \\
&\leq L_f L_V \|\nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)})^{-1}\| \|\nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)}) - \nabla^2_{\theta,\theta} J(x, \pi_{\theta^\star(x)})\| \|\nabla^2_{\theta,\theta} J(x, \pi_{\theta^\star_\tau(x)})^{-1}\| \\
&\leq \frac{L_f L_V}{\underline{\sigma}^2} \Big(\|\nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star_\tau(x)}) - \nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star(x)})\| + \|\nabla^2_{\theta,\theta} J_\tau(x, \pi_{\theta^\star(x)}) - \nabla^2_{\theta,\theta} J(x, \pi_{\theta^\star(x)})\| \Big) \\
&\leq \frac{L_f L_V}{\underline{\sigma}^2} \Big(L_{V,2} \|\pi^\star_\tau(x) - \pi^\star(x)\| + L_V \tau \Big) \\
&\leq \frac{L_\star L_f L_V L_{V,2} \tau}{\underline{\sigma}^2} + \frac{L_f L_V^2 \tau}{\underline{\sigma}^2},
\end{aligned} \tag{143}$$

where the third inequality again follows from Assumption 2, and the fourth inequality is due to (39) of Lemma 3 (note that $J_\tau(x,\pi) - J(x,\pi) = \tau\mathbb{E}_{s\sim d_\rho^\pi}[E(\pi,s)]$). The last inequality again plugs in (141).

For the final term, we have

$$\begin{aligned}
T_4 &\leq \|\nabla^2_{x,\theta}J(x,\pi_{\theta^\star(x)})\nabla^2_{\theta,\theta}J(x,\pi_{\theta^\star(x)})^{-1}\|\|\nabla_\theta f(x,\pi_{\theta^\star_\tau(x)}) - \nabla_\theta f(x,\pi_{\theta^\star(x)})\| \\
&\leq \frac{L_V}{\underline{\sigma}}\|\nabla_\theta f(x,\pi_{\theta^\star_\tau(x)}) - \nabla_\theta f(x,\pi_{\theta^\star(x)})\| \\
&\leq \frac{L_V}{\underline{\sigma}}\|\nabla_\pi f(x,\pi^\star_\tau(x)) - \nabla_\pi f(x,\pi^\star(x))\| \\
&\leq \frac{L_V}{\underline{\sigma}}\|\pi^\star_\tau(x) - \pi^\star(x)\| \\
&\leq \frac{L_\star L_V \tau}{\underline{\sigma}},
\end{aligned} \tag{144}$$

where the third inequality is due to the 1-Lipschitz continuity of softmax function, and the fourth inequality is due to Assumption 3.

Combining (141)-(144) leads to

$$\begin{aligned}
\|\nabla_x\Phi_\tau(x) - \nabla_x\Phi(x)\| &\leq L_\star L_f\tau + \frac{L_\star L_f L_{V,2}\tau}{\underline{\sigma}} + \frac{L_\star L_f L_V L_{V,2}\tau}{\underline{\sigma}^2} + \frac{L_\star L_V^2\tau}{\underline{\sigma}} + \frac{L_\star L_V\tau}{\underline{\sigma}} \\
&\leq \frac{L_\star L_f L_V^2 L_{V,2}\tau}{\underline{\sigma}^2}.
\end{aligned}$$

■

E.13 Proof of Lemma 13

As $\nabla_x J_{\tau_k}(x_k,\pi_{\theta_k}) - \nabla_x J_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)})$ does not depend on the randomness at iteration k, we have

$$\begin{aligned}
&\mathbb{E}[-\langle\nabla_x J_{\tau_k}(x_k,\pi_{\theta_k}) - \nabla_x J_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)}), x_{k+1}-x_k\rangle] \\
&= \zeta_k\mathbb{E}[\langle\nabla_x J_{\tau_k}(x_k,\pi_{\theta_k}) - \nabla_x J_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)}), \mathbb{E}[D_{w_k}(x_k,\pi_k,\pi_k^{\mathcal{L}},s_k,a_k,\bar{s}_k,\bar{a}_k,\xi_k)\mid\mathcal{F}_{k-1}]\rangle] \\
&= \zeta_k\mathbb{E}[\langle\nabla_x J_{\tau_k}(x_k,\pi_{\theta_k}) - \nabla_x J_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)}), \bar{D}_{w_k}(x_k,\pi_k,\pi_k^{\mathcal{L}})\rangle] \\
&= \zeta_k\mathbb{E}[\langle\nabla_x J_{\tau_k}(x_k,\pi_{\theta_k}) - \nabla_x J_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)}), \nabla_x\Phi_{w_k,\tau_k}(x_k)\rangle] \\
&\quad - \zeta_k\mathbb{E}[\langle\nabla_x J_{\tau_k}(x_k,\pi_{\theta_k}) - \nabla_x J_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)}), \bar{D}_{w_k}(x_k,\pi^\star_{\tau_k}(x_k),\pi^\star_{w_k,\tau_k}(x_k)) - \bar{D}_{w_k}(x_k,\pi_k,\pi_k^{\mathcal{L}})\rangle],
\end{aligned}$$

where the third equation is from (29).

By Young's inequality,

$$\begin{aligned}
&\mathbb{E}[-\langle\nabla_x J_{\tau_k}(x_k,\pi_{\theta_k}) - \nabla_x J_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)}), x_{k+1}-x_k\rangle] \\
&\leq \frac{C_L^2\alpha_k\tau_k^2}{128L_V^2}\mathbb{E}[\|\nabla_x J_{\tau_k}(x_k,\pi_{\theta_k}) - \nabla_x J_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)})\|^2] + \frac{32L_V^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2] \\
&\quad + \frac{C_L^2\alpha_k\tau_k^2}{128L_V^2}\mathbb{E}[\|\nabla_x J_{\tau_k}(x_k,\pi_{\theta_k}) - \nabla_x J_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)})\|^2] \\
&\quad + \frac{32L_V^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\bar{D}_{w_k}(x_k,\pi^\star_{\tau_k}(x_k),\pi^\star_{w_k,\tau_k}(x_k)) - \bar{D}_{w_k}(x_k,\pi_k,\pi_k^{\mathcal{L}})\|^2] \\
&\leq \frac{C_L^2\alpha_k\tau_k^2}{64L_V^2}\mathbb{E}[\|\nabla_x J_{\tau_k}(x_k,\pi_{\theta_k}) - \nabla_x J_{\tau_k}(x_k,\pi_{\theta^\star_{\tau_k}(x_k)})\|^2] + \frac{32L_V^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2] \\
&\quad + \frac{64L_D^2L_V^2\zeta_k^2}{C_L^2\alpha_k w_k^2\tau_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]
\end{aligned}$$

$$\leq \frac{C_L^2\alpha_k\tau_k^2}{64}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] + \frac{64L_D^2L_V^2\zeta_k^2}{C_L^2\alpha_k w_k^2\tau_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2] + \frac{32L_V^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2],$$

where the second inequality employs the Lipschitz continuity of $\bar{D}_{w_k}$ established in Lemma 6.

■

E.14 Proof of Lemma 14

As $\nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}}) - \nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)})$ does not depend on the randomness at iteration k, we have

$$\begin{aligned}
&\mathbb{E}[\langle\nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}}) - \nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)}), x_{k+1}-x_k\rangle]\\
&= -\zeta_k\mathbb{E}[\langle\nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}}) - \nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)}), \mathbb{E}[D_{w_k}(x_k,\pi_k,\pi_k^{\mathcal{L}},s_k,a_k,\bar{s}_k,\bar{a}_k,\xi_k)\mid\mathcal{F}_{k-1}]\rangle]\\
&= -\zeta_k\mathbb{E}[\langle\nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}}) - \nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)}), \bar{D}_{w_k}(x_k,\pi_k,\pi_k^{\mathcal{L}})\rangle]\\
&= -\zeta_k\mathbb{E}[\langle\nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}}) - \nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)}), \nabla_x\Phi_{w_k,\tau_k}(x_k)\rangle]\\
&\quad + \zeta_k\mathbb{E}[\langle\nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}}) - \nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)}),\\
&\qquad\qquad \bar{D}_{w_k}(x_k,\pi^\star_{\tau_k}(x_k),\pi^\star_{w_k,\tau_k}(x_k)) - \bar{D}_{w_k}(x_k,\pi_k,\pi_k^{\mathcal{L}})\rangle],
\end{aligned}$$

where the third equation is from (29).

By Young's inequality,

$$\begin{aligned}
&\mathbb{E}[\langle\nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}}) - \nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)}), x_{k+1}-x_k\rangle]\\
&\leq \frac{C_L^2\alpha_k\tau_k^2}{128L_L^2}\mathbb{E}[\|\nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}}) - \nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)})\|^2] + \frac{32L_L^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]\\
&\quad + \frac{C_L^2\alpha_k\tau_k^2}{128L_L^2}\mathbb{E}[\|\nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}}) - \nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)})\|^2]\\
&\quad + \frac{32L_L^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\bar{D}_{w_k}(x_k,\pi^\star_{\tau_k}(x_k),\pi^\star_{w_k,\tau_k}(x_k)) - \bar{D}_{w_k}(x_k,\pi_k,\pi_k^{\mathcal{L}})\|^2]\\
&\leq \frac{C_L^2\alpha_k\tau_k^2}{64L_L^2}\mathbb{E}[\|\nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta_k^{\mathcal{L}}}) - \nabla_x\mathcal{L}^{\text{reweight}}_{w_k,\tau_k}(x_k,\pi_{\theta^\star_{w_k,\tau_k}(x_k)})\|^2] + \frac{32L_L^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]\\
&\quad + \frac{64L_D^2L_L^2\zeta_k^2}{C_L^2\alpha_k w_k^2\tau_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]\\
&\leq \frac{C_L^2\alpha_k\tau_k^2}{64}\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2]\\
&\quad + \frac{64L_D^2L_L^2\zeta_k^2}{C_L^2\alpha_k w_k^2\tau_k^2}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2 + \|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2] + \frac{32L_L^2\zeta_k^2}{C_L^2\alpha_k\tau_k^2}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2],
\end{aligned}$$

where the second inequality employs the Lipschitz continuity of $\bar{D}_{w_k}$ established in Lemma 6.

■

E.15 Proof of Lemma 15

Within the proof of this lemma, we employ the shorthand notation $z_k = [x_k, \theta_k, \tau_k]$, $\ell(z_k) = V_{\tau_k}^{x_k,\pi_{\theta_k}}$, and

$$y_k = \hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}} + \beta_k\bar{G}_{\tau_k}(x_k,\theta_k,\hat{V}_k).$$

As V_τ^{x,π_θ} is smooth in x,θ,τ, we have from the mean-value theorem that there exists $z_{k+1}^m = mz_k + (1-m)z_{k+1}$ for some scalar $m\in[0,1]$ such that

$$\ell(z_k) - \ell(z_{k+1})$$

$$
\begin{aligned}
&= \nabla_z \ell(z_{k+1}^m)^\top \Big(z_k - z_{k+1}\Big) \\
&= \Big(\nabla_x V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \Big(x_k - x_{k+1}\Big) + \Big(\nabla_\theta V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \Big(\theta_k - \theta_{k+1}\Big) \\
&\quad + \Big(\nabla_\tau V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \Big(\tau_k - \tau_{k+1}\Big) \\
&= \zeta_k \Big(\nabla_x V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \bar{D}_{w_k}(x_k, \pi_{\theta_k}, \pi_{\theta_k^{\mathcal{L}}}) \\
&\quad + \zeta_k \Big(\nabla_x V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \Big(D_{w_k}(x_k, \pi_{\theta_k}, \pi_{\theta_k^{\mathcal{L}}}, s_k, a_k, \bar{s}_k, \bar{a}_k, \xi_k) - \bar{D}_{w_k}(x_k, \pi_{\theta_k}, \pi_{\theta_k^{\mathcal{L}}})\Big) \\
&\quad + \alpha_k \Big(\nabla_\theta V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k) \\
&\quad + \alpha_k \Big(\nabla_\theta V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \Big(F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s_k', \xi_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\Big) \\
&\quad + \Big(\nabla_\tau V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \Big(\tau_k - \tau_{k+1}\Big),
\end{aligned} \tag{145}
$$

where we denote $x_{k+1}^m = m x_k + (1-m) x_{k+1}, \theta_{k+1}^m = m\theta_k + (1-m)\theta_{k+1}, \tau_{k+1}^m = m\tau_k + (1-m)\tau_{k+1}$.

Plugging (145) into the cross term of interest, we have

$$
\begin{aligned}
&\langle \hat{V}_k - V_{\tau_k}^{x_k, \pi_{\theta_k}} + \beta_k \bar{G}_{\tau_k}(x_k, \theta_k, \hat{V}_k), V_{\tau_k}^{x_k, \pi_{\theta_k}} - V_{\tau_{k+1}}^{x_{k+1}, \pi_{\theta_{k+1}}} \rangle \\
&= \langle y_k, \ell(z_k) - \ell(x_{k+1}) \rangle \\
&= \alpha_k \langle y_k, \Big(\nabla_\theta V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k) \rangle \\
&\quad + \alpha_k \langle y_k, \Big(\nabla_\theta V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \Big(F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s_k', \xi_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\Big) \rangle \\
&\quad + \zeta_k \langle y_k, \Big(\nabla_x V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \bar{D}_{w_k}(x_k, \pi_{\theta_k}, \pi_{\theta_k^{\mathcal{L}}}) \rangle \\
&\quad + \zeta_k \langle y_k, \Big(\nabla_x V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \Big(D_{w_k}(x_k, \pi_{\theta_k}, \pi_{\theta_k^{\mathcal{L}}}, s_k, a_k, \bar{s}_k, \bar{a}, \xi_k) - \bar{D}_{w_k}(x_k, \pi_{\theta_k}, \pi_{\theta_k^{\mathcal{L}}})\Big) \rangle \\
&\quad + \langle y_k, \Big(\nabla_\tau V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \Big(\tau_k - \tau_{k+1}\Big) \rangle.
\end{aligned} \tag{146}
$$

We bound each term of (146) individually. First, by Young's inequality

$$
\begin{aligned}
&\alpha_k \langle y_k, \Big(\nabla_\theta V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\Big)^\top \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k) \rangle \\
&\le L_V \alpha_k \|y_k\| \|\bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\| \\
&\le \frac{(1-\gamma)\beta_k}{12} \|y_k\|^2 + \frac{3L_V^2 \alpha_k^2}{(1-\gamma)\beta_k} \|\bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\|^2 \\
&\le \frac{(1-\gamma)\beta_k}{12} \|y_k\|^2 + \frac{6L_V^2 \alpha_k^2}{(1-\gamma)\beta_k} \|\bar{F}_{0,\tau_k}(x_k, \theta_k, V_{\tau_k}^{x_k, \pi_{\theta_k}})\|^2 \\
&\quad + \frac{6L_V^2 \alpha_k^2}{(1-\gamma)\beta_k} \|\bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, V_{\tau_k}^{x_k, \pi_{\theta_k}})\|^2 \\
&= \frac{(1-\gamma)\beta_k}{12} \|y_k\|^2 + \frac{6L_V^2 \alpha_k^2}{(1-\gamma)\beta_k} \|\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2 \\
&\quad + \frac{6L_V^2 \alpha_k^2}{(1-\gamma)\beta_k} \|\bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, V_{\tau_k}^{x_k, \pi_{\theta_k}})\|^2 \\
&\le \frac{(1-\gamma)\beta_k}{12} \|y_k\|^2 + \frac{6L_V^2 \alpha_k^2}{(1-\gamma)\beta_k} \|\nabla_\theta J_{\tau_k}(x_k, \pi_{\theta_k})\|^2 + \frac{6L_V^2 L_F^2 \alpha_k^2}{(1-\gamma)\beta_k} \varepsilon_k^V,
\end{aligned} \tag{147}
$$

where the equation follows from the fact that $\bar{F}_{0,\tau}(x, \theta, V_\tau^{x,\pi_\theta}) = \nabla_\theta J_\tau(x, \pi_\theta)$ for any x, θ, τ (see (30)), and the final inequality is due to the Lipschitz continuity of $\bar{F}_{0,\tau_k}$.

For the second term of (146), we take the expectation

$$
\begin{aligned}
&\alpha_k \mathbb{E}\left[\langle y_k, \left(\nabla_\theta V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\right)^\top \left(F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s_k', \xi_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\right)\rangle\right] \\
&= \alpha_k \mathbb{E}[\langle y_k, \left(\nabla_\theta V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}} - \nabla_\theta V_{\tau_k}^{x_k, \pi_{\theta_k}}\right)^\top \left(F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s_k', \xi_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\right)] \\
&\quad + \alpha_k \mathbb{E}[\langle y_k, \left(\nabla_\theta V_{\tau_k}^{x_k, \pi_{\theta_k}}\right)^\top \left(F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s_k', \xi_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\right)] \\
&= \alpha_k \mathbb{E}[\langle y_k, \left(\nabla_\theta V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}} - \nabla_\theta V_{\tau_k}^{x_k, \pi_{\theta_k}}\right)^\top \left(F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s_k', \xi_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\right)] \\
&\leq 2B_F \alpha_k \mathbb{E}[\|y_k\| \|\nabla_\theta V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}} - \nabla_\theta V_{\tau_k}^{x_k, \pi_{\theta_k}}\|] \\
&\leq 2B_F L_V \alpha_k \mathbb{E}[\|y_k\| \left(\|\pi_{\theta_{k+1}^m} - \pi_{\theta_k}\| + \|x_{k+1}^m - x_k\| + |\tau_{k+1}^m - \tau_k|\right)] \\
&\leq 2B_F L_V \alpha_k \mathbb{E}[\|y_k\| \left(\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\| + \|x_{k+1} - x_k\| + |\tau_{k+1} - \tau_k|\right)] \\
&\leq 2B_F L_V \alpha_k \mathbb{E}[\|y_k\|] \left(\alpha_k B_F + \frac{\zeta_k B_D}{w_k} + \frac{8\tau_k}{3(k+1)}\right) \\
&\leq 2B_F L_V \alpha_k \mathbb{E}[\|y_k\|] \cdot \frac{3\tau_0 B_F \alpha_k}{\alpha_0} \\
&\leq \frac{3B_F^2 L_V \tau_0 \alpha_k^2}{\alpha_0} \mathbb{E}[\|y_k\|^2] + \frac{3B_F^2 L_V \tau_0 \alpha_k^2}{\alpha_0},
\end{aligned} \tag{148}
$$

where the fifth inequality follows from the step size conditions $\zeta_k \leq \frac{B_F \alpha_k w_k}{B_D}$ and $\alpha_0 \leq \min\{\tau_0, \frac{3B_F}{8}\}$, and the second equation follows from

$$
\begin{aligned}
&\mathbb{E}[\langle y_k, \left(\nabla_\theta V_{\tau_k}^{x_k, \pi_{\theta_k}}\right)^\top \left(F_{0,\tau_k}(x_k, \theta_k, \hat{V}_k, s_k, a_k, s_k', \xi_k) - \bar{F}_{0,\tau_k}(x_k, \theta_k, \hat{V}_k)\right)] \\
&= \mathbb{E}[\langle y_k, \left(V_{\tau_k}^{x_k, \pi_{\theta_k}}\right)^\top \mathbb{E}\Big[F(\theta_k, \omega_k, \hat{\mu}_k, \hat{V}_{f,k}, s_k, a_k, b_k, s_k') - \bar{F}(\theta_k, \omega_k, \hat{\mu}_k, \hat{V}_{f,k}) \mid \mathcal{F}_{k-1}\Big]] \\
&= 0.
\end{aligned}
$$

The third term of (146) can be bounded similar to the first term,

$$
\begin{aligned}
&\zeta_k \langle y_k, \left(\nabla_x V_{\tau_{k+1}^m}^{x_{k+1}^m, \pi_{\theta_{k+1}^m}}\right)^\top \bar{D}_{w_k}(x_k, \pi_{\theta_k}, \pi_{\theta_k^{\mathcal{L}}})\rangle \\
&\leq L_V \zeta_k \|y_k\| \|\bar{D}_{w_k}(x_k, \pi_{\theta_k}, \pi_{\theta_k^{\mathcal{L}}})\| \\
&\leq \frac{(1-\gamma)\beta_k}{12} \|y_k\|^2 + \frac{3L_V^2 \zeta_k^2}{(1-\gamma)\beta_k} \|\bar{D}_{w_k}(x_k, \pi_{\theta_k}, \pi_{\theta_k^{\mathcal{L}}})\|^2 \\
&\leq \frac{(1-\gamma)\beta_k}{12} \|y_k\|^2 + \frac{6L_V^2 \zeta_k^2}{(1-\gamma)\beta_k} \|\bar{D}_{w_k}(x_k, \pi_{\tau_k}^\star(x_k), \pi_{w_k,\tau_k}^\star(x_k))\|^2 \\
&\quad + \frac{6L_V^2 \zeta_k^2}{(1-\gamma)\beta_k} \|\bar{D}_{w_k}(x_k, \pi_{\theta_k}, \pi_{\theta_k^{\mathcal{L}}}) - \bar{D}_{w_k}(x_k, \pi_{\tau_k}^\star(x_k), \pi_{w_k,\tau_k}^\star(x_k))\|^2 \\
&= \frac{(1-\gamma)\beta_k}{12} \|y_k\|^2 + \frac{6L_V^2 \zeta_k^2}{(1-\gamma)\beta_k} \|\nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2 \\
&\quad + \frac{6L_V^2 \zeta_k^2}{(1-\gamma)\beta_k} \|\bar{D}_{w_k}(x_k, \pi_{\theta_k}, \pi_{\theta_k^{\mathcal{L}}}) - \bar{D}_{w_k}(x_k, \pi_{\tau_k}^\star(x_k), \pi_{w_k,\tau_k}^\star(x_k))\|^2 \\
&\leq \frac{(1-\gamma)\beta_k}{12} \|y_k\|^2 + \frac{6L_V^2 \zeta_k^2}{(1-\gamma)\beta_k} \|\nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2 \\
&\quad + \frac{12L_V^2 L_D^2 \zeta_k^2}{(1-\gamma)\beta_k} \|\pi_k - \pi_{\tau_k}^\star(x_k)\|^2 + \frac{12L_V^2 L_D^2 \zeta_k^2}{(1-\gamma)\beta_k} \|\pi_k^{\mathcal{L}} - \pi_{w_k,\tau_k}^\star(x_k)\|^2,
\end{aligned} \tag{149}
$$

where the equation is due to the condition in (29).

For the fourth term of (146), we again take the expectation and use the technique in (148)

$$
\begin{aligned}
&\zeta_k\mathbb{E}[\langle y_k, \Big(\nabla_x V_{\tau_{k+1}^m}^{x_{k+1}^m,\pi_{\theta_{k+1}^m}}\Big)^\top \Big(D_{w_k}(x_k,\pi_{\theta_k},\pi_{\theta_k^{\mathcal{L}}},s_k,a_k,\bar{s}_k,\bar{a},\xi_k) - \bar{D}_{w_k}(x_k,\pi_{\theta_k},\pi_{\theta_k^{\mathcal{L}}})\Big)\rangle]\\
&= \zeta_k\mathbb{E}[\langle y_k, \Big(\nabla_x V_{\tau_{k+1}^m}^{x_{k+1}^m,\pi_{\theta_{k+1}^m}} - \nabla_x V_{\tau_k}^{x_k,\pi_{\theta_k}}\Big)^\top \Big(D_{w_k}(x_k,\pi_{\theta_k},\pi_{\theta_k^{\mathcal{L}}},s_k,a_k,\bar{s}_k,\bar{a},\xi_k) - \bar{D}_{w_k}(x_k,\pi_{\theta_k},\pi_{\theta_k^{\mathcal{L}}})\Big)]\\
&\quad + \zeta_k\mathbb{E}[\langle y_k, \Big(\nabla_x V_{\tau_k}^{x_k,\pi_{\theta_k}}\Big)^\top \Big(D_{w_k}(x_k,\pi_{\theta_k},\pi_{\theta_k^{\mathcal{L}}},s_k,a_k,\bar{s}_k,\bar{a},\xi_k) - \bar{D}_{w_k}(x_k,\pi_{\theta_k},\pi_{\theta_k^{\mathcal{L}}})\Big)]\\
&= \zeta_k\mathbb{E}[\langle y_k, \Big(\nabla_x V_{\tau_{k+1}^m}^{x_{k+1}^m,\pi_{\theta_{k+1}^m}} - \nabla_x V_{\tau_k}^{x_k,\pi_{\theta_k}}\Big)^\top \Big(D_{w_k}(x_k,\pi_{\theta_k},\pi_{\theta_k^{\mathcal{L}}},s_k,a_k,\bar{s}_k,\bar{a},\xi_k) - \bar{D}_{w_k}(x_k,\pi_{\theta_k},\pi_{\theta_k^{\mathcal{L}}})\Big)]\\
&\le 2B_D\zeta_k\mathbb{E}[\|y_k\|\|\nabla_x V_{\tau_{k+1}^m}^{x_{k+1}^m,\pi_{\theta_{k+1}^m}} - \nabla_x V_{\tau_k}^{x_k,\pi_{\theta_k}}\|]\\
&\le 2B_D L_V\zeta_k\mathbb{E}[\|y_k\|\Big(\|\pi_{\theta_{k+1}^m} - \pi_{\theta_k}\| + \|x_{k+1}^m - x_k\| + |\tau_{k+1}^m - \tau_k|\Big)]\\
&\le 2B_D L_V\zeta_k\mathbb{E}[\|y_k\|\Big(\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\| + \|x_{k+1} - x_k\| + |\tau_{k+1} - \tau_k|\Big)]\\
&\le 2B_D L_V\zeta_k\mathbb{E}[\|y_k\|]\Big(\alpha_k B_F + \frac{\zeta_k B_D}{w_k} + \frac{8\tau_k}{3(k+1)}\Big)\\
&\le 2B_D L_V\zeta_k\mathbb{E}[\|y_k\|]\cdot\frac{3\tau_0 B_F\alpha_k}{\alpha_0}\\
&\le \frac{3B_F^2 L_V\tau_0\alpha_k^2}{\alpha_0}\mathbb{E}[\|y_k\|^2] + \frac{3B_F^2 L_V\tau_0\alpha_k^2}{\alpha_0},
\end{aligned}
\tag{150}
$$

where the last inequality follows from the step size condition $\frac{\zeta_k}{\alpha_k} \le \frac{B_F}{B_D}$.

For the fifth term of (146), we apply Lemma 2

$$
\begin{aligned}
\langle y_k, \Big(\nabla_\tau V_{\tau_{k+1}^m}^{x_{k+1}^m,\pi_{\theta_{k+1}^m}}\Big)^\top \Big(\tau_k - \tau_{k+1}\Big)\rangle &\le L_V\|y_k\||\tau_k - \tau_{k+1}|\\
&\le L_V\|y_k\|\cdot\frac{8\tau_k}{3(k+1)}\\
&\le \frac{(1-\gamma)\beta_k}{6}\|y_k\|^2 + \frac{32L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2}.
\end{aligned}
\tag{151}
$$

Collecting the results in (147)-(151) and substituting into (146),

$$
\begin{aligned}
&\mathbb{E}[\langle \hat{V}_k - V_{\tau_k}^{x_k,\pi_{\theta_k}} + \beta_k\bar{G}_{\tau_k}(x_k,\theta_k,\hat{V}_k), V_{\tau_k}^{x_k,\pi_{\theta_k}} - V_{\tau_{k+1}}^{x_{k+1},\pi_{\theta_{k+1}}}\rangle]\\
&\le \frac{(1-\gamma)\beta_k}{12}\mathbb{E}[\|y_k\|^2] + \frac{6L_V^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2] + \frac{6L_V^2L_F^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\varepsilon_k^V]\\
&\quad + \frac{3B_F^2L_V\tau_0\alpha_k^2}{\alpha_0}\mathbb{E}[\|y_k\|^2] + \frac{3B_F^2L_V\tau_0\alpha_k^2}{\alpha_0}\\
&\quad + \frac{(1-\gamma)\beta_k}{12}\mathbb{E}[\|y_k\|^2] + \frac{6L_V^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]\\
&\quad + \frac{12L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k - \pi_{\tau_k}^\star(x_k)\|^2] + \frac{12L_V^2L_D^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi_{w_k,\tau_k}^\star(x_k)\|^2]\\
&\quad + \frac{3B_F^2L_V\tau_0\alpha_k^2}{\alpha_0}\mathbb{E}[\|y_k\|^2] + \frac{3B_F^2L_V\tau_0\alpha_k^2}{\alpha_0}\\
&\quad + \frac{(1-\gamma)\beta_k}{6}\mathbb{E}[\|y_k\|^2] + \frac{32L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2}\\
&\le \frac{(1-\gamma)\beta_k}{3}\mathbb{E}[\|y_k\|^2] + \frac{6B_F^2L_V\tau_0\alpha_k^2}{\alpha_0}\mathbb{E}[\|y_k\|^2] + \frac{6L_V^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_x\Phi_{w_k,\tau_k}(x_k)\|^2]\\
&\quad + \frac{6L_V^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2] + \frac{6L_V^2L_F^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\varepsilon_k^V]
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{12L_V^2 L_D^2 \zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] + \frac{12L_V^2 L_D^2 \zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2] \\
&+ \frac{6B_F^2 L_V \tau_0 \alpha_k^2}{\alpha_0} + \frac{32L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2} \\
\leq\ & \frac{(1-\gamma)\beta_k}{2}\mathbb{E}[\|y_k\|^2] + \frac{6L_V^2\zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_x \Phi_{w_k,\tau_k}(x_k)\|^2] + \frac{6L_V^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\nabla_\theta J_{\tau_k}(x_k,\pi_{\theta_k})\|^2] + \frac{6L_V^2 L_F^2\alpha_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\varepsilon_k^V] \\
&+ \frac{12L_V^2 L_D^2 \zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k - \pi^\star_{\tau_k}(x_k)\|^2] + \frac{12L_V^2 L_D^2 \zeta_k^2}{(1-\gamma)\beta_k}\mathbb{E}[\|\pi_k^{\mathcal{L}} - \pi^\star_{w_k,\tau_k}(x_k)\|^2] \\
&+ \frac{6B_F^2 L_V \tau_0 \alpha_k^2}{\alpha_0} + \frac{32L_V^2\tau_k^2}{3(1-\gamma)\beta_k(k+1)^2},
\end{aligned}
$$

where the terms are combined in the last inequality under the step size condition $\alpha_k \leq \beta_k$ and $\frac{\alpha_k}{\beta_k} \leq \frac{1-\gamma}{36B_F^2 L_V \tau_0}$.

■

F Simulation Details – GridWorld Goal Placement

The lower-level MDP is defined on a 10×10 grid, where each state corresponds to a position on the grid. At every state, the agent can choose from four possible actions: $\mathcal{A} = \{\textbf{UP}, \textbf{DOWN}, \textbf{LEFT}, \textbf{RIGHT}\}$. Each action moves the agent to the adjacent cell in the corresponding direction. If the current position lies on the boundary and the action would move the agent outside the grid, the state remains unchanged. The upper-level decision variable x sets a goal state for the lower-level problem, shown by the flag in Figure 2. Let $x = (\text{coor}_1(\text{goal}), \text{coor}_2(\text{goal}))$. The reward of state $s = (\text{coor}_1(s), \text{coor}_2(s))$ is

$$
\begin{aligned}
r(s) = &-\Big(\text{coor}_1(s) - \text{coor}_1(\text{goal})\Big)^2 \\
&- \Big(\text{coor}_2(s) - \text{coor}_2(\text{goal})\Big)^2.
\end{aligned}
$$

We choose the upper-level objective f such that $f(x,\pi)$ when x is close to the center of the grid, indicated by green circle in Figure 2 and that π has bias towards DOWN and RIGHT actions. Specifically, with the coordinate of the center cell denoted by $(\text{coor}_1(\text{center}), \text{coor}_2(\text{center}))$, we consider

$$
\begin{aligned}
f(x,\pi) = &\Big(\text{coor}_1(\text{goal}) - \text{coor}_1(\text{center})\Big)^2 \\
&+ \Big(\text{coor}_2(\text{goal}) - \text{coor}_2(\text{center})\Big)^2 \\
&- \lambda \sum_s \Big(\pi(\text{DOWN} \mid s) + \pi(\text{RIGHT} \mid s)\Big),
\end{aligned}
$$

where λ is a non-negative parameter.

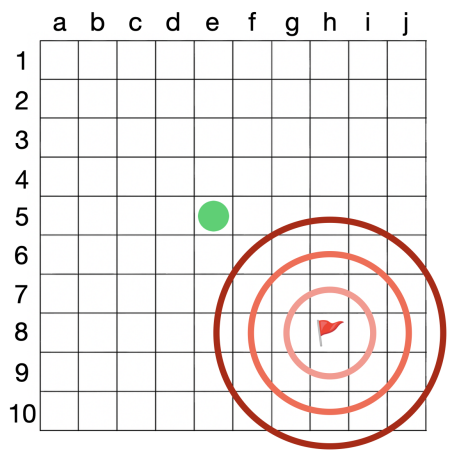

Figure 2: GridWorld Illustration. The red flag is the goal in the lower-level MDP set by the upper-level decision variable. A state further away from the goal incurs a negative reward with higher magnitude. The green circle indicates the center of the grid, which defines a component of the upper-level objective.

We set λ sufficiently large, under which the optimal solution to the bi-level problem is to set the goal on the bottom right corner. This is indeed the learned solution from Algorithm 1.