Softmax Policy Gradient Methods Can Take Exponential Time to Converge

Gen Li* Tsinghua | Yuting Wei† CMU | Yuejie Chi‡ CMU | Yuantao Gu* Tsinghua | Yuxin Chen§ Princeton

February 22, 2021

Abstract

The softmax policy gradient (PG) method, which performs gradient ascent under softmax policy parameterization, is arguably one of the de facto implementations of policy optimization in modern reinforcement learning. For γ-discounted infinite-horizon tabular Markov decision processes (MDPs), remarkable progress has recently been achieved towards establishing global convergence of softmax PG methods in finding a near-optimal policy. However, prior results fall short of delineating clear dependencies of convergence rates on salient parameters such as the cardinality of the state space $\mathcal{S}$ and the effective horizon $\frac{1}{1-\gamma}$, both of which could be excessively large. In this paper, we deliver a pessimistic message regarding the iteration complexity of softmax PG methods, despite assuming access to exact gradient computation. Specifically, we demonstrate that softmax PG methods can take exponential time — in terms of $|\mathcal{S}|$ and $\frac{1}{1-\gamma}$ — to converge, even in the presence of a benign policy initialization and an initial state distribution amenable to exploration. This is accomplished by characterizing the algorithmic dynamics over a carefully-constructed MDP containing only three actions. Our exponential lower bound hints at the necessity of carefully adjusting update rules or enforcing proper regularization in accelerating PG methods.

Keywords: policy gradient methods, exponential lower bounds, softmax parameterization

Contents

*Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.

†Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

‡Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

§Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544, USA.

1 Introduction

Despite their remarkable empirical popularity in modern reinforcement learning (Mnih et al., 2015; Silver et al., 2016), theoretical underpinnings of policy gradient (PG) methods and their variants (Kakade, 2002; Konda and Tsitsiklis, 2000; Peters and Schaal, 2008; Sutton et al., 2000; Williams, 1992) remain severely obscured. Due to the nonconcave nature of value function maximization induced by complicated dynamics of the environments, it is in general highly challenging to pinpoint the computational efficacy of PG methods in finding a near-optimal policy. Motivated by their practical importance, a recent strand of work sought to make progress towards demystifying the effectiveness of policy gradient type methods (e.g., Agarwal et al. (2019); Bhandari (2020); Bhandari and Russo (2019, 2020); Cen et al. (2020); Fazel et al. (2018); Lan (2021); Mei et al. (2020b); Shani et al. (2019); Yang et al. (2020); Zhang et al. (2020a,c); Zhao et al. (2021)), focusing primarily on canonical settings such as tabular Markov decision processes (MDPs) for discrete-state problems and linear quadratic regulators for continuous-state problems.

The current paper studies PG methods with softmax parameterization — commonly referred to as *softmax policy gradient* methods — which are among the de facto implementations of PG methods in practice. An intriguing theoretical result was recently obtained by the work Agarwal et al. (2019), which established asymptotic global convergence of softmax PG methods for infinite-horizon γ-discounted tabular MDPs. Subsequently, Mei et al. (2020b) strengthened the theory by demonstrating that softmax PG methods are capable of finding an ε-optimal policy within an order of $1/\varepsilon$ iterations (see Table 1 for the precise form). While these results take an important step towards understanding softmax PG methods, caution needs to be exercised before declaring fast convergence of the algorithms. In particular, the iteration complexity derived by Mei et al. (2020b) falls short of delineating clear dependencies on important salient parameters of the MDP, such as the dimension of the state space $\mathcal{S}$ and the effective horizon $1/(1-\gamma)$. These parameters are, more often than not, enormous in contemporary RL applications, and might play a pivotal role in determining the computational complexity of softmax PG methods.

Additionally, it is worth noting that existing literature largely concentrated on developing algorithm-dependent upper bounds on the iteration complexities. Nevertheless, upper bounds on the computational complexities of distinct algorithms — without certifying the tightness of them — cannot be directly compared for benchmarking purposes in a decisive manner. As a concrete example, it is of practical interest to benchmark softmax PG methods against natural policy gradient (NPG) methods with softmax parameterization, the latter of which is a variant of policy optimization lying underneath several mainstream RL algorithms such as *proximal policy optimization* (PPO) (Schulman et al., 2017) and *trust region policy optimization* (TRPO) (Schulman et al., 2015). While it is tempting to claim superiority of NPG methods over softmax PG methods — given the appealing convergence properties of NPG methods (Agarwal et al., 2019) (see Table 1) — existing theory fell short to reach such a conclusion due to the lack of explicit and meaningful convergence lower bounds for softmax PG methods.

The above considerations lead to a natural question that we aim to address in the present paper:

Can we develop a lower bound on the iteration complexity of softmax PG methods that reflects explicit dependency on salient parameters of the MDP of interest?

1.1 Main result

As an attempt to address the question posed above, our investigation delivers a somewhat surprising message that can be described in words as follows:

Softmax PG methods can take (super-)exponential time to converge, even in the presence of a benign initialization and an initial state distribution amenable to exploration.

Our finding, which is concerned with a discounted infinite-horizon tabular MDP, is formally stated in the following theorem. Here and throughout, $|\mathcal{S}|$ denotes the size of the state space $\mathcal{S}$, $0<\gamma<1$ stands for the discount factor, $V^\star$ indicates the optimal value function, $\eta>0$ is the learning rate or stepsize, whereas $V^{(t)}$ represents the value function estimate of softmax PG methods in the t-th iteration; see Section 2 for formal descriptions of this set of notation.

Theorem 1. *Assume that the softmax PG method adopts a uniform initial state distribution, a uniform policy initialization, and has access to exact gradient computation. There exist universal constants $c_1, c_2, c_3 > 0$ such that: for any $0.96 < \gamma < 1$ and $|\mathcal{S}| \geq c_3(1-\gamma)^{-6}$, one can find a γ-discounted MDP with state space $\mathcal{S}$ that takes the softmax PG method at least*

$$\frac{c_1}{\eta}|\mathcal{S}|^{2^{\frac{c_2}{1-\gamma}}} \tag{1}$$

iterations to reach $\|V^\star - V^{(t)}\|_\infty \leq 0.15$.

Remark 1. The MDP we construct contains at most three actions for each state.

For simplicity of presentation, Theorem 1 is stated for the long-effective-horizon regime where $\gamma > 0.96$; it continues to hold when $\gamma > c_0$ for some smaller constant $c_0 > 0$. Our result is obtained based on constructing a hard MDP instance — which is a properly augmented chain-like MDP — for which softmax PG methods converge extremely slowly even when perfect model specification is available. Several implications of our result are in order.

Comparisons with prior results. Table 1 provides an extensive comparison of the iteration complexities — including both upper and lower bounds — of PG and NPG methods under softmax parameterization. As suggested by our result, the iteration complexity $O(\mathcal{C}_{\mathsf{spg}}^2(\mathcal{M})\frac{1}{\varepsilon})$ derived in Mei et al. (2020b) (see Table 1) might not be as rosy as it seems for problems with large state space and long effective horizon; in fact, the crucial quantity $\mathcal{C}_{\mathsf{spg}}(\mathcal{M})$ therein could scale in a prohibitive manner with both $|\mathcal{S}|$ and $\frac{1}{1-\gamma}$. Mei et al.

algorithm	iteration complexity	reference
softmax PG upper bound	asymptotic	Agarwal et al. (2019, Thm. 5.1)
softmax PG upper bound	$O\Big(\mathcal{C}^2_{\mathsf{spg}}(\mathcal{M})\Big\|\frac{d_\mu^{\pi^\star}}{\mu}\Big\|_\infty^2\Big\|\frac{1}{\mu}\Big\|_\infty\frac{\lvert\mathcal{S}\rvert}{(1-\gamma)^2\varepsilon}\Big)$	Mei et al. (2020b, Thm. 4)
softmax NPG upper bound	$O\big(\frac{1}{(1-\gamma)^2\varepsilon}\big)$	Agarwal et al. (2019, Thm. 5.3)
softmax PG lower bound	$\frac{(1-\gamma)^5\Delta_\star^2}{12\varepsilon}$	Mei et al. (2020b, Thm. 10)
softmax PG lower bound	$\lvert\mathcal{S}\rvert^{2^{\Omega(\frac{1}{1-\gamma})}}$	**this work**

Table 1: Upper and lower bounds on the iteration complexities of PG and NPG methods with softmax parameterization in finding an ε-optimal policy obeying $\|V^\star - V^{(t)}\|_\infty \le \varepsilon \le 0.15$. We assume exact gradient evaluation, and hide the dependencies that are logarithmic on problem parameters. Here, μ denotes the initial state distribution, $\big\|d_\mu^{\pi^\star}/\mu\big\|_\infty$ is the distribution mismatch coefficient, $a^\star(s)$ is the optimal action at state s according to $\pi^\star$, $\mathcal{C}_{\mathsf{spg}}(\mathcal{M}) := \big(\inf_{s\in\mathcal{S}}\inf_{t\ge 1}\pi^{(t)}(a^\star(s)\,|\,s)\big)^{-1}$ is a quantity depending on both the PG trajectory and salient MDP parameters, whereas $\Delta_\star := \min_{s\in\mathcal{S},a\ne a^\star(s)}\big\{Q^\star(s,a^\star(s)) - Q^\star(s,a)\big\}$ is the optimality gap w.r.t. the optimal Q-function $Q^\star$.

(2020b) also developed a lower bound on the iteration complexity of softmax PG methods, which falls short of capturing the influence of the state space dimension and might become smaller than 1 unless ε is very small (e.g., $\varepsilon \lesssim (1-\gamma)^3$) for problems with long effective horizons. In addition, Mei et al. (2020a) provided some interesting evidence that a poorly-initialized softmax PG algorithm can get stuck at suboptimal policies for a single-state MDP (i.e., the bandit problem). This result, however, fell short of providing a complete runtime analysis and did not look into the influence of a large state space. By contrast, our theory reveals that softmax PG methods can take exponential time to reach even a moderate accuracy level.

Benchmarking with NPG methods. Our algorithm-specific lower bound suggests that softmax PG methods — in their vanilla form — might take a prohibitively long time to converge when the state space and effective horizon are large. This is in stark contrast to the convergence rate of NPG methods, whose iteration complexity is dimension-free and scales only polynomially with the effective horizon. Consequently, our results shed light on the practical superiority of NPG-based algorithms such as PPO (Schulman et al., 2017) and TRPO (Schulman et al., 2015).

Crux of our design. As we shall elucidate momentarily in Section 3, our exponential lower bound is obtained through analyzing the trajectory of softmax PG methods on a carefully-designed MDP instance with no more than 3 actions per state, when a uniform initialization scheme and a uniform initial state distribution are adopted. Our construction underscores the critical challenge of *credit assignments* (Sutton, 1984) in RL compounded by the presence of delayed rewards, long horizon, and intertwined interactions across states.

1.2 Other related work

Non-asymptotic analysis of (natural) policy gradient methods. Moving beyond tabular MDPs, finite-time convergence guarantees of PG methods have recently been studied for control problems (e.g., Fazel et al. (2018); Jansch-Porto et al. (2020); Mohammadi et al. (2019); Tu and Recht (2019); Zhang et al. (2019)), MDPs with linear function approximation (e.g., Agarwal et al. (2019); Cai et al. (2019)), MDPs with neural network approximation (e.g., Agazzi and Lu (2021); Liu et al. (2019); Wang et al. (2019)), constrained MDPs (e.g., Ding et al. (2020); Xu et al. (2020a)), and their use in actor-critic methods (e.g., Wu et al. (2020); Xu et al. (2020b)). In addition, non-asymptotic convergence guarantees of NPG methods and their variants have been derived in Agarwal et al. (2019); Bhandari and Russo (2020); Cen et al. (2020);

Khodadadian et al. (2021); Liu et al. (2020); Shani et al. (2019); Xie et al. (2020); Zhang et al. (2021); Zhao et al. (2021).

Other policy parameterizations. In addition to softmax parameterization, several other policy parameterization schemes have also been investigated in the context of policy optimization and reinforcement learning at large. For example, Agarwal et al. (2019); Zhang et al. (2020b) studied the converge of projected PG methods with direct parameterization, Asadi and Littman (2017) introduced the so-called mallow parameterization, while Mei et al. (2020a) studied the escort parameterization. Part of these parameterizations were proposed in response to the ineffectiveness of softmax parameterization observed in practice.

Lower bounds. Establishing information-theoretic or algorithmic-specific lower bounds on the statistical and computational complexities of RL algorithms — often achieved by constructing hard MDP instances — plays an instrumental role in understanding the bottlenecks of RL algorithms. To give a few examples, Azar et al. (2013) established an information-theoretic lower bound on the sample complexity of learning the optimal policy in a generative model, whereas Khamaru et al. (2020); Pananjady and Wainwright (2020) developed instance-dependent lower bounds for policy evaluation. Additionally, Agarwal et al. (2019) constructed a chain-like MDP whose value function under direct parameterization might contain very flat saddle points under a certain initial state distribution, highlighting the role of distribution mismatch coefficients in policy optimization. Finally, exponential-time convergence of gradient descent has been observed in other nonconvex problems as well (e.g., Du et al. (2017)) despite its asymptotic convergence (Lee et al., 2016), although the context and analysis therein are drastically different from what happens in RL settings.

2 Backgrounds

In this section, we introduce the basics of MDPs, and formally describe the softmax PG method. Here and throughout, we denote by $\Delta(\mathcal{X})$ the probability simplex over a set $\mathcal{X}$, and let $|\mathcal{X}|$ represent the cardinality of the set $\mathcal{X}$. Given two probability distributions p and q over $\mathcal{S}$, we adopt the notation $\big\|\frac{p}{q}\big\|_\infty = \max_{s\in\mathcal{S}}\frac{p(s)}{q(s)}$ and $\big\|\frac{1}{q}\big\|_\infty = \max_{s\in\mathcal{S}}\frac{1}{q(s)}$. Throughout this paper, the notation $f(\mathcal{M}) \gtrsim g(\mathcal{M})$ (resp. $f(\mathcal{M}) \lesssim g(\mathcal{M})$) means there exist some universal constants $c > 0$ independent of the parameters of the MDP $\mathcal{M}$ such that $f(\mathcal{M}) \geq cg(\mathcal{M})$ (resp. $f(\mathcal{M}) \leq cg(\mathcal{M})$), while the notation $f(\mathcal{M}) \asymp g(\mathcal{M})$ means that $f(\mathcal{M}) \gtrsim g(\mathcal{M})$ and $f(\mathcal{M}) \lesssim g(\mathcal{M})$ hold simultaneously.

Infinite-horizon discounted MDP. Let $\mathcal{M} = (\mathcal{S}, \{\mathcal{A}_s\}_{s\in\mathcal{S}}, P, r, \gamma)$ be an infinite-horizon discounted MDP (Puterman, 2014). Here, $\mathcal{S}$ represents the state space, $\mathcal{A}_s$ denotes the action space associated with state $s \in \mathcal{S}$, $\gamma \in (0, 1)$ indicates the discount factor, P is the probability transition kernel (namely, for each state-action pair (s, a), $P(\cdot \,|\, s, a) \in \Delta(\mathcal{S})$ denotes the transition probability from state s to the next state when action a is taken), and r stands for a deterministic reward function (namely, $r(s, a)$ is the immediate reward received in state s upon executing action a). Throughout this paper, we assume normalized rewards such that $-1 \leq r(s, a) \leq 1$ for any state-action pair (s, a). In addition, we concentrate on the scenario where γ is quite close to 1, and often refer to $\frac{1}{1-\gamma}$ as the effective horizon of the MDP.

Policy, value function, Q-function and advantage function. The agent operates by adopting a policy π, which is a (randomized) action selection rule based solely on the current state of the MDP. More precisely, for any state $s \in \mathcal{S}$, we use $\pi(\cdot \,|\, s) \in \Delta(\mathcal{A}_s)$ to specify a probability distribution, with $\pi(a \,|\, s)$ denoting the probability of executing action $a \in \mathcal{A}_s$ when in state s. The value function $V^\pi : \mathcal{S} \to \mathbb{R}$ of a policy π — which indicates the expected discounted cumulative reward induced by policy π — is defined as

$$\forall s \in \mathcal{S}: \qquad V^\pi(s) := \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(s^k, a^k) \,\big|\, s^0 = s\right]. \tag{2}$$

Here, the expectation is taken over the randomness of the MDP trajectory $\{(s^k, a^k)\}_{k\geq 0}$ and the policy, where $s^0 = s$ and, for all $k \geq 0$, $a^k \sim \pi(\cdot \,|\, s^k)$ follows the policy π and $s^{k+1} \sim P(\cdot \,|\, s^k, a^k)$ is generated by the transition kernel P. Analogously, we shall also define the value function $V^{\pi}(\mu)$ of a policy π when the initial state is drawn from a distribution μ over $\mathcal{S}$, namely,

$$V^{\pi}(\mu) := \mathbb{E}_{s\sim\mu}\big[V^{\pi}(s)\big]. \tag{3}$$

Additionally, the Q-function Q^{π} of a policy π — namely, the expected discounted cumulative reward under policy π given an initial state-action pair $(s^0, a^0) = (s, a)$ — is formally defined by

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \qquad Q^{\pi}(s,a) := \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k r(s^k, a^k) \,\big|\, s^0 = s, a^0 = a\right], \tag{4}$$

where the expectation is again over the randomness of the MDP trajectory $\{(s^k, a^k)\}_{k\geq 1}$ when policy π is adopted. In addition, the advantage function of policy π is defined as

$$A^{\pi}(s,a) \coloneqq Q^{\pi}(s,a) - V^{\pi}(s) \tag{5}$$

for every state-action pair (s, a).

A major goal is to find a policy that optimizes the value function and the Q-function. Throughout this paper, we denote respectively by $V^{\star}$ and $Q^{\star}$ the optimal value function and optimal Q-function, namely,

$$V^{\star}(s) \coloneqq \max_{\pi} V^{\pi}(s), \qquad Q^{\star}(s,a) \coloneqq \max_{\pi} Q^{\pi}(s,a), \qquad \text{for all } s\in\mathcal{S} \text{ and } a\in\mathcal{A}_s. \tag{6}$$

Softmax parameterization and policy gradient methods. The family of policy optimization algorithms attempts to identify optimal policies by resorting to optimization-based algorithms. To facilitate differentiable optimization, a widely adopted scheme is to parameterize policies using softmax mappings. Specifically, for any real-valued parameter $\theta = [\theta(s,a)]_{s\in\mathcal{S}, a\in\mathcal{A}_s}$, the corresponding softmax policy $\pi_\theta \coloneqq \mathsf{softmax}(\theta)$ is defined such that

$$\forall s\in\mathcal{S} \text{ and } a\in\mathcal{A}_s: \quad \pi_\theta(a\,|\,s) \coloneqq \frac{\exp(\theta(s,a))}{\sum_{a'\in\mathcal{A}_s}\exp(\theta(s,a'))}. \tag{7}$$

With the aim of maximizing the value function under softmax parameterization, namely,

$$\text{maximize}_\theta \quad V^{\pi_\theta}(\mu), \tag{8}$$

softmax PG methods proceed by adopting gradient ascent update rules w.r.t. θ:

$$\theta^{(t+1)} = \theta^{(t)} + \eta\nabla_\theta V^{(t)}(\mu), \qquad t = 0, 1, \cdots. \tag{9a}$$

Here and throughout, we let $V^{(t)} = V^{\pi^{(t)}}$ and $Q^{(t)} = Q^{\pi^{(t)}}$ abbreviate respectively the value function and Q-function of the policy iterate $\pi^{(t)} \coloneqq \pi_{\theta^{(t)}}$ in the t-th iteration, and $\eta > 0$ denotes the learning rate or stepsize.

Interestingly, the gradient $\nabla_\theta V^{\pi_\theta}$ under softmax parameterization (7) admits a closed-form expression (Agarwal et al., 2019), that is, for any state-action pair (s, a),

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta(s,a)} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\,\pi_\theta(a\,|\,s)\,A^{\pi_\theta}(s,a). \tag{9b}$$

Here, $d_\mu^{\pi_\theta}(s)$ represents the *discounted state visitation distribution* of a policy π given the initial state $s^0 \sim \mu$:

$$\forall s\in\mathcal{S}: \qquad d_\mu^{\pi}(s) \coloneqq (1-\gamma)\mathop{\mathbb{E}}_{s^0\sim\mu}\left[\sum_{k=0}^{\infty}\gamma^k \mathbb{P}(s^k = s\,|\,s^0)\right], \tag{9c}$$

with the expectation taken over the randomness of the MDP trajectory $\{(s^k, a^k)\}_{k\geq 0}$ under the policy π and the initial state distribution μ. In words, $d_\mu^{\pi}(s)$ measures — starting from an initial distribution μ — how frequently state s will be visited in a properly discounted fashion. Throughout this paper, we shall denote $A^{(t)} \coloneqq A^{\pi^{(t)}}$ and $d_\mu^{(t)}(s) \coloneqq d_\mu^{\pi^{(t)}}(s)$ for notation simplicity.

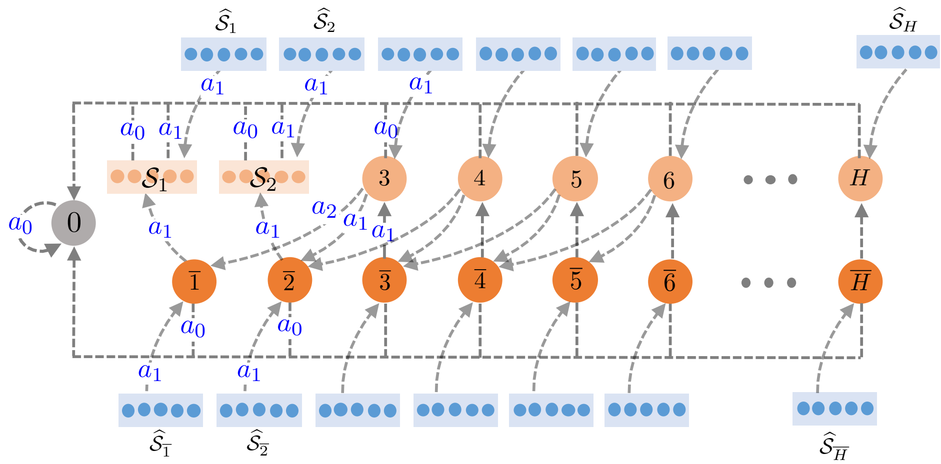

Figure 1: An illustration of the constructed MDP.

3 Construction of the MDP

This section constructs a discounted infinite-horizon MDP $\mathcal{M} = \{\mathcal{S}, \{\mathcal{A}_s\}_{s\in\mathcal{S}}, r, P, \gamma\}$, as depicted in Fig. 1, which forms the basis of the exponential lower bound claimed in this paper. In addition to the basic notation already introduced in Section 2, we remark on the action space as follows.

- *For each state* $s \in \mathcal{S}$, we have $\mathcal{A}_s \subseteq \{a_0, a_1, a_2\}$. For convenience of presentation, we allow the action space to vary with $s \in \mathcal{S}$, but it always comprises no more than 3 actions.

State space partitioning. The states of our MDP exhibit certain group structure. To be precise, we partition the state space $\mathcal{S}$ into a few *disjoint* subsets

$$\mathcal{S} = \{0\} \cup \mathcal{S}_{\mathsf{primary}} \cup \mathcal{S}_{\mathsf{adj}} \cup \mathcal{S}_1 \cup \mathcal{S}_2 \cup \widehat{\mathcal{S}}_1 \cup \cdots \cup \widehat{\mathcal{S}}_H \cup \widehat{\mathcal{S}}_{\overline{1}} \cup \cdots \cup \widehat{\mathcal{S}}_{\overline{H}}, \tag{10}$$

which entails:

- state 0 (an absorbing state);
- two key buffer state subsets $\mathcal{S}_1$ and $\mathcal{S}_2$;
- a set of $H-2$ key primary states $\mathcal{S}_{\mathsf{primary}} \coloneqq \{3, \cdots, H\}$;[1]
- a set of H key adjoint states $\mathcal{S}_{\mathsf{adj}} \coloneqq \{\overline{1}, \overline{2}, \cdots, \overline{H}\}$;
- $2H$ "booster" state subsets $\widehat{\mathcal{S}}_1, \cdots, \widehat{\mathcal{S}}_H, \widehat{\mathcal{S}}_{\overline{1}}, \cdots, \widehat{\mathcal{S}}_{\overline{H}}$.

Remark 2. Our subsequent analysis largely concentrates on the subsets $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_{\mathsf{primary}}$ and $\mathcal{S}_{\mathsf{adj}}$. In particular, each state $s \in \{3, \cdots, H\}$ is paired with what we call an adjoint state $\overline{s}$, whose role will be elucidated shortly. In addition, state $\overline{1}$ (resp. state $\overline{2}$) can be viewed as the adjoint state of the set $\mathcal{S}_1$ (resp. $\mathcal{S}_2$). As we shall make clear momentarily, the set of "booster" states is introduced mainly to boost the initial distribution of the states in $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_{\mathsf{primary}}$, and $\mathcal{S}_{\mathsf{adj}}$.

We shall also specify below the size of these state subsets as well as some key parameters, where the choices of the quantities $c_{\mathrm{h}}, c_{\mathrm{b},1}, c_{\mathrm{b},2}, c_{\mathrm{m}} \asymp 1$ will be made clear in the analysis (cf. (28)).

- H is taken to be on the same order as the "effective horizon" of this discounted MDP, namely,

$$H = \frac{c_{\mathrm{h}}}{1-\gamma}. \tag{11}$$

[1] While we do not include states 1 and 2 here, any state in $\mathcal{S}_1$ (resp. $\mathcal{S}_2$) can essentially be viewed as a (replicated) copy of state 1 (resp. state 2).

- The two buffer state subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ have size

$$|\mathcal{S}_1| = c_{\mathrm{b},1}(1-\gamma)|\mathcal{S}| \qquad \text{and} \qquad |\mathcal{S}_2| = c_{\mathrm{b},2}(1-\gamma)|\mathcal{S}|. \tag{12}$$

- The booster state sets are of the same size, namely,

$$|\widehat{S}_1| = \cdots |\widehat{S}_H| = |\widehat{S}_{\overline{1}}| = \cdots = |\widehat{S}_{\overline{H}}| = c_{\mathrm{m}}(1-\gamma)|\mathcal{S}|. \tag{13}$$

Probability transition kernel and reward function. We now describe the probability transition kernel and the reward function for each state subset. Before continuing, we find it helpful to isolate a few key parameters that will be used frequently in our construction:

$$\tau_s \coloneqq 0.5\gamma^{\frac{2s}{3}}, \tag{14a}$$
$$p \coloneqq c_{\mathrm{p}}(1-\gamma), \tag{14b}$$
$$r_s \coloneqq 0.5\gamma^{\frac{2s}{3}+\frac{5}{6}}, \tag{14c}$$

where $s \in \{1, 2, \cdots, H\}$, and $c_{\mathrm{p}} > 0$ is some small constant that shall be specified later (see (28)). To facilitate understanding, we shall often treat τ_s and r_s ($s \leq H$) as quantities that are all fairly close to 0.5 (which would happen if γ is close to 1 and $H = \frac{c_{\mathrm{h}}}{1-\gamma}$ for c_{h} sufficiently small).

We are now positioned to make precise descriptions of both P and r as follows.

- *Absorbing state* 0*:* singleton action space $\{a_0\}$,

$$P(0 \,|\, 0, a_0) = 1, \qquad\qquad r(0, a_0) = 0. \tag{15}$$

  This is an absorbing state, namely, the MDP will stay in this state permanently once entered. As we shall see below, taking action a_0 in an arbitrary state will enter state 0 immediately.

- *Key primary states $s \in \{3, \cdots, H\}$:* action space $\{a_0, a_1, a_2\}$,

$$P(0 \,|\, s, a_0) = 1, \qquad\qquad r(s, a_0) = r_s + \gamma^2 p \tau_{s-2}, \tag{16a}$$
$$P\big(\overline{s-1} \,|\, s, a_1\big) = 1, \qquad\qquad r(s, a_1) = 0, \tag{16b}$$
$$P(0 \,|\, s, a_2) = 1 - p, \qquad\qquad r(s, a_2) = r_s, \tag{16c}$$
$$P\big(\overline{s-2} \,|\, s, a_2\big) = p, \tag{16d}$$

  where p, τ_s and r_s are all defined in (14).

- *Key adjoint states $\overline{s} \in \{\overline{3}, \cdots, \overline{H}\}$:* action space $\{a_0, a_1\}$,

$$P(0 \,|\, \overline{s}, a_0) = 1, \qquad\qquad r(\overline{s}, a_0) = \gamma\tau_s, \tag{17a}$$
$$P(s \,|\, \overline{s}, a_1) = 1, \qquad\qquad r(\overline{s}, a_1) = 0, \tag{17b}$$

  where τ_s is defined in (14a).

- *Key buffer state subsets $\mathcal{S}_1$ and $\mathcal{S}_2$:* action space $\{a_0, a_1\}$,

$$\forall s_1 \in \mathcal{S}_1: \qquad P(0 \,|\, s_1, a_0) = 1, \qquad\qquad r(s_1, a_0) = -\gamma^2, \tag{18a}$$
$$P(0 \,|\, s_1, a_1) = 1, \qquad\qquad r(s_1, a_1) = \gamma^2, \tag{18b}$$

$$\forall s_2 \in \mathcal{S}_2: \qquad P(0 \,|\, s_2, a_0) = 1, \qquad\qquad r(s_2, a_0) = -\gamma^4, \tag{18c}$$
$$P(0 \,|\, s_2, a_1) = 1, \qquad\qquad r(s_2, a_1) = \gamma^4. \tag{18d}$$

  Given the homogeneity of the states in $\mathcal{S}_1$ (resp. $\mathcal{S}_2$), we shall often use the shorthand notation $P(\cdot \,|\, 1, a)$ (resp. $P(\cdot \,|\, 2, a)$) to abbreviate $P(\cdot \,|\, s_1, a)$ (resp. $P(\cdot \,|\, s_2, a)$) for any $s_1 \in \mathcal{S}_1$ (resp. $s_2 \in \mathcal{S}_2$) for the sake of convenience.

- *Other adjoint states* $\overline{1}$ *and* $\overline{2}$*:* action space $\{a_0, a_1\}$,

$$P(0 \,|\, \overline{1}, a_0) = 1, \qquad r(\overline{1}, a_0) = \gamma\tau_1, \qquad P(s_1 \,|\, \overline{1}, a_1) = \frac{1}{|\mathcal{S}_1|},\ \forall s_1 \in \mathcal{S}_1, \qquad r(\overline{1}, a_1) = 0, \tag{19a}$$

$$P(0 \,|\, \overline{2}, a_0) = 1, \qquad r(\overline{2}, a_0) = \gamma\tau_2, \qquad P(s_2 \,|\, \overline{2}, a_1) = \frac{1}{|\mathcal{S}_2|},\ \forall s_2 \in \mathcal{S}_2, \qquad r(\overline{2}, a_1) = 0, \tag{19b}$$

where τ_1 and τ_2 are defined in (14a).

- *Booster state subsets* $\widehat{\mathcal{S}}_1, \cdots, \widehat{\mathcal{S}}_H, \widehat{\mathcal{S}}_{\overline{1}}, \cdots, \widehat{\mathcal{S}}_{\overline{H}}$*:* singleton action space $\{a_1\}$,

$$\forall s' \in \widehat{\mathcal{S}}_1,\ s \in \mathcal{S}_1 : \qquad P(s \,|\, s', a_1) = 1/|\mathcal{S}_1|, \tag{20a}$$

$$\forall s' \in \widehat{\mathcal{S}}_2,\ s \in \mathcal{S}_2 : \qquad P(s \,|\, s', a_1) = 1/|\mathcal{S}_2|; \tag{20b}$$

for any $s \in \{3, \cdots, H\}$,

$$\forall s' \in \widehat{\mathcal{S}}_s, : \qquad P(s \,|\, s', a_1) = 1, \tag{20c}$$

and for any $\overline{s} \in \{\overline{1}, \cdots, \overline{H}\}$,

$$\forall s' \in \widehat{\mathcal{S}}_{\overline{s}}, : \qquad P(\overline{s} \,|\, s', a_1) = 1. \tag{20d}$$

The rewards in all these cases are set to be 0 (in fact, they will not even appear in the analysis). In addition, any transition probability that has not been specified is equal to zero.

Convenient notation for buffer state subsets $\mathcal{S}_1$ and $\mathcal{S}_2$. By construction, it is easily seen that the states in $\mathcal{S}_1$ (resp. $\mathcal{S}_2$) have identical characteristics; in fact, all states in $\mathcal{S}_1$ (resp. $\mathcal{S}_2$) share exactly the same (soft) value functions and (soft) Q-functions throughout the execution of the (entropy-regularized) PG algorithm. As a result, we introduce the following convenient notation whenever it is clear from the context:

$$V^\pi(s_1) =: V^\pi(1), \qquad Q^\pi(s_1, a) =: Q^\pi(1, a), \qquad A^\pi(s_1) =: A^\pi(1) \qquad \text{for all } s_1 \in \mathcal{S}_1; \tag{21a}$$

$$V^\pi(s_2) =: V^\pi(2), \qquad Q^\pi(s_2, a) =: Q^\pi(2, a), \qquad A^\pi(s_2) =: A^\pi(2) \qquad \text{for all } s_2 \in \mathcal{S}_2; \tag{21b}$$

$$d^\pi_\mu(s_1) =: d^\pi_\mu(1), \qquad \pi(a \,|\, s_1) =: \pi(a \,|\, 1), \qquad \theta(s_1, a) =: \theta(1, a) \qquad \text{for all } s_1 \in \mathcal{S}_1; \tag{21c}$$

$$d^\pi_\mu(s_2) =: d^\pi_\mu(2), \qquad \pi(a \,|\, s_2) =: \pi(a \,|\, 2), \qquad \theta(s_2, a) =: \theta(2, a) \qquad \text{for all } s_2 \in \mathcal{S}_2. \tag{21d}$$

Optimal values and optimal actions. It is instrumental to first determine the optimal value functions and the optimal actions of the constructed MDP as follows, whose proof can be found in Appendix A.3.

Lemma 1. *Suppose that $\gamma^{2H} \geq 2/3$ and $H \geq 2$. Then one has*

$$V^\star(0) = 0, \qquad V^\star(s) = Q^\star(s, a_1) = \gamma^{2s}, \qquad 1 \leq s \leq H, \tag{22a}$$

$$V^\star(\overline{s}) = Q^\star(\overline{s}, a_1) = \gamma^{2s+1}, \quad 1 \leq s \leq H, \tag{22b}$$

and the optimal policy is to take action a_1 in all non-absorbing states. In addition, for any policy π and any state-action pair (s, a), one has $Q^\pi(s, a) \geq -\gamma^2$.

Lemma 1 tells us that for this MDP, the optimal policy for all non-absorbing states takes a simple form: sticking to action a_1. In particular, when $\gamma \approx 1$ and $\gamma^H \approx 1$, Lemma 1 reveals that the optimal values of all non-absorbing major states are fairly close to 1, namely,

$$V^\star(s) \approx 1 \qquad \text{for all } s \in \{1, \cdots, H\} \cup \{\overline{1}, \cdots, \overline{H}\}. \tag{23}$$

Moreover, it indicates that the Q-function (and hence the value function) is always bounded below by -1.

4 Analysis: proof outline

4.1 Preparation: crossing times and choice of constants

Crossing times. To investigate how long it takes for softmax PG methods to converge to the optimal policy, we shall pay particular attention to a family of key quantities: the number of iterations needed for $V^{(t)}(s)$ to surpass a prescribed threshold τ ($\tau < 1$) *before* it reaches its optimal value. To be precise, for each $s \in \{3, \cdots, H\} \cup \{\overline{1}, \cdots, \overline{H}\}$ and any given threshold $\tau > 0$, we introduce the following *crossing time*:

$$t_s(\tau) := \arg\min \big\{ t \,|\, V^{(t)}(s) \geq \tau \big\}. \tag{24}$$

When it comes to the buffer state subsets $\mathcal{S}_1$ and $\mathcal{S}_2$, we define the crossing times analogously as follows

$$t_1(\tau) := \arg\min \big\{ t \,|\, V^{(t)}(1) \geq \tau \big\} \qquad \text{and} \qquad t_2(\tau) := \arg\min \big\{ t \,|\, V^{(t)}(2) \geq \tau \big\}, \tag{25}$$

where we recall the notation $V^{(t)}(1)$ and $V^{(t)}(2)$ introduced in (21).

Monotonicity of crossing times. Recalling the definition (24) of the crossing time $t_s(\cdot)$, we know that

$$V^{(t)}(s) < \tau_s \qquad \text{for all } t < t_s(\tau_s), \tag{26}$$

with τ_s defined in expression (14a). We immediately make note of the following crucial monotonicity property that will be justified later in Remark 5:

$$t_2(\tau_2) \leq t_3(\tau_3) \leq \cdots \leq t_H(\tau_H). \tag{27}$$

It will also be shown in Lemma 4 that $t_1(\tau_1) \leq t_2(\tau_2)$ when the constants $c_{\mathrm{b},1}, c_{\mathrm{b},2}$ and c_{m} are properly chosen.

Choice of parameters. We assume the following choice of parameters throughout the proof:

$$\gamma > 0.96,\ c_{\mathrm{m}} < 1,\ c_{\mathrm{h}} < 0.19,\ \eta < \frac{(1-\gamma)^2}{5},\ \frac{c_{\mathrm{b},1}}{c_{\mathrm{m}}} \leq \frac{1}{79776},\ 8 \leq \frac{c_{\mathrm{b},2}}{c_{\mathrm{m}}} \leq 15,\ c_{\mathrm{p}} < \frac{1}{2016}. \tag{28}$$

In the sequel, we outline the key steps that underlie the proof of our main results, with the proofs of the key lemmas postponed to the appendix.

4.2 A high-level picture

While our proof is highly technical, it is prudent to point out some key features that help paint a high-level picture about the slow convergence of the algorithm. The chain-like structure of our MDP underscores a sort of *sequential dependency*: the dynamic of any primary state $s \in \{3, \cdots, H\}$ depends heavily on what happens in those states prior to s — particularly state $s-1$, state $s-2$ as well as the associated adjoint states. By carefully designing the immediate rewards, we can ensure that for any $s \in \{3, \cdots, H\}$, the iterate $\pi^{(t)}(a_1 \,|\, s)$ corresponding to the optimal action a_1 keeps decreasing before $\pi^{(t)}(a_1 \,|\, s-2)$ gets reasonably close to 1. As illustrated in Figure 2, this feature implies that the time taken for $\pi^{(t)}(a_1 \,|\, s)$ to get close to 1 grows (at least) geometrically as s increases, as will be formalized in (39).

Furthermore, we summarize below the typical dynamics of the iterates $\theta^{(t)}(s, a)$ before they converge, which are helpful for the reader to understand the proof. We start with the key buffer state sets $\mathcal{S}_1$ and $\mathcal{S}_2$, which are the easiest to describe.

Dynamics of $\theta^{(t)}(s, a)$ (for key buffer state sets $\mathcal{S}_1$ and $\mathcal{S}_2$):

1. Initialization: $\theta^{(0)}(1, a_0) = \theta^{(0)}(1, a_1) = 0$ and $\theta^{(0)}(2, a_0) = \theta^{(0)}(2, a_1) = 0$
2. All iterations (Lemma 4):
   - $\theta^{(t)}(1, a_1)$ and $\theta^{(t)}(2, a_1)$ keep increasing and remains the largest
   - $\theta^{(t)}(1, a_0)$ and $\theta^{(t)}(2, a_0)$ keep decreasing and remains the smallest

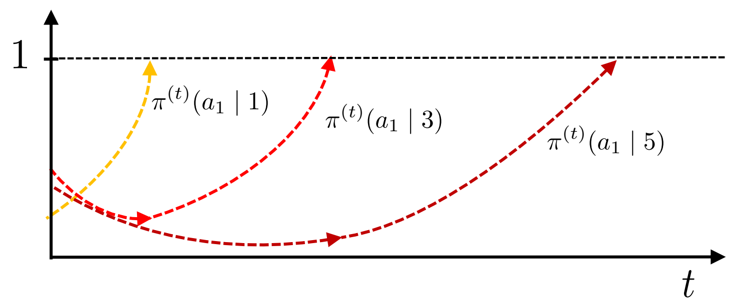

Figure 2: An illustration of the dynamics of $\pi^{(t)}(a_1 \,|\, s)$.

Next, the dynamics of $\theta^{(t)}(s,a)$ for the key primary states $3 \le s \le H$ are much more complicated, and rely heavily on the status of several prior states $s-1$, $s-2$ and $\overline{s-1}$. This motivates us to divide the dynamics into several stages based on the crossing times of these prior states, which are illustrated in Figure 3 as well. Here, we remind the reader of the definition of τ_s in (14).

Dynamics of $\theta^{(t)}(s,a)$ (for key primary states $3 \le s \le H$):

1. Initialization: $\theta^{(0)}(s,a_0) = \theta^{(0)}(s,a_1) = \theta^{(0)}(s,a_2) = 0$
2. Initial stage: $t < t_{s-2}(\tau_{s-2})$ (Lemma 5)
   - $\theta^{(t)}(s,a_1)$ keeps decreasing and remains the smallest
   - $\theta^{(t)}(s,a_0)$ keeps increasing and remains the largest
   - $\theta^{(t)}(s,a_2)$ keeps increasing
3. Intermediate stage: $t_{s-2}(\tau_{s-2}) \le t \le t_{\overline{s-1}}(\tau_s)$ (Lemma 6)
   - $\theta^{(t)}(s,a_1)$ keeps decreasing and remains the smallest
   - $\theta^{(t)}(s,a_2)$ keeps increasing
4. Final stage (part 1): $t_{\overline{s-1}}(\tau_s) < t < t_{\mathsf{ref}}$ (Lemma 7)
   - $\theta^{(t)}(s,a_1)$ increases a little
   - $\theta^{(t)}(s,a_0)$ keeps decreasing and approaches $\theta^{(t)}(s,a_1)$
   - $\theta^{(t)}(s,a_2)$ keeps increasing and becomes the largest
5. Final stage (part 2): $t \ge t_{\mathsf{ref}}$ (Lemma 7)
   - $\theta^{(t)}(s,a_1)$ keeps increasing and becomes the largest
   - $\theta^{(t)}(s,a_2)$ decreases a lot

4.3 Proof outline

We are now in a position to outline the main steps of the proof of Theorem 1, with details deferred to the appendix.

Step 1: bounding the discounted state visitation distributions

In view of the PG update rule (9), the size of the policy gradient relies heavily on the discounted state visitation distribution $d_\mu^{(t)}(s)$. In light of this observation, this step aims to quantify the magnitudes of

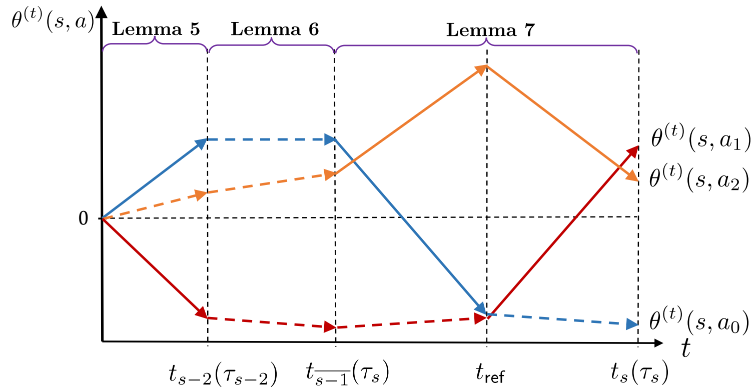

Figure 3: An illustration of the dynamics of $\{\theta^{(t)}(s,a)\}_{a\in\{a_0,a_1,a_2\}}$ vs. iteration number t. Here, we use solid lines to emphasize the time ranges for which the dynamics of $\theta^{(t)}(s,a)$ play the most crucial roles in our lower bound analysis.

$d_\mu^{(t)}(s)$, for which we start with several universal lower bounds regardless of the policy in use.

Lemma 2. *For any policy π, the following lower bounds hold true:*

$$d_\mu^\pi(s) \ge c_{\mathrm{m}}\gamma(1-\gamma)^2, \qquad \text{if } s\in\{3,\cdots,H\}, \tag{29a}$$

$$d_\mu^\pi(\overline{s}) \ge c_{\mathrm{m}}\gamma(1-\gamma)^2, \qquad \text{if } \overline{s}\in\{\overline{1},\cdots,\overline{H}\}, \tag{29b}$$

$$d_\mu^\pi(1) \ge \frac{c_{\mathrm{m}}\gamma(1-\gamma)^2}{|\mathcal{S}_1|} = \gamma(1-\gamma)\frac{c_{\mathrm{m}}}{c_{\mathrm{b},1}}\cdot\frac{1}{|\mathcal{S}|}, \tag{29c}$$

$$d_\mu^\pi(2) \ge \frac{c_{\mathrm{m}}\gamma(1-\gamma)^2}{|\mathcal{S}_2|} = \gamma(1-\gamma)\frac{c_{\mathrm{m}}}{c_{\mathrm{b},2}}\cdot\frac{1}{|\mathcal{S}|}. \tag{29d}$$

As it turns out, the above lower bounds are order-wise tight estimates prior to certain crucial crossing times. This is formalized in the following lemma, where we recall the definition of τ_s in (14).

Lemma 3. *Under the assumption (28), the following results hold:*

$$\forall 3\le s\le H,\ t\le t_s(\tau_s): \qquad d_\mu^{(t)}(s) \le 14c_{\mathrm{m}}(1-\gamma)^2, \tag{30a}$$

$$\forall 2\le s\le H,\ t\le t_s(\tau_s): \qquad d_\mu^{(t)}(\overline{s}) \le 14c_{\mathrm{m}}(1-\gamma)^2, \tag{30b}$$

$$\forall t\le t_2(\tau_2): \qquad d_\mu^{(t)}(2) \le \frac{1-\gamma}{|\mathcal{S}|}\left(1+\frac{8c_{\mathrm{m}}}{c_{\mathrm{b},2}}\right), \tag{30c}$$

$$\forall t\le t_2(\tau_2): \qquad d_\mu^{(t)}(\overline{1}) \le 14c_{\mathrm{m}}(1-\gamma)^2, \tag{30d}$$

$$\forall t\le \min\{t_1(\tau_1), t_2(\tau_2)\}: \qquad d_\mu^{(t)}(1) \le \frac{1-\gamma}{|\mathcal{S}|}\left(1+\frac{17c_{\mathrm{m}}}{c_{\mathrm{b},1}}\right). \tag{30e}$$

Remark 3. As will be demonstrated in Lemma 4, one has $t_1(\tau_1)\le t_2(\tau_2)$ for properly chosen constants $c_{\mathrm{b},1}, c_{\mathrm{b},2}$ and c_{m}. Therefore, we shall bear in mind that the properties (30d) and (30e) hold for any $t\le t_1(\tau_1)$.

The proofs of these two lemmas are deferred to Appendix B. The sets of booster states, whose cardinality is controlled by c_{m}, play an important role in sandwiching the initial distribution of the states in $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_{\mathsf{primary}}$, and $\mathcal{S}_{\mathsf{adj}}$. Combining these bounds, we uncover the following properties happening before $V^{(t)}(s)$ exceeds τ_s:

- For any key primary state $s\in\{3,\cdots,H\}$ or any adjoint state $s\in\{\overline{1},\cdots,\overline{H}\}$, one has

$$d_\mu^{(t)}(s) \asymp (1-\gamma)^2.$$

- For any state s contained in the buffer state subsets $\mathcal{S}_1$ and $\mathcal{S}_2$, we have

$$d_\mu^{(t)}(1) \asymp \frac{(1-\gamma)^2}{|\mathcal{S}_1|} \qquad \text{and} \qquad d_\mu^{(t)}(2) \asymp \frac{(1-\gamma)^2}{|\mathcal{S}_2|},$$

  where we recall the size of $\mathcal{S}_1$ and $\mathcal{S}_2$ in (12). In other words, the discounted state visitation probability of any buffer state is substantially smaller than that of any key primary state $3, \cdots, H$ or adjoint state. In principle, the size of each buffer state subset plays a crucial role in determining the associated $d_\mu^{(t)}(s)$ — the larger the size of the buffer state subset, the smaller the resulting state visitation probability.

- Further, the aggregate discounted state visitation probability of the above states is no more than the order of

$$(1-\gamma)^2 \cdot H \asymp 1-\gamma = o(1),$$

  which is vanishingly small. In fact, state 0 and the booster states account for the dominant fraction of state visitations at the initial stage of the algorithm.

Step 2: characterizing the crossing times for the first few states ($\mathcal{S}_1$, $\mathcal{S}_2$, and $\overline{1}$)

Armed with the bounds on $d_\mu^{(t)}$ developed in Step 1, we can move forward to study the crossing times for the key states. In this step, we pay attention to the crossing times for the buffer states $\mathcal{S}_1, \mathcal{S}_2$ as well as the first adjoint state $\overline{1}$, which forms a crucial starting point towards understanding the behavior of the subsequent states. Specifically, the following lemma develops lower and upper bounds regarding these quantities, whose proof can be found in Appendix C.

Lemma 4. *Suppose that (28) holds. If $|\mathcal{S}| \geq 1/(1-\gamma)^2$, then the crossing times satisfy*

$$\frac{\log 3}{1+17c_{\mathrm{m}}/c_{\mathrm{b},1}} \frac{|\mathcal{S}|}{\eta} \leq t_1(\tau_1) \leq t_1\big(\gamma^2 - 1/4\big) \leq t_2(\tau_2) \leq t_2\big(\gamma^4 - 1/4\big) \leq \frac{15c_{\mathrm{b},2}}{c_{\mathrm{m}}} \frac{|\mathcal{S}|}{\eta}. \tag{31a}$$

In addition, if $|\mathcal{S}| \geq \frac{320\gamma^3}{c_{\mathrm{m}}(1-\gamma)^2}$, then one has

$$t_2(\tau_2) > t_{\overline{1}}\big(\gamma^3 - 1/4\big). \tag{31b}$$

For properly chosen constants $c_{\mathrm{b},1}$, $c_{\mathrm{b},2}$ and c_{m}, Lemma 4 delivers the following important messages:

- The cross times of these first few states are already fairly large; for instance,

$$t_1(\tau_1) \asymp t_2(\tau_2) \asymp \frac{|\mathcal{S}|}{\eta}, \tag{32}$$

  which scale linearly with the state space dimension. As we shall see momentarily, while $t_1(\tau_1)$ and $t_2(\tau_2)$ remain polynomially large, these play a pivotal role in ensuring rapid explosion of the crossing times of the states that follow (namely, the states $\{3, \cdots, H\}$).

- By adjusting the quantities $c_{\mathrm{b},1}$ and $c_{\mathrm{b},2}$, we can guarantee a strict ordering such that the crossing time of state 2 is at least as large as that of both state 1 and state $\overline{1}$. This property is helpful as well for subsequent analysis.

Step 3: understanding the dynamics of $\theta^{(t)}(s,a)$ before $t_{s-2}(\tau_{s-2})$

With the above characterization of the crossing times for the first few states, we are ready to investigate the dynamics of $\theta^{(t)}(s,a)$ $(3 \leq s \leq H)$ at the initial stage, that is, the duration prior to the threshold $t_{s-2}(\tau_{s-2})$. Our finding for this stage is summarized in the following lemma, with the proof deferred to Appendix D.

Lemma 5. *Suppose that (28) holds. For any $3 \leq s \leq H$ and any $0 \leq t \leq t_{s-2}(\tau_{s-2})$, one has*

$$\theta^{(t)}(s,a_1) \leq -\frac{1}{2}\log\Big(1+\frac{c_{\mathrm{m}}\gamma}{35}\eta(1-\gamma)^2 t\Big) \tag{33}$$

$$\textit{and} \qquad \theta^{(t)}(s,a_0) \geq \theta^{(t)}(s,a_2) \geq 0. \tag{34}$$

Lemma 5 makes clear the behavior of $\theta^{(t)}(s,a)$ during this initial stage:

- The iterate $\theta^{(t)}(s,a_1)$ associated with the optimal action a_1 keeps dropping at a rate of $\log\big(O(\frac{1}{\sqrt{t}})\big)$, and remains the smallest compared to the ones with other actions.
- The other two iterates $\theta^{(t)}(s,a_0)$ and $\theta^{(t)}(s,a_2)$ stay non-negative throughout this stage, with a_0 being perceived as more favorable than the other two actions.
- In fact, a closer inspection of the proof in Section D reveals that $\theta^{(t)}(s,a_2)$ remains increasing — even though at a rate slower than that of $\theta^{(t)}(s,a_0)$ — throughout this stage (see (122) and the gradient expression (9b)).

In particular, around the threshold $t_{s-2}(\tau_{s-2})$, the iterate $\theta^{(t)}(s,a_1)$ becomes as small as

$$\exp\Big(\theta^{(t)}(s,a_1)\Big) \leq O\left(\frac{1}{\sqrt{\eta(1-\gamma)^2 t_{s-2}(\tau_{s-2})}}\right).$$

In fact, an inspection of the proof of this lemma reveals that

$$\pi^{(t)}(a_1 \,|\, s) \leq O\left(\frac{1}{\eta(1-\gamma)^2 t_{s-2}(\tau_{s-2})}\right) \qquad \text{when } t = t_{s-2}(\tau_{s-2}).$$

This means that $\pi^{(t)}(a_1 \,|\, s)$ becomes smaller for a larger $t_{s-2}(\tau_{s-2})$, making it more difficult to return/converge to 1 afterward.

Step 4: understanding the dynamics of $\theta^{(t)}(s,a)$ between $t_{s-2}(\tau_{s-2})$ and $t_{\overline{s-1}}(\tau_s)$

Next, we investigate, for any $3 \leq s \leq H$, the behavior of the iterates during an "intermediate" stage, namely, the duration when the iteration count t is between $t_{s-2}(\tau_{s-2})$ and $t_{\overline{s-1}}(\tau_s)$. This is summarized in the following lemma, whose proof can be found in Appendix E.

Lemma 6. *Consider any $3 \leq s \leq H$. Assume that (28) holds. Suppose that*

$$t_{s-1}(\tau_{s-1}) > t_{\overline{s-2}}(\tau_{s-1}) + \frac{2444s}{c_{\mathrm{m}}\gamma\eta(1-\gamma)^2}, \tag{35a}$$

$$t_3(\tau_3) > t_2(\gamma^4 - 1/4). \tag{35b}$$

Then one has

$$\theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_1) \leq \theta^{(t_{s-2}(\tau_{s-2}))}(s,a_1) \qquad \text{and} \qquad \theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_2) \geq 0. \tag{36}$$

In particular, when $s=3$, the results in (36) hold true without requiring the assumption (35).

Remark 4. Condition (35a) only requires $t_{s-1}(\tau_{s-1})$ to be slightly larger than $t_{\overline{s-2}}(\tau_{s-1})$, which will be justified using an induction argument when proving the main theorem.

As revealed by the claim (36) of Lemma 6, the iterate $\theta^{(t)}(s,a_2)$ remains sufficiently large during this intermediate stage. In the meantime, Lemma 6 guarantees that $\theta^{(t)}(s,a_1)$ does not grow during this stage, lying below the level of $\theta^{(t_{s-2}(\tau_{s-2}))}(s,a_1)$ that has been pinned down in Lemma 5 (which has been shown to be quite small). Both of these properties make clear that the iterates $\theta^{(t)}(s,a)$ remain far from optimal at the end of this intermediate stage.

Step 5: establishing a blowing-up phenomenon

The next lemma, which plays a pivotal role in developing the desired exponential convergence lower bound, demonstrates that the cross times explode at a super fast rate. The proof is postponed to Appendix F.

Lemma 7. *Consider any $3 \leq s \leq H$. Suppose that (28) holds and*

$$t_{s-2}(\tau_{s-2}) \geq \Big(\frac{6300e}{c_{\mathrm{p}}(1-\gamma)}\Big)^4 \frac{1}{\frac{c_{\mathrm{m}}\gamma}{35}\eta(1-\gamma)^2}. \tag{37}$$

Then there exists a time instance t_{ref} obeying $t_{\overline{s-1}}(\tau_s) \leq t_{\mathsf{ref}} < t_s(\tau_s)$ such that

$$\theta^{(t_{\mathsf{ref}})}(s,a_0) \leq \theta^{(t_{\mathsf{ref}})}(s,a_1) - \log\Big(\frac{c_{\mathrm{p}}}{16128}(1-\gamma)\Big), \tag{38a}$$

$$\theta^{(t_{\mathsf{ref}})}(s,a_1) \leq \; -\frac{1}{2}\log\Big(1 + \frac{c_{\mathrm{m}}\gamma}{35}\eta(1-\gamma)^2 t_{s-2}(\tau_{s-2})\Big) + 1, \tag{38b}$$

and at the same time,

$$t_s(\tau_s) - t_{\mathsf{ref}} \geq 10^{-10} c_{\mathrm{p}} c_{\mathrm{m}}^{0.5} \eta^{0.5}(1-\gamma)^2 \Big(t_{s-2}(\tau_{s-2})\Big)^{1.5}. \tag{38c}$$

The most important message of Lemma 7 lies in property (38c). In a nutshell, this property uncovers that the crossing time $t_s(\tau_s)$ is substantially larger than $t_{s-2}(\tau_{s-2})$, namely,

$$t_s(\tau_s) \gtrsim \eta^{0.5}(1-\gamma)^2 \Big(t_{s-2}(\tau_{s-2})\Big)^{1.5}, \tag{39}$$

thus leading to explosion at a super-linear rate. By contrast, the other two properties unveil some important features happening between $t_{\overline{s-1}}(\tau_s)$ and $t_s(\tau_s)$ that in turn lead to property (38c). In words, property (38a) requires $\theta^{(t_{\mathsf{ref}})}(s,a_0)$ to be not much larger than $\theta^{(t_{\mathsf{ref}})}(s,a_1)$; property (38b) indicates that: when $t_{s-2}(\tau_{s-2})$ is large, both $\theta^{(t_{\mathsf{ref}})}(s,a_1)$ and $\theta^{(t_{\mathsf{ref}})}(s,a_0)$ are fairly small, with $\theta^{(t_{\mathsf{ref}})}(s,a_2)$ being the dominant one.

The reader might naturally wonder what the above results imply about $\pi^{(t_{\mathsf{ref}})}(a_1 \,|\, s)$ (as opposed to $\theta^{(t_{\mathsf{ref}})}(s,a_1)$). Towards this end, we make the observation that

$$\begin{aligned}\pi^{(t_{\mathsf{ref}})}(a_1 \,|\, s) &= \frac{\exp\big(\theta^{(t_{\mathsf{ref}})}(s,a_1)\big)}{\sum_a \exp\big(\theta^{(t_{\mathsf{ref}})}(s,a)\big)} \leq \frac{\exp\big(\theta^{(t_{\mathsf{ref}})}(s,a_1)\big)}{\exp\big(\theta^{(t_{\mathsf{ref}})}(s,a_2)\big)} \overset{\text{(i)}}{=} \exp\big(2\theta^{(t_{\mathsf{ref}})}(s,a_1) + \theta^{(t_{\mathsf{ref}})}(s,a_0)\big) \\ &\overset{\text{(ii)}}{\lesssim} \frac{1}{(1-\gamma)\Big(\eta(1-\gamma)^2 t_{s-2}(\tau_{s-2})\Big)^{1.5}} \asymp \frac{1}{\eta^{1.5}(1-\gamma)^4\big(t_{s-2}(\tau_{s-2})\big)^{1.5}},\end{aligned} \tag{40}$$

where (i) holds true since $\sum_a \theta^{(t_{\mathsf{ref}})}(s,a) = 0$ (a well-known property of policy gradient methods as recorded in Lemma 8(vii)), and (ii) follows from the properties (38a) and (38b). In other words, $\pi^{(t_{\mathsf{ref}})}(s,a_1)$ is inversely proportional to $\big(t_{s-2}(\tau_{s-2})\big)^{3/2}$. As we shall see, the time taken for $\pi^{(t_{\mathsf{ref}})}(a_1 \,|\, s)$ to converge to 1 is proportional to the inverse policy iterate $\big(\pi^{(t)}(s,a_1)\big)^{-1}$, meaning that it is expected to take an order of $\big(t_{s-2}(\tau_{s-2})\big)^{3/2}$ iterations to increase from $\pi^{(t_{\mathsf{ref}})}(s,a_1)$ to 1.

Step 6: putting all this together

With the above steps in place, we are ready to combine them to establish the following result. As can be easily seen, Theorem 1 is an immediate consequence of Theorem 2.

Theorem 2. *Suppose that (28) holds. There exist some universal constants $c_1, c_2, c_3 > 0$ such that*

$$t_s(0.5) \geq c_1 \frac{|\mathcal{S}|^{\frac{2}{3}}}{\eta}\Big(c_2|\mathcal{S}|\Big)^{\frac{1}{3}\cdot 1.5^{\lfloor s/2 \rfloor}}, \tag{41}$$

provided that

$$|\mathcal{S}| \geq \frac{c_3}{(1-\gamma)^6}. \tag{42}$$

Proof of Theorem 2. Let us define two universal constants $C_1 \coloneqq \frac{\log 3}{1+17c_{\mathrm{m}}/c_{\mathrm{b},1}}$ and $C_2 \coloneqq \frac{10^{-20}c_{\mathrm{p}}^2 c_{\mathrm{m}} \log 3}{1+17c_{\mathrm{m}}/c_{\mathrm{b},1}}$. We claim that if one can show that

$$t_s(\tau_s) \geq C_1 \frac{|\mathcal{S}|}{\eta} \Big(C_2(1-\gamma)^4 |\mathcal{S}|\Big)^{1.5^{\lfloor (s-1)/2 \rfloor}-1}, \tag{43}$$

then the desired bound (41) holds true directly. In order to see this, recall that $\tau_s \leq 1/2$ by definition, and therefore,

$$t_s(0.5) \geq t_s(\tau_s) \overset{\text{(i)}}{\geq} C_1 \frac{|\mathcal{S}|}{\eta} \Big(C_2 \sqrt[3]{|\mathcal{S}|}\Big)^{1.5^{\lfloor (s-1)/2 \rfloor}-1} \overset{\text{(ii)}}{\geq} c_1 \frac{|\mathcal{S}|^{\frac{2}{3}}}{\eta} \Big(c_2 |\mathcal{S}|\Big)^{\frac{1}{3}\cdot 1.5^{\lfloor s/2 \rfloor}}.$$

Here, (i) follows from (43) in conjunction with the assumption (42), whereas (ii) holds true by setting $c_1 = C_1/C_2$ and $c_2 = C_2^3$.

It is then sufficient to prove the inequality (43), towards which we shall resort to mathematical induction in conjunction with the following induction hypothesis

$$t_s(\tau_s) > t_{\overline{s-1}}(\tau_s) + \frac{2444(s+1)}{c_{\mathrm{m}}\gamma\eta(1-\gamma)^2}, \qquad \text{for } s \geq 3. \tag{44}$$

- We start with the cases with $s = 1, 2, 3$. It follows from Lemma 4 that

$$t_2(\tau_2) \geq t_1(\tau_1) \geq \frac{\log 3}{1+17c_{\mathrm{m}}/c_{\mathrm{b},1}} \frac{|\mathcal{S}|}{\eta} = C_1 \frac{|\mathcal{S}|}{\eta}, \tag{45}$$

  which validates the above claim (43) for $s = 1$ and $s = 2$. In addition, Lemma 7 ensures that

$$\begin{aligned} t_3(\tau_3) - \max\Big\{t_{\overline{1}}(\gamma^3 - 1/4),\ t_{\overline{2}}(\tau_3)\Big\} &\geq 10^{-10} c_{\mathrm{p}} c_{\mathrm{m}}^{0.5} \eta^{0.5} (1-\gamma)^2 \Big(t_1(\tau_1)\Big)^{1.5} \\ &\geq \frac{9776}{c_{\mathrm{m}}\gamma\eta(1-\gamma)^2}, \end{aligned} \tag{46}$$

  where the last inequality is guaranteed by (45) and the assumption $|\mathcal{S}| \geq \max\Big\{\frac{4888}{C_1 c_{\mathrm{m}}\gamma(1-\gamma)^2}, \frac{4}{C_2(1-\gamma)^4}\Big\}$. This implies that the inequality (44) is satisfied when $s = 3$.

- Next, suppose that the inequality (43) holds true up to state $s-1$ and the inequality (44) holds up to s for some $3 \leq s \leq H$. To invoke the induction argument, it suffices to show that the inequality (43) continues to hold for state s and the inequality (44) remains valid for $s+1$. This will be accomplished by taking advantage of Lemma 7.

  Given that the inequality (43) holds true for every state up to $s-1$, one has

$$t_{s-1}(\tau_{s-1}) \geq t_{s-2}(\tau_{s-2}) \geq C_1 \frac{|\mathcal{S}|}{\eta} \Big(C_2(1-\gamma)^4|\mathcal{S}|\Big)^{1.5^{\lfloor (s-3)/2 \rfloor}-1} \geq \Big(\frac{6300e}{c_{\mathrm{p}}(1-\gamma)}\Big)^4 \frac{1}{\frac{c_{\mathrm{m}}\gamma}{35}\eta(1-\gamma)^2},$$

  where the last inequality is satisfied provided that $|\mathcal{S}| > \max\Big\{\big(\frac{6300e}{c_{\mathrm{p}}}\big)^4 \frac{35}{C_1 c_{\mathrm{m}}\gamma(1-\gamma)^6}, \frac{4}{C_2(1-\gamma)^4}\Big\}$. Therefore, Lemma 7 is applicable for both s and $s+1$, thus leading to

$$\begin{aligned} t_s(\tau_s) - t_{\overline{s-1}}(\tau_s) &\geq 10^{-10} c_{\mathrm{p}} c_{\mathrm{m}}^{0.5} \eta^{0.5} (1-\gamma)^2 \Big(t_{s-2}(\tau_{s-2})\Big)^{1.5} \\ &\geq 10^{-10} c_{\mathrm{p}} c_{\mathrm{m}}^{0.5} \eta^{0.5} (1-\gamma)^2 \left(C_1 \frac{|\mathcal{S}|}{\eta} \Big(C_2(1-\gamma)^4|\mathcal{S}|\Big)^{1.5^{\lfloor (s-3)/2 \rfloor}-1}\right)^{1.5} \\ &\geq C_1 \frac{|\mathcal{S}|}{\eta} \Big(C_2(1-\gamma)^4|\mathcal{S}|\Big)^{1.5^{\lfloor (s-1)/2 \rfloor}-1}. \end{aligned}$$

Here, the last step relies on the condition $10^{-10}c_{\mathrm{p}}c_{\mathrm{m}}^{0.5}\eta^{0.5}(1-\gamma)^2(C_1\frac{|\mathcal{S}|}{\eta})^{0.5} \geq 1$. This in turn establishes the property (43) for state s (given that $t_{\overline{s-1}}(\tau_s) \geq 0$). In addition, Lemma 7 — when applied to $s+1$ — gives

$$\begin{aligned}
t_{s+1}(\tau_{s+1}) - t_{\overline{s}}(\tau_{s+1}) &\geq 10^{-10}c_{\mathrm{p}}c_{\mathrm{m}}^{0.5}\eta^{0.5}(1-\gamma)^2\Big(t_{s-1}(\tau_{s-1})\Big)^{1.5} \\
&\geq 10^{-10}c_{\mathrm{p}}c_{\mathrm{m}}^{0.5}\eta^{0.5}(1-\gamma)^2\left(C_1\frac{|\mathcal{S}|}{\eta}\Big(C_2(1-\gamma)^4|\mathcal{S}|\Big)^{1.5^{\lfloor (s-2)/2\rfloor}-1}\right)^{1.5} \\
&\geq C_1\frac{|\mathcal{S}|}{\eta}\Big(C_2(1-\gamma)^4|\mathcal{S}|\Big)^{1.5^{\lfloor s/2\rfloor}-1} \\
&\geq \frac{2444(s+2)}{c_{\mathrm{m}}\gamma\eta(1-\gamma)^2},
\end{aligned}$$

where the last step follows as long as $|\mathcal{S}| > \max\left\{\frac{4888}{C_1c_{\mathrm{m}}\gamma(1-\gamma)^2}, \frac{4}{C_2(1-\gamma)^4}\right\}$. We have thus established the property (44) for state $s+1$.

Putting all the above pieces together, we arrive at the inequality (43), thus establishing Theorem 2. $\square$

5 Discussions

This paper develops an algorithmic-specific lower bound on the iteration complexity of softmax PG methods, obtained by analyzing its trajectory on a carefully-designed hard MDP instance. It is shown that the iteration complexity of softmax PG methods can scale pessimistically, in fact (super-)exponentially, with the dimension of the state space and the effective horizon of the discounted MDP. Our finding makes apparent the potential inefficiency of softmax PG methods in solving large-dimensional and long-horizon problems, which in turns suggests the necessity of carefully adjusting update rules and/or enforcing proper regularization in accelerating PG methods.

Acknowledgements

G. Li and Y. Gu are supported in part by the grant NSFC-61971266. Y. Wei is supported in part by the grants NSF CCF-2007911 and DMS-2015447. Y. Chi is supported in part by the grants ONR N00014-18-1-2142 and N00014-19-1-2404, ARO W911NF-18-1-0303, NSF CCF-1806154 and CCF-2007911. Y. Chen is supported in part by the grants AFOSR YIP award FA9550-19-1-0030, ONR N00014-19-1-2120, ARO YIP award W911NF-20-1-0097, ARO W911NF-18-1-0303, NSF CCF-1907661, DMS-2014279 and IIS-1900140, and the Princeton SEAS Innovation Award.

A Preliminary facts

A.1 Basic properties of the constructed MDP

In this section, we provide more basic properties about the MDP we have constructed (see Section 3). Specifically, we present a miscellaneous collection of basic relations regarding more general policies, postponing the proof to Appendix A.4.

Lemma 8. *Consider any policy π, and recall the quantities defined in (14). Suppose that $\gamma^{2H} \geq 1/2$ and $0 < c_{\mathrm{p}} \leq 1/6$.*

(i) For any state $s \in \{3, \cdots, H\}$, one has

$$\gamma^{\frac{3}{2}}\tau_{s-1} \leq Q^{\pi}(s, a_0) = r_s + \gamma^2 p\tau_{s-2} \leq \gamma^{\frac{1}{2}}\tau_s, \tag{47a}$$

$$Q^{\pi}(s, a_1) = \gamma V^{\pi}(\overline{s-1}), \tag{47b}$$

$$Q^{\pi}(s,a_2) = r_s + \gamma p V^{\pi}(\overline{s-2}) \le \gamma^{\frac{1}{2}}\tau_s. \tag{47c}$$

If one further has $V^{\pi}(\overline{s-2}) \ge 0$, *then* $Q^{\pi}(s,a_2) \ge \gamma^{\frac{3}{2}}\tau_{s-1}$.

(ii) If $V^{\pi}(s) \ge \tau_s$ *for some* $s \in \{3, \cdots, H\}$, *then we necessarily have*

$$\pi(a_1 \,|\, s) \ge \frac{1-\gamma}{2}. \tag{48}$$

(iii) For any $\overline{s} \in \{\overline{1}, \cdots, \overline{H}\}$, *one has*

$$Q^{\pi}(\overline{s}, a_0) = \gamma\tau_s \qquad \textit{and} \qquad Q^{\pi}(\overline{s}, a_1) = \gamma V^{\pi}(s), \tag{49}$$

where we recall the definition of $V^{\pi}(1)$ *and* $V^{\pi}(2)$ *in* (21)*. In addition, if* $\pi(a_1 \,|\, \overline{s}) > 0$, *then*

$$V^{\pi}(\overline{s}) \ge \gamma\tau_s \qquad \textit{holds if and only if} \qquad V^{\pi}(s) \ge \tau_s. \tag{50}$$

This means that: if $\pi^{(t)}(a_1 \,|\, \overline{s}) > 0$ *holds for all* $t \ge 0$, *then one necessarily has*

$$t_{\overline{s}}(\gamma\tau_s) = t_s(\tau_s). \tag{51}$$

(iv) For any policy π, *we have*

$$Q^{\pi}(1,a_0) = -\gamma^2, \quad Q^{\pi}(1,a_1) = \gamma^2, \quad V^{\pi}(1) = -\gamma^2\pi(a_0 \,|\, 1) + \gamma^2\pi(a_1 \,|\, 1), \tag{52a}$$

$$Q^{\pi}(2,a_0) = -\gamma^4, \quad Q^{\pi}(2,a_1) = \gamma^4, \quad V^{\pi}(2) = -\gamma^4\pi(a_0 \,|\, 2) + \gamma^4\pi(a_1 \,|\, 2). \tag{52b}$$

(v) Consider any policy π *obeying* $\min_{a,s}\pi(a \,|\, s) > 0$. *For every* $s \in \{3, \cdots, H\}$, *if* $V^{\pi}(s) \ge \gamma^{\frac{1}{2}}\tau_s$ *occurs, then one necessarily has* $V^{\pi}(s-1) \ge \tau_{s-1}$.

(vi) If $V^{\pi}(s-2) < \tau_{s-2}$ *and* $\pi(a_1 \,|\, \overline{s-2}) > 0$, *then*

$$Q^{\pi}(s,a_0) - Q^{\pi}(s,a_2) = \gamma p\big(\gamma\tau_{s-2} - V^{\pi}(\overline{s-2})\big) > 0.$$

If $V^{\pi}(s-1) \le \tau_{s-1}$ *and* $V^{\pi}(\overline{s-2}) \ge 0$, *then*

$$\min\big\{Q^{\pi}(s,a_0), Q^{\pi}(s,a_2)\big\} - Q^{\pi}(s,a_1) \ge (1-\gamma)/8.$$

(vii) Consider the softmax PG update rule (9)*. One has for any* $s \in \mathcal{S}$ *and any* θ,

$$\sum_a \frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta(s,a)} = 0 \qquad \textit{and} \qquad \sum_a \theta^{(t)}(s,a) = 0 \tag{53}$$

Remark 5. As it turns out, invoking Part (v) of Lemma 8 recursively reveals that: for any $2 \le s \le H$ and any $t < t_s(\tau_s)$, we have

$$V^{(t)}(s') < \gamma^{1/2}\tau_{s'} < \tau_{s'} \qquad \text{for all } s' \text{ obeying } s \le s' \le H. \tag{54}$$

This in turn implies that $t_2(\tau_2) \le t_3(\tau_3) \le \cdots \le t_H(\tau_H)$ according to the definition (24).

Let us point out some implications of Lemma 8 that help guide our lower-bound analysis. Once again, it is helpful to look at the results of this lemma when $\gamma \approx 1$ and $\gamma^H \approx 1$. In this case, the quantities defined in (14) obey $\tau_s \approx r_s \approx 1/2$, allowing us to obtain the following messages:

- Lemma 8(i) implies that, under mild conditions,

  $$Q^{\pi}(s,a_0) \approx Q^{\pi}(s,a_2) \approx 1/2$$

  holds any $s \in \{3, \cdots, H\}$ and any policy π. In comparison to the optimal values (23), this result uncovers the strict sub-optimality of actions a_0 and a_2, and indicates that one cannot possibly approach the optimal values unless $\pi(a_1 \,|\, s) \approx 1$.

- As further revealed by Lemma 8(ii), one needs to ensure a sufficiently large $\pi(a_1 \,|\, s)$ — i.e., $\pi(a_1 \,|\, s) \geq (1-\gamma)/2$ — in order to achieve $V^{\pi}(s) \gtrapprox 1/2$.

- Lemma 8(iii) establishes an intimate connection between $V^{\pi}(s)$ and $V^{\pi}(\overline{s})$: if we hope to attain $V^{\pi}(\overline{s}) \gtrapprox 1/2$ for an adjoint state $\overline{s}$, then one needs to first ensure that its associated primary state achieves $V^{\pi}(s) \gtrapprox 1/2$. The equivalence property (51) allows one to propagate the crossing time of state s to that of state $\overline{s}$.

- In Lemma 8(iv), we make clear that the Q-functions w.r.t. the buffer states are independent of the policy in use.

- Lemma 8(v) further establishes an intriguing connection between the crossing time of state s and that of the preceding state $s-1$.

- Lemma 8(vi) uncovers that: (a) if $V^{\pi}(s-2)$ is not sufficiently large, then the Q-value associated with (s, a_0) dominates the one associated with (s, a_2); (b) if $V^{\pi}(s-1)$ is not large enough, then the Q-value associated with (s, a_1) is dominated by that of the other two.

- As indicated by Lemma 8(vii), the sum of the iterate $\theta^{(t)}(s,a)$ over a remains unchanged throughout the execution of the algorithm.

Another key feature that permeates our analysis is a certain monotonicity property of value function estimates as the iteration count t increases, which we discuss in the sequel. To begin with, akin to the monotonicity properties of gradient descent (Beck, 2017), the unregularized PG update is known to achieve monotonic performance improvement in a pointwise manner, as summarized in the following lemma. The interested reader is referred to Agarwal et al. (2019, Lemma C.2) or details.

Lemma 9. *Consider the softmax PG method (9). One has*

$$V^{(t+1)}(s) \geq V^{(t)}(s) \qquad \text{and} \qquad Q^{(t+1)}(s,a) \geq Q^{(t)}(s,a)$$

for any state-action pair (s,a) and any $t \geq 0$, provided that $0 < \eta < (1-\gamma)^2/5$.

The preceding monotonicity feature, in conjunction with the uniform initialization scheme, ensures non-negativity of value function estimates throughout the execution of the algorithm.

Lemma 10. *Consider the softmax PG method (9), and suppose the initial policy $\pi^{(0)}(\cdot \,|\, s)$ for any $s \in \mathcal{S}$ is given by a uniform distribution over the action space $\mathcal{A}_s$ and $0 < \eta < (1-\gamma)^2/5$. Then one has*

$$\forall t \geq 0, \;\; \forall s \in \mathcal{S}: \qquad V^{(t)}(s) \geq 0.$$

Proof. The only negative rewards in our constructed MDP are $r(s_1, a_0)$ for $s_1 \in \mathcal{S}_1$ and $r(s_2, a_0)$ for $s_2 \in \mathcal{S}_2$. When $\pi^{(0)}(\cdot \,|\, s_1)$ is uniformly distributed, the MDP specification (18) gives

$$\forall s_1 \in \mathcal{S}_1: \qquad V^{(0)}(s_1) = 0.5 r(s_1, a_0) + 0.5 r(s_1, a_1) = 0.$$

Similarly, one has $V^{(0)}(s_2) = 0$ for all $s_2 \in \mathcal{S}_2$. Applying Lemma 9, we can demonstrate that $V^{(t)}(s) \geq V^{(0)}(s) \geq 0$ for any $s \in \mathcal{S}_1 \cup \mathcal{S}_2$ and any $t \geq 0$. From the Bellman equation, it is easily seen that the value function $V^{(t)}$ of any other state is a linear combination of $\{r(s,a) \,|\, s \notin \mathcal{S}_1, s \notin \mathcal{S}_2\}$, $\{V^{(t)}(s_1) \,|\, s_1 \in \mathcal{S}_1\}$ and $\{V^{(t)}(s_2) \,|\, s_2 \in \mathcal{S}_2\}$, which are all non-negative. It thus follows that $V^{(t)}(s) \geq 0$ for any $s \in \mathcal{S}$ and any $t \geq 0$. □

A.2 A type of recursive relations

In addition, we make note of a sort of recursive relations that appear commonly when studying the dynamics of gradient descent (Beck, 2017). The proof of the following lemma can be found in Appendix A.5.

Lemma 11. *Consider a positive sequence $\{x_t\}_{t \geq 0}$.*

(i) Suppose that $x_t \le x_{t-1}$ *for all* $t > 0$*. If there exists some quantity* $c_{\mathrm{l}} > 0$ *obeying* $c_{\mathrm{l}} x_0 \le 1/2$ *and*

$$x_t \ge x_{t-1} - c_{\mathrm{l}} x_{t-1}^2 \qquad \text{for all } t > 0, \tag{55a}$$

then one has

$$x_t \ge \frac{1}{2c_{\mathrm{l}} t + \frac{1}{x_0}} \qquad \text{for all } t \ge 0. \tag{55b}$$

(ii) If there exists some quantity $c_{\mathrm{u}} > 0$ *obeying*

$$x_t \le x_{t-1} - c_{\mathrm{u}} x_{t-1}^2 \qquad \text{for all } t > 0, \tag{56a}$$

then it follows that

$$x_t \le \frac{1}{c_{\mathrm{u}} t + \frac{1}{x_0}} \qquad \text{for all } t \ge 0. \tag{56b}$$

(iii) Suppose that $0 < x_t < c_x$ *for all* $t < t_0$ *and* $x_{t_0} \ge c_x$ *for some quantity* $c_x > 0$*. Assume that*

$$x_t \ge x_{t-1} + c_{-} x_{t-1}^2 \qquad \text{for all } 0 < t \le t_0 \tag{57a}$$

for some quantity $c_{-} > 0$*. Then one necessarily has*

$$t_0 \le \frac{1 + c_{-} c_x}{c_{-} x_0}. \tag{57b}$$

(iv) Suppose that

$$0 \le x_t \le x_{t-1} + c_{+} x_{t-1}^2 \qquad \text{for all } 0 < t \le t_0 \tag{58a}$$

for some quantity $c_{+} > 0$*. Then one necessarily has*

$$t_0 \ge \frac{\frac{1}{x_0} - \frac{1}{x_{t_0}}}{c_{+}}. \tag{58b}$$

A.3 Proof of Lemma 1

(i) Let us start with state 0. Given that this is an absorbing state and that $r(0, a_0) = 0$, we have $V^\star(0) = 0$.

(ii) Next, we turn to the buffer states in $\mathcal{S}_1$ and $\mathcal{S}_2$. For any $s_1 \in \mathcal{S}_1$, the Bellman equation gives

$$Q^\star(s_1, a_0) = r(s_1, a_0) + \gamma V^\star(0) = -\gamma^2; \tag{59a}$$

$$Q^\star(s_1, a_1) = r(s_1, a_1) + \gamma V^\star(0) = \gamma^2. \tag{59b}$$

This in turn implies that $V^\star(s_1) = Q^\star(s_1, a_1) = \gamma^2$. Repeating the same argument, we arrive at $V^\star(s_2) = Q^\star(s_2, a_1) = r(s_2, a_1) = \gamma^4$ for any $s_2 \in \mathcal{S}_2$.

(iii) We then move on to the adjoint states $\overline{1}$ and $\overline{2}$. From the construction (19), the Bellman equation yields

$$Q^\star(\overline{1}, a_0) = r(\overline{1}, a_0) + \gamma V^\star(0) = \gamma \tau_1 < \gamma/2,$$
$$Q^\star(\overline{1}, a_1) = r(\overline{1}, a_1) + \frac{\gamma}{|\mathcal{S}_1|} \sum_{s_1 \in \mathcal{S}_1} V^\star(s_1) = \frac{\gamma}{|\mathcal{S}_1|} \sum_{s_1 \in \mathcal{S}_1} V^\star(s_1) = \gamma^3,$$

where the last identity follows since $V^\star(s_1) = \gamma^2$. This in turn indicates that $V^\star(\overline{1}) = \max\{Q^\star(\overline{1}, a_0), Q^\star(\overline{1}, a_1)\} = \gamma^3$, provided that $\gamma^2 \ge 1/2$. Similarly, repeating this argument shows that $V^\star(\overline{2}) = \gamma^5$, as long as $\gamma^4 \ge 1/2$. As before, the optimal action in state $\overline{1}$ (resp. $\overline{2}$) is a_1.

(iv) The next step is to determine $V^\star(s)$ for any $s \in \{3, \cdots, H\}$. Suppose that $V^\star(\overline{s-2}) = \gamma^{2s-3}$ and $V^\star(\overline{s-1}) = \gamma^{2s-1}$. Then the construction (16) together with the Bellman equation yields

$$
\begin{aligned}
Q^\star(s, a_0) &= r(s, a_0) + \gamma V^\star(0) = r_s + \gamma^2 p \tau_{s-2} < 2/3; \\
Q^\star(s, a_1) &= r(s, a_1) + \gamma V^\star(\overline{s-1}) = \gamma \gamma^{2s-1} = \gamma^{2s}; \\
Q^\star(s, a_2) &= r(s, a_2) + \gamma(1-p) V^\star(0) + \gamma p V^\star(\overline{s-2}) = r_s + p\gamma^{2s-2} < 2/3.
\end{aligned}
$$

Consequently, one has $V^\star(s) = Q^\star(s, a_1) = \gamma^{2s}$ — namely, a_1 is the optimal action — as long as $\gamma^{2s} \geq 2/3$.

(v) We then turn attention to $V^\star(\overline{s})$ for any $\overline{s} \in \{\overline{3}, \cdots, \overline{H}\}$. Suppose that $V^\star(s) = \gamma^{2s}$. In view of the construction (17) and the Bellman equation, one has

$$
\begin{aligned}
Q^\star(\overline{s}, a_0) &= r(\overline{s}, a_0) + \gamma V^\star(0) = \gamma \tau_s < 1/2; \\
Q^\star(\overline{s}, a_1) &= r(\overline{s}, a_1) + \gamma V^\star(s) = \gamma^{2s+1}.
\end{aligned}
$$

Hence, we have $V^\star(\overline{s}) = Q^\star(\overline{s}, a_1) = \gamma^{2s+1}$ — with the optimal action being a_1 — provided that $\gamma^{2s+1} \geq 1/2$.

(vi) Applying an induction argument based on Steps (iii), (iv) and (v), we conclude that

$$
V^\star(s) = \gamma^{2s} \qquad \text{and} \qquad V^\star(\overline{s}) = \gamma^{2s+1} \tag{60}
$$

for all $3 \leq s \leq H$, with the proviso that $\gamma^{2H} \geq 2/3$ and $\gamma^{2H+1} \geq 1/2$.

(vii) In view of our MDP construction, a negative immediate reward (which is either $-\gamma^2$ or $-\gamma^4$) is accrued only when the current state lies in the buffer sets $\mathcal{S}_1$ and $\mathcal{S}_2$ and when action a_0 is executed. However, once a_0 is taken, the MDP will transition to the absorbing state 0, with all subsequent rewards frozen to 0. In conclusion, the entire MDP trajectory cannot receive negative immediate rewards more than once, thus indicating that $Q^\pi(s, a) \geq \min\{-\gamma^2, -\gamma^4\} = -\gamma^2$ irrespective of π and (s, a).

A.4 Proof of Lemma 8

Proof of Part (i). Before proceeding, we make note of a straightforward fact

$$
V^\pi(0) = 0, \tag{61}
$$

given that state 0 is an absorbing state and $r(0, a_0) = 0$.

For any $s \in \{3, \cdots, H\}$, the construction (16) together with (61) and the Bellman equation yields

$$
Q^\pi(s, a_0) = r(s, a_0) + \gamma V^\pi(0) = r_s + \gamma^2 p \tau_{s-2}; \tag{62a}
$$

$$
Q^\pi(s, a_1) = r(s, a_1) + \gamma V^\pi(\overline{s-1}) = \gamma V^\pi(\overline{s-1}); \tag{62b}
$$

$$
Q^\pi(s, a_2) = r(s, a_2) + \gamma(1-p) V^\pi(0) + \gamma p V^\pi(\overline{s-2}) = r_s + \gamma p V^\pi(\overline{s-2}). \tag{62c}
$$

Recalling the choices of τ_s, r_s and p in (14), we can continue the derivation in (62a) to reach

$$
\begin{aligned}
& Q^\pi(s, a_0) = 0.5\gamma^{\frac{2s}{3}+\frac{5}{6}} + c_{\mathrm{p}}(1-\gamma)\gamma^{\frac{2s}{3}+\frac{2}{3}} \\
\Longrightarrow \qquad & \gamma^{\frac{3}{2}} \tau_{s-1} = 0.5\gamma^{\frac{2s}{3}+\frac{5}{6}} \leq Q^\pi(s, a_0) \leq 0.5\gamma^{\frac{2s}{3}+\frac{1}{2}} = \gamma^{\frac{1}{2}} \tau_s.
\end{aligned}
$$

Here, the last inequality is valid when $c_{\mathrm{p}} \leq 1/6$, given that $\gamma^{\frac{1}{3}} + \frac{1-\gamma}{6}\gamma^{\frac{1}{6}} \leq 1$ holds for any $\gamma < 1$.

In addition, combining (62c) with (60), we arrive at

$$
Q^\pi(s, a_2) \leq r_s + \gamma p V^\star(\overline{s-2}) = 0.5\gamma^{\frac{2s}{3}+\frac{5}{6}} + c_{\mathrm{p}}(1-\gamma)\gamma^{2s-2} \leq 0.5\gamma^{\frac{2s}{3}+\frac{1}{2}} = \gamma^{\frac{1}{2}} \tau_s.
$$

This is guaranteed to hold when $c_{\mathrm{p}} \leq 1/6$, given that $\gamma^{\frac{1}{3}} + \frac{1-\gamma}{3}\gamma^{\frac{4s}{3}-\frac{5}{2}} \leq \gamma^{\frac{1}{3}} + \frac{1-\gamma}{3}\gamma^{\frac{3}{2}} \leq 1$ is valid for all $\gamma < 1$ and $s \geq 3$. Moreover, if one further has $V^\pi(\overline{s-2}) \geq 0$, then it is seen from (62c) that

$$
Q^\pi(s, a_2) \geq r_s = 0.5\gamma^{\frac{2s}{3}+\frac{5}{6}} = \gamma^{\frac{3}{2}} \tau_{s-1}. \tag{63}
$$

Proof of Part (ii). By virtue of the construction (16), we can invoke the Bellman equation to show that

$$
\begin{aligned}
V^{\pi}(s) &= \pi(a_0 \,|\, s)Q^{\pi}(s, a_0) + \pi(a_1 \,|\, s)Q^{\pi}(s, a_1) + \pi(a_2 \,|\, s)Q^{\pi}(s, a_2) \\
&= \pi(a_1 \,|\, s) \cdot \gamma V^{\pi}(\overline{s-1}) + \pi(a_0 \,|\, s)Q^{\pi}(s, a_0) + \pi(a_2 \,|\, s)Q^{\pi}(s, a_2) \\
&\leq \pi(a_1 \,|\, s)\gamma^{2s} + \big\{\pi(a_0 \,|\, s) + \pi(a_2 \,|\, s)\big\}\gamma^{\frac{1}{2}}\tau_s \\
&= \gamma^{2s}\pi(a_1 \,|\, s) + \gamma^{\frac{1}{2}}\tau_s\big(1 - \pi(a_1 \,|\, s)\big). \qquad (64)
\end{aligned}
$$

Here, the second identity comes from (62b), the penultimate line follows from (47), (60), as well as the facts $V^{\pi}(\overline{s-1}) \leq V^{\star}(\overline{s-1})$, while the last inequality exploits the fact $\pi(a_0 \,|\, s) + \pi(a_2 \,|\, s) = 1 - \pi(a_1 \,|\, s)$.

If $V^{\pi}(s) \geq \tau_s$, then this together with the upper bound (64) necessarily requires that

$$
\tau_s \leq \gamma^{2s}\pi(a_1 \,|\, s) + \gamma^{\frac{1}{2}}\tau_s\big(1 - \pi(a_1 \,|\, s)\big),
$$

which is equivalent to saying that

$$
\pi(a_1 \,|\, s) \geq \frac{\tau_s - \gamma^{\frac{1}{2}}\tau_s}{\gamma^{2s} - \gamma^{\frac{1}{2}}\tau_s} = \frac{1 - \gamma^{\frac{1}{2}}}{2\gamma^{\frac{4s}{3}} - \gamma^{\frac{1}{2}}} \geq \frac{1 - \gamma^{\frac{1}{2}}}{\gamma^{\frac{4s}{3}}} = \frac{1 - \gamma}{\gamma^{\frac{4s}{3}}(1 + \gamma^{\frac{1}{2}})} \geq \frac{1 - \gamma}{2}. \qquad (65)
$$

Putting these arguments together establishes the advertised result (48).

Proof of Part (iii). For any $\overline{s} \in \{\overline{3}, \cdots, \overline{H}\}$, in view of the construction (17) and the Bellman equation, one has

$$
\begin{aligned}
Q^{\pi}(\overline{s}, a_0) &= r(\overline{s}, a_0) + \gamma V^{\pi}(0) = \gamma\tau_s; \\
Q^{\pi}(\overline{s}, a_1) &= r(\overline{s}, a_1) + \gamma V^{\pi}(s) = \gamma V^{\pi}(s).
\end{aligned}
$$

Regarding state $\overline{1}$, we have

$$
\begin{aligned}
Q^{\pi}(\overline{1}, a_0) &= r(\overline{1}, a_0) + \gamma V^{\pi}(0) = \gamma\tau_1; \\
Q^{\pi}(\overline{1}, a_1) &= r(\overline{1}, a_1) + \gamma\frac{1}{|\mathcal{S}_1|}\sum\nolimits_{s' \in \mathcal{S}_1} V^{\pi}(s') = \gamma V^{\pi}(1).
\end{aligned}
$$

Similarly, one obtains $Q^{\pi}(\overline{2}, a_0) = \gamma\tau_2$ and $Q^{\pi}(\overline{2}, a_1) = \gamma V^{\pi}(2)$.

Next, let us decompose $V^{\pi}(\overline{s})$ as follows:

$$
\begin{aligned}
V^{\pi}(\overline{s}) &= \pi(a_0 \,|\, \overline{s})Q^{\pi}(\overline{s}, a_0) + \pi(a_1 \,|\, \overline{s})Q^{\pi}(\overline{s}, a_1) \\
&= \gamma\tau_s\pi(a_0 \,|\, \overline{s}) + \gamma\pi(a_1 \,|\, \overline{s})V^{\pi}(s) = \gamma\tau_s + \gamma\pi(a_1 \,|\, \overline{s})\big(V^{\pi}(s) - \tau_s\big),
\end{aligned}
$$

where we have used $\pi(a_0 \,|\, \overline{s}) + \pi(a_1 \,|\, \overline{s}) = 1$. From this relation and the assumption $\pi(a_1 \,|\, \overline{s}) > 0$, it is straightforward to see that $V^{\pi}(\overline{s}) \geq \gamma\tau_s$ if and only if $V^{\pi}(s) \geq \tau_s$. The claim (51) regarding $t_s(\tau_s)$ and $t_{\overline{s}}(\gamma\tau_s)$ then follows directly from the definition of t_s (see (24) and (25)).

Proof of Part (iv). For any $s_1 \in \mathcal{S}_1$, the Bellman equation yields

$$
\begin{aligned}
Q^{\pi}(s_1, a_0) &= r(s_1, a_0) + \gamma V^{\pi}(0) = -\gamma^2 + 0 = -\gamma^2, \\
Q^{\pi}(s_1, a_1) &= r(s_1, a_1) + \gamma V^{\pi}(0) = \gamma^2 + 0 = \gamma^2,
\end{aligned}
$$

and hence

$$
V^{\pi}(s_1) = \pi(a_0 \,|\, s_1)Q^{\pi}(s_1, a_0) + \pi(a_1 \,|\, s_1)Q^{\pi}(s_1, a_1) = -\gamma^2\pi(a_0 \,|\, s_1) + \gamma^2\pi(a_1 \,|\, s_1).
$$

A similar argument immediately yields that for any $s_2 \in \mathcal{S}_2$,

$$
Q^{\pi}(s_2, a_0) = -\gamma^4, \qquad Q^{\pi}(s_2, a_1) = \gamma^4, \qquad \text{and} \qquad V^{\pi}(s_2) = -\gamma^4\pi(a_0 \,|\, s_2) + \gamma^4\pi(a_1 \,|\, s_2).
$$

These together with our notation convention (21) establish (52).

Proof of Part (v). Suppose instead that $V^{\pi}(s-1) < \tau_{s-1}$. In view of the basic property (50) in Lemma 8, this necessarily requires that

$$V^{\pi}(\overline{s-1}) < \gamma\tau_{s-1}. \tag{66}$$

Taking (66) together with the relation (47b) allows us to reach

$$Q^{\pi}(s, a_1) = \gamma V^{\pi}(\overline{s-1}) < \gamma^2 \tau_{s-1} = \gamma^{\frac{4}{3}}\tau_s. \tag{67}$$

In addition, the properties (47a) and (47c) imply that

$$Q^{\pi}(s, a_0) < \gamma^{\frac{1}{2}}\tau_s \qquad \text{and} \qquad Q^{\pi}(s, a_2) < \gamma^{\frac{1}{2}}\tau_s.$$

Putting everything together implies that

$$V^{\pi}(s) \leq \max\Big\{Q^{\pi}(s, a_0), Q^{\pi}(s, a_1), Q^{\pi}(s, a_2)\Big\} < \gamma^{\frac{1}{2}}\tau_s,$$

which contractiсts the assumption $V^{\pi}(s) \geq \gamma^{\frac{1}{2}}\tau_s$. This establishes the claimed result for any $s \in \{3, \cdots, H\}$.

Proof of Part (vi). First, due to explicit expressions of the Q functions (62a) and (62c), one has

$$Q^{\pi}(s, a_0) - Q^{\pi}(s, a_2) = \gamma^2 p\tau_{s-2} - \gamma p V^{\pi}(\overline{s-2}) = \gamma p\big(\gamma\tau_{s-2} - V^{\pi}(\overline{s-2})\big) > 0,$$

where the last relation holds since $V^{\pi}(\overline{s-2}) < \gamma\tau_{s-2}$ when $V^{\pi}(s-2) < \tau_{s-2}$ (see (50)).

In addition, following the same derivation as for (67), we see that the condition $V^{\pi}(s-1) \leq \tau_{s-1}$ implies

$$Q^{\pi}(s, a_1) \leq \gamma^2\tau_{s-1}.$$

It is also seen from Part (i) of this lemma that

$$Q^{\pi}(s, a_0) \geq \gamma^{3/2}\tau_{s-1} \qquad \text{and} \qquad Q^{\pi}(s, a_2) \geq \gamma^{3/2}\tau_{s-1},$$

provided that $V^{\pi}(\overline{s-2}) \geq 0$. Combining these two inequalities, we arrive at the claimed bound

$$\min\big\{Q^{\pi}(s, a_0), Q^{\pi}(s, a_2)\big\} - Q^{\pi}(s, a_1) \geq \gamma^{3/2}\tau_{s-1} - \gamma^2\tau_{s-1} = \frac{\gamma^{3/2}\,(1-\gamma)\,\tau_{s-1}}{1+\gamma^{1/2}} \geq (1-\gamma)/8,$$

where the last inequality holds if $\gamma^{2s/3+5/6} \geq \gamma^s \geq 1/2$.

Proof of Part (vii). According to the update rule (9), we have — for any policy π — that

$$\begin{aligned}
\sum_a \frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta(s,a)} &= \sum_a \frac{1}{1-\gamma} d_{\mu}^{\pi_\theta}(s)\pi_\theta(a\,|\,s)\big(Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)\big) \\
&= \frac{1}{1-\gamma} d_{\mu}^{\pi_\theta}(s)\Big(\sum_a \pi_\theta(a\,|\,s)Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)\sum_a \pi_\theta(a\,|\,s)\Big) = 0,
\end{aligned}$$

where we have used the identities $\sum_a \pi_\theta(a\,|\,s) = 1$ and $V^{\pi}(s) = \sum_a \pi(a\,|\,s)Q^{\pi}(s,a)$. As a result, if $\sum_a \theta^{(0)}(s,a) = 0$, then it follows from the PG update rule that $\sum_a \theta^{(t)}(s,a) = 0$.

A.5 Proof of Lemma 11

Proof of Part (i). Dividing both sides of (55a) by $x_t x_{t-1}$, we obtain

$$\frac{1}{x_{t-1}} \geq \frac{1}{x_t} - \frac{c_1 x_{t-1}}{x_t}.$$

If $c_l x_0 \le 1/2$, then the monotonicity assumption gives $c_l x_t \le 1/2$ for all $t \ge 0$. It then follows that

$$\frac{x_t}{x_{t-1}} \ge 1 - c_l x_{t-1} \ge \frac{1}{2} \qquad \Longrightarrow \qquad \frac{1}{x_{t-1}} \ge \frac{1}{x_t} - \frac{c_l x_{t-1}}{x_t} \ge \frac{1}{x_t} - 2c_l.$$

Apply this relation recursively to deduce that

$$\frac{1}{x_t} \le \frac{1}{x_{t-1}} + 2c_l \le \cdots \le \frac{1}{x_0} + 2c_l t.$$

This readily concludes the proof of (55b).

Proof of Part (ii). Similarly, divide both sides of (56a) by $x_t x_{t-1}$ to derive

$$\frac{1}{x_{t-1}} \le \frac{1}{x_t} - \frac{c_u x_{t-1}}{x_t} \le \frac{1}{x_t} - c_u,$$

given the monotonicity and positivity assumption $0 < x_t \le x_{t-1}$. Invoking this inequality recursively gives

$$\frac{1}{x_t} \ge \frac{1}{x_{t-1}} + c_u \ge \cdots \ge \frac{1}{x_0} + c_u t,$$

thus establishing the advertised bound (56b).

Proof of Part (iii). We now turn attention to (57b). As is clearly seen, the non-negative sequence $\{x_t\}$ majorizes another sequence $\{y_t\}$ generated as follows (in the sense that $x_t \ge y_t$ for all $0 < t \le t_0$)

$$y_0 = x_0 \qquad \text{and} \qquad y_t = y_{t-1} + c_- y_{t-1}^2 \quad \text{for all } 0 < t \le t_0 \tag{68}$$

Dividing both sides of the second equation of (68) by $y_{t-1} y_t$, we reach

$$\frac{1}{y_{t-1}} = \frac{1}{y_t} + c_- \frac{y_{t-1}}{y_t} \ge \frac{1}{y_t} + \frac{c_-}{1 + c_- c_x}.$$

To see why the last inequality holds, note that, according to the first equation of (68) and the assumption $x_{t-1} < c_x$ (and hence $y_{t-1} \le x_{t-1} < c_x$), we have

$$\frac{y_t}{y_{t-1}} = 1 + c_- y_{t-1} \le 1 + c_- c_x.$$

As a result, we can apply the preceding inequalities recursively to derive

$$\frac{1}{y_0} \ge \frac{1}{y_1} + \frac{c_-}{1 + c_- c_x} \ge \cdots \ge \frac{1}{y_{t_0}} + \frac{c_-}{1 + c_- c_x} t_0 \ge \frac{c_-}{1 + c_- c_x} t_0,$$

and hence we arrive at (57b),

$$t_0 \le \frac{1 + c_- c_x}{c_- y_0} = \frac{1 + c_- c_x}{c_- x_0}.$$

Proof of Part (iv). The proof of (58b) is quite similar to that of (57b). Let us construct another non-negative sequence $\{z_t\}$ as follows

$$z_0 = x_0 \qquad \text{and} \qquad z_t = z_{t-1} + c_+ z_{t-1}^2 \quad \text{for all } 0 < t \le t_0. \tag{69}$$

Comparing this with (58a) clearly reveals that $z_t \ge x_t$. Divide both sides of (69) by $z_t z_{t-1}$ to reach

$$\frac{1}{z_{t-1}} = \frac{1}{z_t} + c_+ \frac{z_{t-1}}{z_t} \le \frac{1}{z_t} + c_+,$$

where the last inequality is valid since, by construction, $z_t \ge z_{t-1}$. Applying this relation recursively yields

$$\frac{1}{z_0} \le \frac{1}{z_{t_0}} + c_+ t_0,$$

which taken together with the fact $z_0 = x_0$ and $z_{t_0} \ge x_{t_0}$ leads to

$$t_0 \ge \frac{\frac{1}{z_0} - \frac{1}{z_{t_0}}}{c_+} \ge \frac{\frac{1}{x_0} - \frac{1}{x_{t_0}}}{c_+}.$$

B Discounted state visitation probability (Lemmas 2-3)

In this section, we establish our bounds concerning the discounted state visitation probability, as claimed in Lemma 2 and Lemma 3. Throughout this section, we denote by $\mathbb{P}(\cdot \,|\, \pi)$ the probability distribution when policy π is adopted. Also, we recall that μ is taken to be a uniform distribution over all states.

B.1 Lower bounds: proof of Lemma 2

Consider an arbitrary policy π, and let $\{s^k\}_{k\geq 0}$ represent an MDP trajectory. For any $s \in \{3, \cdots, H\}$, it follows from the definition (9c) of d_μ^π that

$$\begin{aligned}
d_\mu^\pi(s) &= (1-\gamma)\sum_{k=0}^{\infty}\gamma^k \mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi\big) \\
&\geq (1-\gamma)\gamma\mathbb{P}\big(s^1 = s \,|\, s^0 \sim \mu, \pi\big) \geq (1-\gamma)\gamma \sum_{s' \in \widehat{\mathcal{S}}_s} \mathbb{P}\big(s^1 = s \,|\, s^0 = s', \pi\big)\mathbb{P}(s_0 = s' \,|\, s^0 \sim \mu) \\
&= (1-\gamma)\gamma \cdot \frac{|\widehat{\mathcal{S}}_s|}{|\mathcal{S}|} = c_{\mathrm{m}}\gamma(1-\gamma)^2.
\end{aligned} \tag{70}$$

Here, the penultimate identity is valid due to the construction (20) and the assumption that μ is uniformly distributed, whereas the last identity results from the assumption (13). This establishes (29a). Repeating the same argument also reveals that

$$d_\mu^\pi(\overline{s}) \geq c_{\mathrm{m}}\gamma(1-\gamma)^2$$

for any $\overline{s} \in \{\overline{1}, \cdots, \overline{H}\}$, thus validating the lower bound (29b).

In addition, for any $s \in \mathcal{S}_1$, the MDP construction (20) allows one to derive

$$\begin{aligned}
d_\mu^\pi(s) &= (1-\gamma)\sum_{k=0}^{\infty}\gamma^k \mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi\big) \geq \gamma(1-\gamma)\mathbb{P}\big(s^1 = s \,|\, s^0 \sim \mu, \pi\big) \\
&\geq \gamma(1-\gamma)\mathbb{P}\big(s^1 = s \,|\, s^0 \in \widehat{\mathcal{S}}_1\big)\, \mathbb{P}\big(s_0 \in \widehat{\mathcal{S}}_1 \,|\, s^0 \sim \mu\big) \\
&= \gamma(1-\gamma)\cdot \frac{1}{|\mathcal{S}_1|}\cdot \frac{|\widehat{\mathcal{S}}_1|}{|\mathcal{S}|} = \gamma(1-\gamma)\frac{c_{\mathrm{m}}}{c_{\mathrm{b},1}}\cdot\frac{1}{|\mathcal{S}|}.
\end{aligned}$$

Here, the last line holds due to the fact that μ is uniformly distributed and the assumptions (12) and (13). We have thus concluded the proof for (29c). The proof for (29d) follows from an identical argument and is hence omitted.

B.2 Upper bounds: proof of Lemma 3

B.2.1 Preliminary facts

Before embarking on the proof, we collect several basic yet useful properties that happen when $t < t_s(\tau_s)$. The first-time readers can proceed directly to Appendix B.2.2.

Properties about $Q^{(t)}(\overline{s}, a)$. Combine the property (49) in Lemma 8 with (26) to yield that: for any $1 \leq s \leq H$ and any $t < t_s(\tau_s)$, one has

$$Q^{(t)}(\overline{s}, a_1) = \gamma V^{(t)}(s) < \gamma\tau_s = Q^{(t)}(\overline{s}, a_0) \tag{71}$$

In addition, combining the property (49) in Lemma 8 with (54) yields: for any $2 \leq s \leq H$,

$$V^{(t)}(\overline{s'}) \leq \max\Big\{Q^{(t)}\big(\overline{s'}, a_0\big), Q^{(t)}\big(\overline{s'}, a_1\big)\Big\} = \max\Big\{\gamma\tau_{s'}, \gamma V^{(t)}(s')\Big\} = \gamma\tau_{s'} \tag{72}$$

holds for all s' obeying $s \leq s' \leq H$ and all $t < t_s(\tau_s)$. As a remark, (71) indicates that a_1 remains unfavored (according to the current estimate $Q^{(t)}$) before the iteration number hits $t_s(\tau_s)$.

Properties about $Q^{(t)}(s+1, a)$ and $Q^{(t)}(s+2, a)$. First, combining (72) with the relation (62) reveals that: for any $2 \le s \le H-1$ and any $t < t_s(\tau_s)$,

$$Q^{(t)}(s+1, a_0) = r_{s+1} + \gamma^2 p \tau_{s-1} \ge r_{s+1}, \tag{73a}$$
$$Q^{(t)}(s+1, a_1) = \gamma V^{(t)}(\overline{s}) \le \gamma^2 \tau_s = \gamma^{1/2} r_{s+1}, \tag{73b}$$
$$Q^{(t)}(s+1, a_2) = r_{s+1} + \gamma p V^{(t)}(\overline{s-1}) \ge r_{s+1} \tag{73c}$$

hold as long as $V^{(t)}(\overline{s-1}) \ge 0$ (which is guaranteed by Lemma 10). Similarly, (72) and (62) also give

$$Q^{(t)}(s+2, a_0) = r_{s+2} + \gamma^2 p \tau_s$$
$$Q^{(t)}(s+2, a_1) = \gamma V^{(t)}(\overline{s+1}) = \gamma^2 \tau_{s+1} = \gamma^{1/2} r_{s+2},$$
$$Q^{(t)}(s+2, a_2) = r_{s+2} + \gamma p V^{(t)}(\overline{s}) \le r_{s+2} + \gamma^2 p \tau_s$$

for any $1 \le s \le H-2$ and any $t < t_s(\tau_s)$. Consequently, we have

$$Q^{(t)}(s+1, a_1) \le \min\left\{Q^{(t)}(s+1, a_0), Q^{(t)}(s+1, a_2)\right\}, \qquad \text{if } 2 \le s \le H-1 \tag{74a}$$
$$Q^{(t)}(s+2, a_2) \le Q^{(t)}(s+2, a_0), \qquad \text{if } 1 \le s \le H-2 \tag{74b}$$

for all $t < t_s(\tau_s)$. In other words, the above two inequalities reveal that actions a_1 and a_2 are perceived as suboptimal (based on the current Q-function estimates) before the iteration count surpasses $t_s(\tau_s)$.

Next, consider any $2 \le s \le H-1$ and any $t < t_s(\tau_s)$. It has already been shown above that

$$Q^{(t)}(s+1, a) \ge Q^{(t)}(s+1, a_1), \qquad a \in \{a_0, a_2\}. \tag{75a}$$

A similar argument also implies that, for any $t < t_s(\tau_s)$,

$$Q^{(t)}(s+2, a_0) \ge Q^{(t)}(s+2, a_2). \tag{75b}$$

B.2.2 Proof of the upper bounds (30a) and (30b)

We now turn attention to upper bounding $d_\mu^{(t)}(s)$ for any $s \in \{3, \cdots, H\}$. By virtue of the expansion (70), upper bounding $d_\mu^{(t)}(s)$ requires controlling $\mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi^{(t)}\big)$ for all $k \ge 0$. In light of this, our analysis consists of (i) developing upper bounds on the inter-related quantities $\mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi^{(t)}\big)$ and $\mathbb{P}\big(s^k = \overline{s} \,|\, s^0 \sim \mu, \pi^{(t)}\big)$ for any $k \ge 0$, and (ii) combining these upper bounds to control $d_\mu^{(t)}(s)$. At the core of our analysis is the following upper bounds on the t-th policy iterate, which will be established in Appendix B.2.6.

Lemma 12. *Under the assumption (28), for any $2 \le s \le H$ and any $t < t_s(\tau_s)$, one has*

$$\pi^{(t)}(a_1 \,|\, \overline{s}) \le \pi^{(t)}(a_0 \,|\, \overline{s}) \qquad \text{and} \qquad \pi^{(t)}(a_1 \,|\, \overline{s}) \le 1/2. \tag{76a}$$

Furthermore,

$$\pi^{(t)}(a_1 \,|\, s+1) \le \min\left\{\pi^{(t)}(a_0 \,|\, s+1), \pi^{(t)}(a_2 \,|\, s+1)\right\} \quad \text{and} \quad \pi^{(t)}(a_1 \,|\, s+1) \le 1/3 \tag{76b}$$

hold if $2 \le s \le H-1$, and

$$\pi^{(t)}(a_2 \,|\, s+2) \le \pi^{(t)}(a_0 \,|\, s+2) \qquad \text{and} \qquad \pi^{(t)}(a_2 \,|\, s+2) \le 1/2 \tag{76c}$$

hold if $1 \le s \le H-2$.

In words, Lemma 12 posits that, at the beginning, the policy iterate $\pi^{(t)}$ does not assign too much probability mass on actions that are currently perceived as suboptimal (see the remarks in Appendix B.2.1). With this lemma in place, we are positioned to establish the advertised upper bound.

Step 1: bounding $\mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi^{(t)}\big)$. For any $t < t_s(\tau_s)$ and any $s \in \{3, \cdots, H\}$, making use of the upper bound (76a) and the MDP construction in Section 3 yields

$$\begin{aligned}
&\mathbb{P}\big(s^0 = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) = 1/|\mathcal{S}|,\\
&\mathbb{P}\big(s^1 = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq \mathbb{P}\big(s^0 \in \widehat{\mathcal{S}}_s\big) + \pi^{(t)}(a_1 \,|\, \overline{s})\, \mathbb{P}(s^0 = \overline{s}) \leq \frac{|\widehat{\mathcal{S}}_s|}{|\mathcal{S}|} + \frac{1}{2}\cdot\frac{1}{|\mathcal{S}|} \leq \frac{2|\widehat{\mathcal{S}}_s|}{|\mathcal{S}|} = 2c_{\mathrm{m}}(1-\gamma),\\
&\mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) = \pi^{(t)}(a_1 \,|\, \overline{s})\, \mathbb{P}\big(s^{k-1} = \overline{s} \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq \frac{1}{2}\mathbb{P}\big(s^{k-1} = \overline{s} \,|\, s^0 \sim \mu, \pi^{(t)}\big)
\end{aligned}$$

for all $k \geq 2$. Note that the above calculation exploits the fact that μ is a uniform distribution.

Step 2: bounding $\mathbb{P}\big(s^k = \overline{s} \,|\, s^0 \sim \mu, \pi^{(t)}\big)$. Given that μ is a uniform distribution, one has

$$\mathbb{P}\big(s^0 = \overline{s} \,|\, s^0 \sim \mu, \pi^{(t)}\big) = 1/|\mathcal{S}| \tag{78a}$$

for any $s \in \mathcal{S}$. With (76b) and (76c) in mind, the MDP construction in Section 3 allows one to show that

$$\begin{aligned}
\mathbb{P}\big(s^1 = \overline{s} \,|\, s^0 \sim \mu, \pi^{(t)}\big) &\leq \mathbb{P}\big(s^0 \in \widehat{\mathcal{S}}_{\overline{s}}\big) + \pi^{(t)}(a_1 \,|\, s+1)\, \mathbb{P}(s^0 = s+1) + \pi^{(t)}(a_2 \,|\, s+2)\, \mathbb{P}(s^0 = s+2)\\
&\leq \frac{|\widehat{\mathcal{S}}_{\overline{s}}|}{|\mathcal{S}|} + \frac{1}{3|\mathcal{S}|} + \frac{1}{2|\mathcal{S}|} \leq \frac{2|\widehat{\mathcal{S}}_{\overline{s}}|}{|\mathcal{S}|} = 2c_{\mathrm{m}}(1-\gamma)
\end{aligned} \tag{78b}$$

holds for any $2 \leq s \leq H-2$ and any $t < t_s(\tau_s)$, and in addition,

$$\begin{aligned}
&\mathbb{P}\big(s^k = \overline{s} \,|\, s_0 \sim \mu, \pi^{(t)}\big)\\
&\quad\leq \pi^{(t)}(a_1 \,|\, s+1)\, \mathbb{P}\big(s^{k-1} = s+1 \,|\, s^0 \sim \mu, \pi^{(t)}\big) + \pi^{(t)}(a_2 \,|\, s+2)\, \mathbb{P}\big(s^{k-1} = s+2 \,|\, s^0 \sim \mu, \pi^{(t)}\big)\\
&\quad\leq \frac{1}{3}\mathbb{P}\big(s^{k-1} = s+1 \,|\, s_0 \sim \mu, \pi^{(t)}\big) + \frac{1}{2}\mathbb{P}\big(s^{k-1} = s+2 \,|\, s^0 \sim \mu, \pi^{(t)}\big)
\end{aligned} \tag{78c}$$

hold for any $k \geq 2$, $2 \leq s \leq H-2$, and any $t < t_s(\tau_s)$. Moreover, invoking (76b) and the MDP construction once again reveals that

$$\begin{aligned}
\mathbb{P}\big(s^k = \overline{H-1} \,|\, s^0 \sim \mu, \pi^{(t)}\big) &\leq \pi^{(t)}(a_1 \,|\, H)\, \mathbb{P}\big(s^{k-1} = H \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq \frac{1}{3}\mathbb{P}\big(s^{k-1} = H \,|\, s^0 \sim \mu, \pi^{(t)}\big)\\
\mathbb{P}\big(s^k = \overline{H} \,|\, s^0 \sim \mu, \pi^{(t)}\big) &= 0
\end{aligned}$$

hold for any $k \geq 2$ and any $t < t_s(\tau_s)$. In addition, it is seen that

$$\begin{aligned}
\mathbb{P}\big(s^1 = \overline{H-1} \,|\, s^0 \sim \mu, \pi^{(t)}\big) &\leq \mathbb{P}\big(s^0 \in \widehat{\mathcal{S}}_{\overline{H-1}}\big) + \mathbb{P}(s^0 = H) = \frac{\big|\widehat{\mathcal{S}}_{\overline{H-1}}\big|}{|\mathcal{S}|} + \frac{1}{|\mathcal{S}|} \leq \frac{2\big|\widehat{\mathcal{S}}_{\overline{H-1}}\big|}{|\mathcal{S}|} = 2c_{\mathrm{m}}(1-\gamma),\\
\mathbb{P}\big(s^1 = \overline{H} \,|\, s^0 \sim \mu, \pi^{(t)}\big) &\leq \mathbb{P}\big(s^0 \in \widehat{\mathcal{S}}_{\overline{H}}\big) = \frac{\big|\widehat{\mathcal{S}}_{\overline{H}}\big|}{|\mathcal{S}|} = c_{\mathrm{m}}(1-\gamma).
\end{aligned}$$

Step 3: putting all this together. Combining the preceding upper bounds on both $\mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi^{(t)}\big)$ and $\mathbb{P}\big(s^k = \overline{s} \,|\, s^0 \sim \mu, \pi^{(t)}\big)$ $(k \geq 1)$ and recognizing the monotonicity property (27), we immediately arrive at the following crude bounds

$$\begin{aligned}
&\max_{3 \leq s \leq H, t < t_s(\tau_s)} \Big\{\mathbb{P}\big(s^0 = s \,|\, s^0 \sim \mu, \pi^{(t)}\big), \mathbb{P}\big(s^1 = s \,|\, s^0 \sim \mu, \pi^{(t)}\big)\Big\} \leq 1/|\mathcal{S}| \leq 2c_{\mathrm{m}}(1-\gamma)\\
&\max_{2 \leq s \leq H, t < t_s(\tau_s)} \Big\{\mathbb{P}\big(s^0 = \overline{s} \,|\, s^0 \sim \mu, \pi^{(t)}\big), \mathbb{P}\big(s^1 = \overline{s} \,|\, s^0 \sim \mu, \pi^{(t)}\big)\Big\} \leq 2c_{\mathrm{m}}(1-\gamma)\\
&\max_{3 \leq s \leq H, t < t_s(\tau_s)} \mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq \frac{5}{6} \max_{2 \leq s \leq H, t < t_s(\tau_s)} \mathbb{P}\big(s^{k-1} = \overline{s} \,|\, s^0 \sim \mu, \pi^{(t)}\big)\\
&\max_{2 \leq s \leq H, t < t_s(\tau_s)} \mathbb{P}\big(s^k = \overline{s} \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq \frac{5}{6} \max_{3 \leq s \leq H, t < t_s(\tau_s)} \mathbb{P}\big(s^{k-1} = s \,|\, s^0 \sim \mu, \pi^{(t)}\big)
\end{aligned}$$

for any $k \geq 2$. It is then straightforward to deduce that

$$\max_{3 \leq s \leq H, t < t_s(\tau_s)} \mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq \left(\frac{5}{6}\right)^{k-1} 2c_{\mathrm{m}}(1-\gamma) \tag{79a}$$

$$\max_{2 \leq s \leq H, t < t_s(\tau_s)} \mathbb{P}\big(s^k = \overline{s} \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq \left(\frac{5}{6}\right)^{k-1} 2c_{\mathrm{m}}(1-\gamma) \tag{79b}$$

for any $k \geq 1$. In turn, these bounds give rise to

$$\begin{aligned} d_\mu^{(t)}(s) &= (1-\gamma)\sum_{k=0}^{\infty}\gamma^k \mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq (1-\gamma)\left\{2c_{\mathrm{m}}(1-\gamma) + \sum_{k=1}^{\infty}\left(\frac{5}{6}\right)^{k-1} 2c_{\mathrm{m}}(1-\gamma)\right\} \\ &\leq 2c_{\mathrm{m}}(1-\gamma)^2 + \frac{1}{1-5/6}\cdot 2c_{\mathrm{m}}(1-\gamma)^2 = 14c_{\mathrm{m}}(1-\gamma)^2 \end{aligned} \tag{80a}$$

for any $3 \leq s \leq H$ and any $t < t_s(\tau_s)$. This establishes the claimed upper bound (30a) as long as Lemma 12 is valid. Further, replacing s with $\overline{s}$ in (80) also reveals that

$$d_\mu^{(t)}(\overline{s}) \leq 14c_{\mathrm{m}}(1-\gamma)^2 \tag{80b}$$

for any $2 \leq s \leq H$ and any $t < t_s(\tau_s)$, thus concluding the proof of (30b).

B.2.3 Proof of the upper bound (30c)

We now consider any $s \in \mathcal{S}_2$. From our MDP construction, we have

$$\begin{aligned} \mathbb{P}\big(s^0 = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) &= 1/|\mathcal{S}|, \\ \mathbb{P}\big(s^1 = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) &\leq \mathbb{P}\big(s^1 = s \,|\, s^0 \in \widehat{\mathcal{S}}_2, \pi^{(t)}\big)\mathbb{P}\big(s^0 \in \widehat{\mathcal{S}}_2\big) + \mathbb{P}\big(s^1 = s \,|\, s^0 = \overline{2}, \pi^{(t)}\big)\,\mathbb{P}\big(s^0 = \overline{2}\big) \\ &\leq \frac{1}{|\mathcal{S}_2|}\frac{|\widehat{\mathcal{S}}_2|}{|\mathcal{S}|} + \frac{1}{|\mathcal{S}_2|}\frac{1}{|\mathcal{S}|} \leq \frac{2|\widehat{\mathcal{S}}_2|}{|\mathcal{S}_2|\,|\mathcal{S}|} = \frac{2c_{\mathrm{m}}}{c_{\mathrm{b},2}|\mathcal{S}|}, \\ \mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) &\leq \mathbb{P}\big(s^k = s \,|\, s^{k-1} = \overline{2}, \pi^{(t)}\big)\,\mathbb{P}\big(s^{k-1} = \overline{2} \,|\, s^0 \sim \mu, \pi^{(t)}\big) \\ &\leq \frac{1}{2|\mathcal{S}_2|}\mathbb{P}\big(s^{k-1} = \overline{2} \,|\, s^0 \sim \mu, \pi^{(t)}\big) \end{aligned}$$

for any $k \geq 2$ and any $s \in \mathcal{S}_2$. In addition, our bound in (79b) gives

$$\mathbb{P}\big(s^{k-1} = \overline{2} \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq \left(\frac{5}{6}\right)^{k-2} 2c_{\mathrm{m}}(1-\gamma)$$

for any $k \geq 2$ and any $t < t_2(\tau_2)$. Consequently, we arrive at

$$\mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq \frac{1}{|\mathcal{S}_2|}\mathbb{P}\big(s^{k-1} = \overline{2} \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq \frac{c_{\mathrm{m}}(1-\gamma)}{|\mathcal{S}_2|}\left(\frac{5}{6}\right)^{k-2} = \frac{c_{\mathrm{m}}}{c_{\mathrm{b},2}|\mathcal{S}|}\left(\frac{5}{6}\right)^{k-2}. \tag{82}$$

Armed with the preceding inequalities, we can derive

$$\begin{aligned} d_\mu^{(t)}(s) &= (1-\gamma)\sum_{k=0}^{\infty}\gamma^k \mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) \\ &\leq (1-\gamma)\left\{\frac{1}{|\mathcal{S}|} + \gamma\cdot\frac{2c_{\mathrm{m}}}{c_{\mathrm{b},2}|\mathcal{S}|} + \sum_{k=2}^{\infty}\gamma^k\frac{c_{\mathrm{m}}}{c_{\mathrm{b},2}|\mathcal{S}|}\left(\frac{5}{6}\right)^{k-2}\right\} \\ &\leq \frac{1-\gamma}{|\mathcal{S}|}\left(1 + \frac{2c_{\mathrm{m}}}{c_{\mathrm{b},2}}\right) + \frac{c_{\mathrm{m}}(1-\gamma)}{(1-5/6)c_{\mathrm{b},2}|\mathcal{S}|} = \frac{1-\gamma}{|\mathcal{S}|}\left(1 + \frac{8c_{\mathrm{m}}}{c_{\mathrm{b},2}}\right) \end{aligned}$$

for any $s \in \mathcal{S}_2$ and any $t < t_2(\tau_2)$, thus concluding the advertised upper bound for $s \in \mathcal{S}_2$.

B.2.4 Proof of the upper bound (30d)

It follows from our MDP construction that

$$\mathbb{P}\big(s^0 = \overline{1} \,|\, s^0 \sim \mu, \pi^{(t)}\big) = 1/|\mathcal{S}|,$$
$$\mathbb{P}\big(s^1 = \overline{1} \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq \mathbb{P}\big(s^0 \in \widehat{\mathcal{S}}_{\overline{1}}\big) + \mathbb{P}\big(s^0 = 3\big) = \frac{|\widehat{\mathcal{S}}_{\overline{1}}|}{|\mathcal{S}|} + \frac{1}{|\mathcal{S}|}.$$

Moreover, for any $k \geq 2$ and any $t < t_3(\tau_3)$, one can derive

$$\mathbb{P}\big(s^k = \overline{1} \,|\, s^0 \sim \mu, \pi^{(t)}\big) = \pi^{(t)}(a_2 \,|\, 3)\, \mathbb{P}\big(s^{k-1} = 3 \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq \left(\frac{5}{6}\right)^{k-2} 2c_{\mathrm{m}}(1-\gamma), \tag{83}$$

where the last inequality arises from (79a). Putting these bounds together leads to

$$\begin{aligned} d_\mu^{(t)}(\overline{1}) &= (1-\gamma)\sum_{k=0}^{\infty}\gamma^k \mathbb{P}\big(s^k = \overline{1} \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq (1-\gamma)\left\{\frac{1}{|\mathcal{S}|} + \gamma\left(\frac{|\widehat{\mathcal{S}}_{\overline{1}}|}{|\mathcal{S}|} + \frac{1}{|\mathcal{S}|}\right) + \sum_{k=2}^{\infty}\left(\frac{5}{6}\right)^{k-2} 2c_{\mathrm{m}}(1-\gamma)\right\} \\ &\leq (1-\gamma)\left\{\frac{2|\widehat{\mathcal{S}}_{\overline{1}}|}{|\mathcal{S}|} + \frac{1}{1-5/6}2c_{\mathrm{m}}(1-\gamma)\right\} = 14c_{\mathrm{m}}(1-\gamma)^2, \end{aligned}$$

where we have used the assumption that $|\widehat{\mathcal{S}}_{\overline{1}}| = c_{\mathrm{m}}(1-\gamma)|\mathcal{S}|$. When $t < t_2(\tau_2)$, the monotonicity property (27) indicates that $t < t_3(\tau_3)$, thus concluding the proof of (30d).

B.2.5 Proof of the upper bound (30e)

In view of our MDP construction, for any $s \in \mathcal{S}_1$ and any $t < \min\{t_1(\tau_1), t_2(\tau_2)\}$ we have

$$\begin{aligned} \mathbb{P}\big(s^0 = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) &= 1/|\mathcal{S}|, \\ \mathbb{P}\big(s^1 = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) &\leq \mathbb{P}\big(s^1 = s \,|\, s^0 \in \widehat{\mathcal{S}}_1, \pi^{(t)}\big)\mathbb{P}\big(s^0 \in \widehat{\mathcal{S}}_1\big) + \mathbb{P}\big(s^1 = s \,|\, s^0 = \overline{1}, \pi^{(t)}\big)\, \mathbb{P}\big(s^0 = \overline{1}\big) \\ &\leq \frac{1}{|\mathcal{S}_1|}\frac{|\widehat{\mathcal{S}}_1|}{|\mathcal{S}|} + \frac{1}{|\mathcal{S}_1|}\frac{1}{|\mathcal{S}|} \leq \frac{1}{|\mathcal{S}_1|}\frac{2|\widehat{\mathcal{S}}_1|}{|\mathcal{S}|} = \frac{2c_{\mathrm{m}}}{c_{\mathrm{b},1}|\mathcal{S}|}, \\ \mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) &\leq \mathbb{P}\big(s^k = s \,|\, s^{k-1} = \overline{1}, \pi^{(t)}\big)\, \mathbb{P}\big(s^{k-1} = \overline{1} \,|\, s^0 \sim \mu, \pi^{(t)}\big) \\ &\leq \frac{1}{|\mathcal{S}_1|}\mathbb{P}\big(s^{k-1} = \overline{1} \,|\, s^0 \sim \mu, \pi^{(t)}\big) \leq \frac{2c_{\mathrm{m}}}{c_{\mathrm{b},1}|\mathcal{S}|}\left(\frac{5}{6}\right)^{k-3}, \end{aligned}$$

where k is any integer obeying $k \geq 2$. Here, the last inequality comes from (83). These bounds taken collectively demonstrate that

$$\begin{aligned} d_\mu^{(t)}(s) &= (1-\gamma)\sum_{k=0}^{\infty}\gamma^k \mathbb{P}\big(s^k = s \,|\, s^0 \sim \mu, \pi^{(t)}\big) \\ &\leq (1-\gamma)\left\{\frac{1}{|\mathcal{S}|} + \gamma \cdot \frac{2c_{\mathrm{m}}}{c_{\mathrm{b},1}|\mathcal{S}|} + \sum_{k=2}^{\infty}\gamma^k \frac{2c_{\mathrm{m}}}{c_{\mathrm{b},1}|\mathcal{S}|}\left(\frac{5}{6}\right)^{k-3}\right\} \\ &\leq \frac{1-\gamma}{|\mathcal{S}|}\left(1 + \frac{2c_{\mathrm{m}}}{c_{\mathrm{b},1}}\right) + \frac{\frac{6}{5}\cdot 2c_{\mathrm{m}}(1-\gamma)}{(1-5/6)c_{\mathrm{b},1}|\mathcal{S}|} \leq \frac{1-\gamma}{|\mathcal{S}|}\left(1 + \frac{17c_{\mathrm{m}}}{c_{\mathrm{b},1}}\right) \end{aligned}$$

for any $s \in \mathcal{S}_1$ and any $t < \min\{t_1(\tau_1), t_2(\tau_2)\}$. This completes the proof.

B.2.6 Proof of Lemma 12

In order to prove this lemma, we are in need of the following auxiliary result, whose proof can be found in Appendix B.2.7.

Lemma 13. *Consider any state $1 \le s \le H$. Suppose that $0 < \eta \le (1-\gamma)/2$.*

(i) If the following conditions

$$Q^{(t)}(s,a_0) - Q^{(t)}(s,a_1) \ge 0, \qquad Q^{(t)}(s,a_2) - Q^{(t)}(s,a_1) \ge 0$$
$$\pi^{(t-1)}(a_1 \,|\, s) \le \min\left\{\pi^{(t-1)}(a_0 \,|\, s), \pi^{(t-1)}(a_2 \,|\, s)\right\}$$

hold, then one has $\pi^{(t)}(a_1 \,|\, s) \le 1/3$ and $\pi^{(t)}(a_1 \,|\, s) \le \min\left\{\pi^{(t)}(a_0 \,|\, s), \pi^{(t)}(a_2 \,|\, s)\right\}$.

(ii) If the following conditions

$$Q^{(t)}(s,a_0) - Q^{(t)}(s,a_2) \ge 0 \quad \textit{and} \quad \pi^{(t-1)}(a_2 \,|\, s) \le \pi^{(t-1)}(a_0 \,|\, s)$$

hold, then one has $\pi^{(t)}(a_2 \,|\, s) \le 1/2$ and $\pi^{(t)}(a_2 \,|\, s) \le \pi^{(t)}(a_0 \,|\, s)$.

(iii) If the following conditions

$$Q^{(t)}(\overline{s},a_0) - Q^{(t)}(\overline{s},a_1) \ge 0 \quad \textit{and} \quad \pi^{(t-1)}(a_1 \,|\, \overline{s}) \le \pi^{(t-1)}(a_0 \,|\, \overline{s})$$

hold, then one has $\pi^{(t)}(a_1 \,|\, \overline{s}) \le 1/2$ and $\pi^{(t)}(a_1 \,|\, \overline{s}) \le \pi^{(t)}(a_0 \,|\, \overline{s})$.

Remark 6. In words, Lemma 13 develops nontrivial upper bounds on the policy associated with actions that are currently perceived as suboptimal. As we shall see, such upper bounds — which are strictly below 1 — translate to some contraction factors that enable the advertised result of this lemma.

With Lemma 13 in place, we proceed to prove Lemma 12 by induction. Let us start from the base case with $t = 0$. Given that the initial policy is chosen to be uniformly distributed, we have

$$\begin{aligned} &\pi^{(0)}(a_1 \,|\, s) = \pi^{(0)}(a_0 \,|\, s) = \pi^{(0)}(a_2 \,|\, s), && 3 \le s \le H; \\ &\pi^{(0)}(a_1 \,|\, \overline{s}) = \pi^{(0)}(a_0 \,|\, \overline{s}), && 1 \le s \le H. \end{aligned}$$

Therefore, the claim (76) trivially holds for $t = 0$.

Next, we move on to the induction step. Suppose that the induction hypothesis (76) holds for the t-th iteration, and we intend to establish it for the $(t+1)$-th iteration. Apply Lemma 13 with Conditions (71) and (76a) to yield

$$\pi^{(t+1)}(a_1 \,|\, \overline{s}) \le \pi^{(t+1)}(a_0 \,|\, \overline{s})$$

with the proviso that $0 < \eta \le (1-\gamma)/2$. Clearly, this also implies that $\pi^{(t+1)}(a_1 \,|\, \overline{s}) \le 1/2$. Further, invoke Lemma 13 once again with Condition (75) and the induction hypothesis (76) to arrive at

$$\begin{aligned} &\pi^{(t+1)}(a_1 \,|\, s+1) \le \min\left\{\pi^{(t+1)}(a_0 \,|\, s+1), \pi^{(t+1)}(a_2 \,|\, s+1)\right\}, && \text{if } 2 \le s \le H-1; \\ &\pi^{(t+1)}(a_2 \,|\, s+2) \le \pi^{(t+1)}(a_0 \,|\, s+2), && \text{if } 1 \le s \le H-2. \end{aligned}$$

A straightforward consequence is $\pi^{(t+1)}(a_1 \,|\, s+1) \le 1/3$ and $\pi^{(t+1)}(a_2 \,|\, s+2) \le 1/2$. The proof is thus complete by induction.

B.2.7 Proof of Lemma 13

First of all, suppose that $Q^{(t)}(s,a_0) - Q^{(t)}(s,a_1) \ge 0$ and $\pi^{(t-1)}(a_0 \,|\, s) \ge \pi^{(t-1)}(a_1 \,|\, s)$ hold true. Combining this result with the PG update rule (9) gives

$$\begin{aligned} \theta^{(t)}(s,a_1) &= \theta^{(t-1)}(s,a_1) + \frac{\eta}{1-\gamma} d_\mu^{(t-1)}(s)\, \pi^{(t-1)}(a_1 \,|\, s)\, A^{(t-1)}(s,a_1) \\ &\le \theta^{(t-1)}(s,a_1) + \frac{\eta}{1-\gamma} d_\mu^{(t-1)}(s)\, \pi^{(t-1)}(a_1 \,|\, s)\, A^{(t-1)}(s,a_0). \end{aligned}$$

Consequently, applying this inequality and using the PG update rule (9) yield

$$\theta^{(t)}(s,a_1) - \theta^{(t)}(s,a_0)$$

$$
\begin{aligned}
&\le \theta^{(t-1)}(s,a_1) + \frac{\eta}{1-\gamma} d_\mu^{(t-1)}(s)\,\pi^{(t-1)}(a_1 \,|\, s)\,A^{(t-1)}(s,a_0) \\
&\qquad - \theta^{(t-1)}(s,a_0) - \frac{\eta}{1-\gamma} d_\mu^{(t-1)}(s)\,\pi^{(t-1)}(a_0 \,|\, s)\,A^{(t-1)}(s,a_0) \\
&\le \Big\{\theta^{(t-1)}(s,a_1) - \theta^{(t-1)}(s,a_0)\Big\} + \Big\{\pi^{(t-1)}(a_0 \,|\, s) - \pi^{(t-1)}(a_1 \,|\, s)\Big\}\left|\frac{\eta}{1-\gamma} d_\mu^{(t-1)}(s) A^{(t-1)}(s,a_0)\right|, \quad (85)
\end{aligned}
$$

where the last line arises by combining terms and invoking the assumption $\pi^{(t-1)}(a_0 \,|\, s) \ge \pi^{(t-1)}(a_1 \,|\, s)$.

Additionally, it is seen from the definition of the (unregularized) advantage function that

$$
\big|A^{(t-1)}(s,a_0)\big| \le \max_{\pi,a} \big|Q^\pi(s,a)\big| + \max_\pi \big|V^\pi(s)\big| \le 2, \tag{86}
$$

where the last inequality follows from Lemma 1. Recognizing that $d_\mu^{(t-1)}(s) \le 1$, one obtains

$$
\left|\frac{\eta}{1-\gamma} d_\mu^{(t-1)}(s) A^{(t-1)}(s,a_0)\right| \le \frac{\eta}{1-\gamma} \cdot 2 \le 1, \tag{87}
$$

with the proviso that $0 < \eta \le (1-\gamma)/2$.

Substituting (87) into (85) then yields

$$
\begin{aligned}
(85) &\le \Big\{\theta^{(t-1)}(s,a_1) - \theta^{(t-1)}(s,a_0)\Big\} + \Big\{\pi^{(t-1)}(a_0 \,|\, s) - \pi^{(t-1)}(a_1 \,|\, s)\Big\} \\
&\le \Big\{\theta^{(t-1)}(s,a_1) - \theta^{(t-1)}(s,a_0)\Big\} - \Big\{\theta^{(t-1)}(s,a_0) - \theta^{(t-1)}(s,a_1)\Big\} = 0, \quad (88)
\end{aligned}
$$

where both the first line and the last identity rely on the fact that $\theta^{(t-1)}(s,a_1) \le \theta^{(t-1)}(s,a_0)$ — an immediate consequence of the assumption $\pi^{(t-1)}(a_1 \,|\, s) \le \pi^{(t-1)}(a_0 \,|\, s)$. To see why the inequality (88) holds, it suffices to make note of the following consequence of softmax parameterization:

$$
\begin{aligned}
\pi^{(t-1)}(a_0 \,|\, s) - \pi^{(t-1)}(a_1 \,|\, s) &= \pi^{(t-1)}(a_1 \,|\, s)\Big\{\exp\big[\theta^{(t-1)}(s,a_0) - \theta^{(t-1)}(s,a_1)\big] - 1\Big\} \\
&\overset{\text{(a)}}{\le} \frac{\exp\big[\theta^{(t-1)}(s,a_0) - \theta^{(t-1)}(s,a_1)\big] - 1}{\exp\big[\theta^{(t-1)}(s,a_0) - \theta^{(t-1)}(s,a_1)\big] + 1} \\
&\overset{\text{(b)}}{\le} \theta^{(t-1)}(s,a_0) - \theta^{(t-1)}(s,a_1),
\end{aligned}
$$

where (b) follows since $\frac{e^x-1}{e^x+1} \le x$ for all $x \ge 0$, and the validity of (a) is guaranteed since

$$
\begin{aligned}
\pi^{(t-1)}(a_1 \,|\, s) &= \frac{\exp\big(\theta^{(t-1)}(s,a_1)\big)}{\sum_a \exp\big(\theta^{(t-1)}(s,a)\big)} \le \frac{\exp\big(\theta^{(t-1)}(s,a_1)\big)}{\exp\big(\theta^{(t-1)}(s,a_0)\big) + \exp\big(\theta^{(t-1)}(s,a_1)\big)} \\
&= \frac{1}{\exp\big[\theta^{(t-1)}(s,a_0) - \theta^{(t-1)}(s,a_1)\big] + 1}.
\end{aligned}
$$

To conclude, the above result (88) implies that

$$
\pi^{(t)}(a_0 \,|\, s) \ge \pi^{(t)}(a_1 \,|\, s). \tag{89}
$$

Repeating the above argument immediately reveals that: if

$$
Q^{(t-1)}(s,a_2) \ge Q^{(t-1)}(s,a_1) \quad \text{and} \quad \pi^{(t-1)}(a_2 \,|\, s) \ge \pi^{(t-1)}(a_1 \,|\, s),
$$

then one has $\pi^{(t)}(a_2 \,|\, s) \ge \pi^{(t)}(a_1 \,|\, s)$, which together with (89) indicates that

$$
\begin{aligned}
&\pi^{(t)}(a_1 \,|\, s) \le \min\big\{\pi^{(t)}(a_0 \,|\, s), \pi^{(t)}(a_2 \,|\, s)\big\} \\
\Longrightarrow \qquad &\pi^{(t)}(a_1 \,|\, s) \le \frac{\pi^{(t)}(a_0 \,|\, s) + \pi^{(t)}(a_1 \,|\, s) + \pi^{(t)}(a_2 \,|\, s)}{3} = \frac{1}{3}.
\end{aligned}
$$

This establishes Part (i) of Lemma 13.

The proof of Parts (ii) and (iii) follows from exactly the same argument, and is hence omitted.

C Crossing times of the first few states (Lemma 4)

This section presents the proof of Lemma 4 regarding the crossing times w.r.t. $\mathcal{S}_1$, $\mathcal{S}_2$, and state $\overline{1}$.

C.1 Crossing times for the buffer states in $\mathcal{S}_1$ and $\mathcal{S}_2$

We first present the proof of the relation (31a) regarding several quantities about t_1 and t_2.

Step 1: characterize the policy gradients. Our analysis largely relies on understanding the policy gradient dynamics, towards which we need to first characterize the gradient. Recalling that the gradient of $V^{(t)}$ w.r.t. $\theta_t(1, a_1)$ (cf. (9b)) is given by

$$\begin{aligned}
\frac{\partial V^{(t)}(\mu)}{\partial \theta(1, a_1)} &= \frac{1}{1-\gamma} d_\mu^{(t)}(1)\pi^{(t)}(a_1 \,|\, 1)\Big\{Q^{(t)}(1, a_1) - V^{(t)}(1)\Big\} \\
&= \frac{1}{1-\gamma} d_\mu^{(t)}(1)\pi^{(t)}(a_1 \,|\, 1)\Big\{Q^{(t)}(1, a_1) - \pi^{(t)}(a_0 \,|\, 1)Q^{(t)}(1, a_0) - \pi^{(t)}(a_1 \,|\, 1)Q^{(t)}(1, a_1)\Big\} \\
&= \frac{1}{1-\gamma} d_\mu^{(t)}(1)\pi^{(t)}(a_1 \,|\, 1)\pi^{(t)}(a_0 \,|\, 1)\Big\{Q^{(t)}(1, a_1) - Q^{(t)}(1, a_0)\Big\} \\
&= \frac{2\gamma^2}{1-\gamma} d_\mu^{(t)}(1)\pi^{(t)}(a_1 \,|\, 1)\pi^{(t)}(a_0 \,|\, 1) > 0,
\end{aligned} \tag{90}$$

where in the last step we use $Q^{(t)}(1, a_1) - Q^{(t)}(1, a_0) = 2\gamma^2$ (see (52)). The same calculation also yields

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(2, a_1)} = \frac{2\gamma^4}{1-\gamma} d_\mu^{(t)}(2)\pi^{(t)}(a_1 \,|\, 2)\pi^{(t)}(a_0 \,|\, 2) > 0. \tag{91}$$

As an immediate consequence, the PG update rule (9a) reveals that both $\theta^{(t)}(1, a_1)$ (resp. $\pi^{(t)}(1, a_1)$) and $\theta^{(t)}(2, a_1)$ (resp. $\pi^{(t)}(2, a_1)$) are monotonically increasing with t throughout the execution of the algorithm, which together with the initial condition $\pi^{(0)}(a_0 \,|\, 1) = \pi^{(0)}(a_1 \,|\, 1) = \pi^{(0)}(a_0 \,|\, 2) = \pi^{(0)}(a_1 \,|\, 2)$ as well as the identities $\theta^{(t)}(1, a_1) = -\theta^{(t)}(1, a_0)$ and $\theta^{(t)}(2, a_1) = -\theta^{(t)}(2, a_0)$ (due to (53)) gives

$$\pi^{(t)}(a_0 \,|\, 1) \le \pi^{(t)}(a_1 \,|\, 1) \qquad \text{and} \qquad \pi^{(t)}(a_0 \,|\, 2) \le \pi^{(t)}(a_1 \,|\, 2) \qquad \text{for all } t \ge 0. \tag{92}$$

Step 2: determine the range of $\pi^{(t)}(\cdot \,|\, 1)$ and $\pi^{(t)}(\cdot \,|\, 2)$. From the basic property (52), the value function of the buffer states in $\mathcal{S}_1$ — abbreviated by $V^{(t)}(1)$ as in the notation convention (21) — satisfies

$$V^{(t)}(1) = -\gamma^2 \pi^{(t)}(a_0 \,|\, 1) + \gamma^2 \pi^{(t)}(a_1 \,|\, 1) = -\gamma^2 + 2\gamma^2 \pi^{(t)}(a_1 \,|\, 1), \tag{93}$$

given that $\pi^{(t)}(a_0 \,|\, 1) + \pi^{(t)}(a_1 \,|\, 1) = 1$. Therefore, for any $t < t_1(\gamma^2 - 1/4)$ — which means $V^{(t)}(1) < \gamma^2 - 1/4$ according to the definition (25) — one has the following upper bound:

$$V^{(t)}(1) = -\gamma^2 + 2\gamma^2 \pi^{(t)}(a_1 \,|\, 1) < \gamma^2 - 1/4.$$

This is equivalent to requiring that

$$\pi^{(t)}(a_1 \,|\, 1) < 1 - (8\gamma^2)^{-1} \le 7/8 \tag{94}$$

and, consequently, $\pi^{(t)}(a_0 \,|\, 1) = 1 - \pi^{(t)}(a_1 \,|\, 1) \ge 1/8$ for any $t < t_1(\gamma^2 - 1/4)$. Putting this and (92) together further implies — for every $t < t_1(\gamma^2 - 1/4)$ — that:

$$1/8 \le \pi^{(t)}(a_0 \,|\, 1) \le \pi^{(t)}(a_1 \,|\, 1) \le 7/8. \tag{95}$$

Step 3: determine the range of policy gradients. In addition to showing the non-negativity of $\frac{\partial V^{(t)}(\mu)}{\partial \theta(1,a_1)}$ and $\frac{\partial V^{(t)}(\mu)}{\partial \theta(2,a_1)}$ for all $t \geq 0$, we are also in need of bounding their magnitudes. Towards this, invoke the property (95) to bound the derivative (90) by

$$\frac{7\gamma^2}{32(1-\gamma)} d_\mu^{(t)}(1) \leq \frac{\partial V^{(t)}(\mu)}{\partial \theta(1,a_1)} \leq \frac{\gamma^2}{2(1-\gamma)} d_\mu^{(t)}(1) \tag{96}$$

for any $t < t_1(\gamma^2 - 1/4)$, where we have used the elementary facts

$$\min_{1/8 \leq x \leq 7/8} x(1-x) = 7/64 \qquad \text{and} \qquad \max_{0 \leq x \leq 1} x(1-x) = 1/4.$$

Similarly, repeating the above argument with the gradient expression (91) leads to

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(2,a_1)} \leq \frac{\gamma^4}{2(1-\gamma)} d_\mu^{(t)}(2) \qquad \text{for all } t \geq 0; \tag{97a}$$

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(2,a_1)} \geq \frac{7\gamma^4}{32(1-\gamma)} d_\mu^{(t)}(2) \qquad \text{for all } 0 \leq t < t_2(\gamma^4 - 1/4). \tag{97b}$$

Further, note that Lemma 2 and Lemma 3 deliver upper and lower bounds on the quantities $d_\mu^{(t)}(1)$ and $d_\mu^{(t)}(2)$, which allow us to deduce that

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(1,a_1)} \geq \frac{7\gamma^3 c_{\mathrm{m}}}{32 c_{\mathrm{b},1}|\mathcal{S}|} \qquad \text{for all } t < t_1\big(\gamma^2 - 1/4\big); \tag{98a}$$

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(2,a_1)} \geq \frac{7\gamma^5 c_{\mathrm{m}}}{32 c_{\mathrm{b},2}|\mathcal{S}|} \qquad \text{for all } t < t_2(\gamma^4 - 1/4); \tag{98b}$$

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(1,a_1)} \leq \frac{\gamma^2(1 + 17c_{\mathrm{m}}/c_{\mathrm{b},1})}{2|\mathcal{S}|} \qquad \text{for all } t < \min\big\{t_1(\tau_1), t_2(\tau_2)\big\}; \tag{98c}$$

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(2,a_1)} \leq \frac{\gamma^4(1 + 8c_{\mathrm{m}}/c_{\mathrm{b},2})}{2|\mathcal{S}|} \qquad \text{for all } t < t_2(\tau_2). \tag{98d}$$

Step 4: develop an upper bound on $t_1(\gamma^2 - 1/4)$. The preceding bounds allow us to develop an upper bound on $t_1(\gamma^2 - 1/4)$. To do so, it is first observed from the fact $\theta^{(t)}(1,a_0) = -\theta^{(t)}(1,a_1)$ (due to (53)) that

$$\pi^{(t)}(a_1 \,|\, 1) = \frac{\exp\big(\theta^{(t)}(1,a_1)\big)}{\exp\big(\theta^{(t)}(1,a_0)\big) + \exp\big(\theta^{(t)}(1,a_1)\big)} = 1 - \frac{1}{1 + \exp\big(2\theta^{(t)}(1,a_1)\big)}.$$

Recognizing that $V^{(t)}(1) < \gamma^2 - 1/4$ occurs if and only if $\pi^{(t)}(a_1 \,|\, 1) < 1 - (8\gamma^2)^{-1}$ (see (94)), we can easily demonstrate that

$$\theta^{(t)}(1,a_1) \leq \frac{1}{2}\log\big(8\gamma^2 - 1\big) \leq \frac{1}{2}\log 7 \qquad \text{for all } t < t_1\big(\gamma^2 - 1/4\big). \tag{99}$$

If $t_1\big(\gamma^2 - 1/4\big) \geq \big\lceil \frac{32\log(7)c_{\mathrm{b},1}|\mathcal{S}|}{7\gamma^3 c_{\mathrm{m}}\eta} \big\rceil$, then taking $t = \big\lceil \frac{32\log(7)c_{\mathrm{b},1}|\mathcal{S}|}{7\gamma^3 c_{\mathrm{m}}\eta} \big\rceil$ together with (98) and (9a) yields

$$\theta^{(t)}(1,a_1) \geq \theta^{(0)}(1,a_1) + \eta\frac{7\gamma^3 c_{\mathrm{m}}}{32 c_{\mathrm{b},1}|\mathcal{S}|} t = \eta\frac{7\gamma^3 c_{\mathrm{m}}}{32 c_{\mathrm{b},1}|\mathcal{S}|} t \geq \log 7,$$

thus leading to contradiction with (99). As a result, one arrives at the following upper bound:

$$t_1(\tau_1) \leq t_1\big(\gamma^2 - 1/4\big) \leq \frac{32\log(7)c_{\mathrm{b},1}|\mathcal{S}|}{7\gamma^3 c_{\mathrm{m}}\eta} \leq \frac{15 c_{\mathrm{b},1}|\mathcal{S}|}{c_{\mathrm{m}}\eta}, \tag{100}$$

with the proviso that $\gamma \geq 0.85$ (so that $\tau_1 \leq \gamma^2 - 1/4$).

An upper bound on $t_2(\gamma^4 - 1/4)$ (and hence $t_2(\tau_2)$) can be obtained in a completely analogous manner

$$t_2(\tau_2) \leq t_2\big(\gamma^4 - 1/4\big) \leq \frac{15 c_{\mathrm{b},2}|\mathcal{S}|}{c_{\mathrm{m}}\eta},$$

provided that $\gamma \geq 0.95$ (so that $\tau_2 \leq \gamma^4 - 1/4$). We omit the proof of this part for the sake of brevity.

Step 5: develop a lower bound on $t_2(\tau_2)$. Repeating the argument in (94) and (99), we see that $V^{(t)}(2) \geq \tau_2$ if and only if $\pi^{(t)}(a_1 \,|\, 2) \geq \frac{1}{2} + \frac{\tau_2}{2\gamma^4}$, which is also equivalent to

$$\theta^{(t)}(2, a_1) \geq \frac{1}{2} \log \left(\frac{1}{\frac{1}{2} - \frac{\tau_2}{2\gamma^4}} - 1 \right) > \frac{1}{2} \log 3,$$

as long as $2\tau_2 > \gamma^4$. Of necessity, this implies that $\theta^{(t)}(2, a_1) > \frac{1}{2}\log 3$ when $t = t_2(\tau_2)$. If $t_2(\tau_2) \leq \frac{|\mathcal{S}| \log 3}{2\eta\gamma^4(1+8c_{\mathrm{m}}/c_{\mathrm{b},2})}$, then invoking (98) and (9a) and taking $t = t_2(\tau_2)$ yield

$$\theta^{(t)}(2, a_1) \leq \theta^{(0)}(2, a_1) + \eta \frac{\gamma^4}{2|\mathcal{S}|} \left(1 + \frac{8c_{\mathrm{m}}}{c_{\mathrm{b},2}} \right) t = \frac{\eta \gamma^4 t}{2|\mathcal{S}|} \left(1 + \frac{8c_{\mathrm{m}}}{c_{\mathrm{b},2}} \right) \leq \frac{1}{2} \log 3,$$

thus resulting in contradiction. We can thus conclude that

$$t_2(\tau_2) > \frac{|\mathcal{S}| \log 3}{\eta\gamma^4 \left(1 + 8c_{\mathrm{m}}/c_{\mathrm{b},2}\right)} > \frac{|\mathcal{S}| \log 3}{\eta(1 + 8c_{\mathrm{m}}/c_{\mathrm{b},2})}. \tag{101}$$

As an important byproduct, comparing (101) with (100) immediately reveals that

$$t_2(\tau_2) \geq t_1\big(\gamma^2 - 1/4\big) \geq t_1(\tau_1), \tag{102}$$

with the proviso that $\frac{\log 3}{1+8c_{\mathrm{m}}/c_{\mathrm{b},2}} \geq \frac{15c_{\mathrm{b},1}}{c_{\mathrm{m}}}$ and $\gamma \geq 0.87$ (so that $\gamma^2 - 1/4 > \tau_1$).

Step 6: develop a lower bound on $t_1(\tau_1)$. Repeat the analysis in (94) and (99) to show that: $V^{(t)}(1) \geq \tau_1$ if and only if

$$\theta^{(t)}(1, a_1) \geq \frac{1}{2} \log \left(\frac{1}{\frac{1}{2} - \frac{\tau_1}{2\gamma^2}} - 1 \right) > \frac{1}{2} \log 3.$$

Clearly, this lower bound should hold if $t = t_1(\tau_1)$. In addition, in view of (102), one has $\min\{t_1(\tau_1), t_2(\tau_2)\} = t_1(\tau_1)$. If $t_1(\tau_1) \leq \frac{|\mathcal{S}| \log 3}{\eta\gamma^2(1+17c_{\mathrm{m}}/c_{\mathrm{b},1})}$, then setting $t = t_1(\tau_1) = \min\{t_1(\tau_1), t_2(\tau_2)\}$ and applying (98) and (9a) lead to

$$\theta^{(t)}(1, a_1) \leq \theta^{(0)}(1, a_1) + \eta \frac{\gamma^2 (1 + 17c_{\mathrm{m}}/c_{\mathrm{b},1})}{2|\mathcal{S}|} t = \frac{\eta\gamma^2 t (1 + 17c_{\mathrm{m}}/c_{\mathrm{b},1})}{2|\mathcal{S}|} \leq \frac{1}{2} \log 3,$$

which is contradictory to the preceding lower bound. This in turn implies that

$$t_1(\tau_1) \geq \frac{|\mathcal{S}| \log 3}{\eta\gamma^2 (1 + 17c_{\mathrm{m}}/c_{\mathrm{b},1})} > \frac{|\mathcal{S}| \log 3}{\eta(1 + 17c_{\mathrm{m}}/c_{\mathrm{b},1})}. \tag{103}$$

C.2 Crossing times for the adjoint state $\overline{1}$

We now move on to the proof of (31b). Note that we have developed a lower bound on $t_2(\tau_2)$ in (101). In order to justify the advertised result $t_2(\tau_2) > t_{\overline{1}}\big(\gamma^3 - 1/4\big)$, it thus suffices to demonstrate that

$$t_{\overline{1}}\big(\gamma^3 - 1/4\big) \leq \frac{|\mathcal{S}| \log 3}{\eta(1 + 8c_{\mathrm{m}}/c_{\mathrm{b},2})}, \tag{104}$$

a goal we aim to accomplish in this subsection.

To do so, we divide into two cases. In the scenario where $t_1(\tau_1) \geq t_{\overline{1}}\big(\gamma^3 - 1/4\big)$, the bound (100) derived previously immediately leads to the desired bound:

$$t_{\overline{1}}\big(\gamma^3 - 1/4\big) \leq t_1(\tau_1) \leq \frac{15c_{\mathrm{b},1}|\mathcal{S}|}{c_{\mathrm{m}}\eta} \leq \frac{|\mathcal{S}| \log 3}{\eta(1 + 8c_{\mathrm{m}}/c_{\mathrm{b},2})},$$

with the proviso that $\frac{15c_{\mathrm{b},1}}{c_{\mathrm{m}}} \leq \frac{\log 3}{1+8c_{\mathrm{m}}/c_{\mathrm{b},2}}$. Consequently, the subsequent analysis concentrates on establishing (104) for the case where

$$t_1(\tau_1) < t_{\overline{1}}\big(\gamma^3 - 1/4\big).$$

In what follows, we divide into three stages and investigate each one separately, after presenting some basic gradient calculations that shall be invoked frequently.

Gradient characterizations. To begin with, observe from (9) that

$$\begin{aligned}\frac{\partial V^{(t)}(\mu)}{\partial\theta(\overline{1},a_1)} &= \frac{1}{1-\gamma}d_\mu^{(t)}(\overline{1})\pi^{(t)}\big(a_1|\overline{1}\big)\Big(Q^{(t)}(\overline{1},a_1)-V^{(t)}(\overline{1})\Big)\\
&= \frac{1}{1-\gamma}d_\mu^{(t)}(\overline{1})\pi^{(t)}\big(a_1\,|\,\overline{1}\big)\Big(Q^{(t)}(\overline{1},a_1)-\sum_{a\in\{a_0,a_1\}}\pi^{(t)}\big(a\,|\,\overline{1}\big)Q^{(t)}(\overline{1},a)\Big)\\
&= \frac{1}{1-\gamma}d_\mu^{(t)}(\overline{1})\pi^{(t)}\big(a_1\,|\,\overline{1}\big)\pi^{(t)}\big(a_0\,|\,\overline{1}\big)\Big(Q^{(t)}(\overline{1},a_1)-Q^{(t)}(\overline{1},a_0)\Big),\end{aligned}\tag{105a}$$

which makes use of the fact $\pi^{(t)}\big(a_0\,|\,\overline{1}\big)+\pi^{(t)}\big(a_1\,|\,\overline{1}\big)=1$. Analogously, we have

$$\frac{\partial V^{(t)}(\mu)}{\partial\theta(\overline{1},a_0)} = -\frac{1}{1-\gamma}d_\mu^{(t)}(\overline{s})\pi^{(t)}(a_1\,|\,\overline{1})\pi^{(t)}(a_0\,|\,\overline{1})\Big(Q^{(t)}(\overline{1},a_1)-Q^{(t)}(\overline{1},a_0)\Big).\tag{105b}$$

Stage 1: any t obeying $t<t_1(\tau_1)$. We start by looking at each term in the gradient expression (105a) separately. First, note that when $t<t_1(\tau_1)$, one has $V^{(t)}(1)<\tau_1$, which combined with (49) in Lemma 8 indicates that $Q^{(t)}(\overline{1},a_1)=\gamma V^{(t)}(1)<\gamma\tau_1=Q^{(t)}(\overline{1},a_0)$. In fact, from the definition (14a) of τ_1, the property (49) and Lemma 10, we have

$$1/2\geq Q^{(t)}(\overline{1},a_0)>Q^{(t)}(\overline{1},a_1)=\gamma V^{(t)}(1)\geq 0.$$

Additionally, recall that $t_1(\tau_1)<t_2(\tau_2)$ (see (102)). Lemma 3 then tells us that $d_\mu^{(t)}(\overline{1})\leq 14c_{\mathrm{m}}(1-\gamma)^2$ during this stage. Substituting these into (105a) and using $\pi^{(t)}(a_0\,|\,\overline{1})\leq 1$, we arrive at

$$0\geq\frac{\partial V^{(t)}(\mu)}{\partial\theta(\overline{1},a_1)}\geq -7c_{\mathrm{m}}(1-\gamma)\pi^{(t)}\big(a_1\,|\,\overline{1}\big),\tag{106}$$

which together with the PG update rule (9) also indicates that $\theta^{(t)}(\overline{1},a_1)$ (and hence $\pi^{(t)}\big(a_1\,|\,\overline{1}\big)$) is monotonically non-increasing with t in this stage. Invoke the auxiliary fact in Lemma 14 to reach

$$\pi^{(t+1)}\big(a_1\,|\,\overline{1}\big)-\pi^{(t)}\big(a_1\,|\,\overline{1}\big)\geq 2\eta\pi^{(t)}\big(a_1\,|\,\overline{1}\big)\frac{\partial V^{(t)}(\mu)}{\partial\theta(\overline{1},a_1)}\geq -14\eta c_{\mathrm{m}}(1-\gamma)\Big[\pi^{(t)}\big(a_1\,|\,\overline{1}\big)\Big]^2.$$

Taking the preceding recursive relation together with Lemma 11 and recalling the initialization $\pi^{(0)}\big(a_1\,|\,\overline{1}\big)=1/2$, we can guarantee that

$$\pi^{(t)}\big(a_1\,|\,\overline{1}\big)\geq\frac{1}{28\eta c_{\mathrm{m}}(1-\gamma)t+2}\qquad\text{for all } t\leq t_1(\tau_1)\tag{107}$$

provided that $14\eta c_{\mathrm{m}}(1-\gamma)\leq 1$. In conclusion, the above calculation precludes $\pi^{(t)}\big(a_1\,|\,\overline{1}\big)$ from decaying to zero too quickly, an observation that is particularly useful for our analysis in Stage 3.

Stage 2: any t obeying $t_1(\tau_1)\leq t<t_1(\gamma^2-1/4)$. The only step lies in extending the lower bound (107) to this stage. From the definition (25) of $t_1(\tau_1)$ as well as the monotonicity of $V^{(t)}(1)$ (see Lemma 9), we know that

$$V^{(t)}(1)\geq V^{(t_1(\tau_1))}(1)\geq\tau_1\qquad\text{for all } t\geq t_1(\tau_1),$$

provided that $\eta<(1-\gamma)^2/5$. This taken together with the property (49) in Lemma 8 reveals that

$$Q^{(t)}(\overline{1},a_1)-Q^{(t)}(\overline{1},a_0)\geq 0\qquad\text{for all } t\geq t_1(\tau_1),$$

and hence $\pi^{(t)}\big(a_1\,|\,\overline{1}\big)$ is non-decreasing in t during this stage. Therefore, we have

$$\pi^{(t)}\big(a_1\,|\,\overline{1}\big)\geq\pi^{(t_1(\tau_1))}\big(a_1\,|\,\overline{1}\big)\geq\frac{1}{28\eta c_{\mathrm{m}}(1-\gamma)t_1(\tau_1)+2},\qquad t\geq t_1(\tau_1),\tag{108}$$

where the first inequality follows from the non-decreasing property established above, and the second inequality follows from the lower bound (107). In fact, we have established a lower bound on $\pi^{(t)}\big(a_1\,|\,\overline{1}\big)$ that holds for the entire trajectory of the algorithm.

Stage 3: any t obeying $t_1(\gamma^2 - 1/4) \le t \le t_{\overline{1}}(\gamma^3 - 1/4)$. To facilitate analysis, we single out a time threshold t' as follows:

$$t' := \min\left\{t \,|\, \pi^{(t)}(a_0 \,|\, \overline{1}) < 1/2\right\}. \tag{109}$$

We begin by developing an upper bound on $\pi^{(t)}\big(a_0 \,|\, \overline{1}\big)$ for any $t \ge \max\{t', t_1(\gamma^2 - 1/4)\}$. Towards this, with the help of (49) in Lemma 8 we make the observation that: for any $t \ge t_1(\gamma^2 - 1/4)$, one has

$$Q^{(t)}(\overline{1}, a_1) - Q^{(t)}(\overline{1}, a_0) = \gamma V^{(t)}(1) - \gamma\tau_1 \ge \gamma\big(\gamma^2 - 1/4\big) - \gamma\tau_1 \ge 0.1 \tag{110}$$

as long as $\gamma \ge 0.92$, which combined with (105b) indicates that

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(\overline{1}, a_0)} < 0 \qquad \text{for all } t \ge t_1(\gamma^2 - 1/4). \tag{111}$$

Recognizing that $d_\mu^\pi(\overline{1}) \ge c_{\mathrm{m}}\gamma(1-\gamma)^2$ (see Lemma 2), we can continue the derivation (105b) to derive

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(\overline{1}, a_0)} \le -\frac{1}{1-\gamma}\gamma c_{\mathrm{m}}(1-\gamma)^2 \cdot \frac{1}{2} \cdot \pi^{(t)}(a_0 \,|\, \overline{1}) \cdot 0.1 = -0.05 c_{\mathrm{m}}\gamma(1-\gamma)\pi^{(t)}(a_0 \,|\, \overline{1})$$

for any $t \ge \max\{t', t_1(\gamma^2 - 1/4)\}$, which implies

$$\pi^{(t)}(a_1 \,|\, \overline{1}) \ge \pi^{(t')}(a_1 \,|\, \overline{1}) = 1 - \pi^{(t')}(a_0 \,|\, \overline{1}) \ge 1/2 \qquad \text{for any } t \ge t'.$$

Invoke Lemma 14 to arrive at

$$\pi^{(t+1)}\big(a_0 \,|\, \overline{1}\big) - \pi^{(t)}\big(a_0 \,|\, \overline{1}\big) \le \frac{\eta}{2}\pi^{(t)}\big(a_0 \,|\, \overline{1}\big)\frac{\partial V^{(t)}(\mu)}{\partial \theta(\overline{1}, a_0)} \le -\frac{\eta}{40}c_{\mathrm{m}}\gamma(1-\gamma)\Big[\pi^{(t)}(a_0 \,|\, \overline{1})\Big]^2,$$

provided that $2\eta\frac{\partial V^{(t)}(\mu)}{\partial \theta(\overline{1}, a_0)} \ge -1$, which is guaranteed by $\eta < (1-\gamma)/2$. Recalling that $\pi^{(t)}\big(a_0 \,|\, \overline{1}\big) \le 1/2$ for this entire stage, one can apply Lemma 11 to obtain

$$\pi^{(t)}\big(a_0 \,|\, \overline{1}\big) \le \frac{1}{\frac{\eta}{40}c_{\mathrm{m}}\gamma(1-\gamma)\Big(t - \max\big\{t', t_1(\gamma^2 - 1/4)\big\}\Big) + 2} \tag{112}$$

for any $t \ge \max\{t', t_1(\gamma^2 - 1/4)\}$.

With the above upper bound (112) in place, we are capable of showing that the target quantity $t_{\overline{1}}\big(\gamma^3 - 1/4\big)$ is not much larger than $\max\{t', t_1(\gamma^2 - 1/4)\}$. To show this, we first note that the value function of the adjoint state $\overline{1}$ obeys (see Part (iii) in Lemma 8)

$$\begin{aligned}
V^{(t)}(\overline{1}) &= \pi^{(t)}(a_0 \,|\, \overline{1})Q^{(t)}(\overline{1}, a_0) + \pi^{(t)}(a_1 \,|\, \overline{1})Q^{(t)}(\overline{1}, a_1) = \gamma\tau_1\pi^{(t)}(a_0 \,|\, \overline{1}) + \gamma\pi^{(t)}(a_1 \,|\, \overline{1})V^{(t)}(1) \\
&= \gamma\tau_1\pi^{(t)}(a_0 \,|\, \overline{1}) + \gamma V^{(t)}(1)\Big\{1 - \pi(a_0 \,|\, \overline{1})\Big\} \ge \gamma\tau_1\pi^{(t)}(a_0 \,|\, \overline{1}) + \gamma\left(\gamma^2 - 1/4\right)\Big\{1 - \pi^{(t)}(a_0 \,|\, \overline{1})\Big\} \\
&= \gamma\left\{\tau_1 - \gamma^2 + 1/4\right\}\pi^{(t)}(a_0 \,|\, \overline{1}) + \gamma^3 - \gamma/4,
\end{aligned}$$

where the inequality holds since $V^{(t)}(1) \ge \gamma^2 - 1/4$ in this stage (given that $t \ge t_1(\gamma^2 - 1/4)$). Recognizing that $0.5\gamma^{2/3} - \gamma^2 + 1/4 < 0$ for any $\gamma \ge 0.85$, we can rearrange terms to demonstrate that $V^{(t)}(\overline{1}) \ge \gamma^3 - 1/4$ holds once

$$\pi^{(t)}(a_0 \,|\, \overline{1}) \le \frac{1-\gamma}{4\gamma\left(\gamma^2 - 1/4 - 0.5\gamma^{2/3}\right)}.$$

In fact, for any $\gamma \ge 0.85$, the above inequality is guaranteed to hold as long as $\pi^{(t)}(a_0 \,|\, \overline{1}) \le 1 - \gamma$ since $4\gamma\left(\gamma^2 - 1/4 - 0.5\gamma^{2/3}\right) < 1$. In view of (112), we can achieve $\pi^{(t)}(a_0 \,|\, \overline{1}) \le 1 - \gamma$ as soon as $t - \max\big\{t', t_1(\gamma^2 - 1/4)\big\}$ surpasses $\frac{40}{c_{\mathrm{m}}\gamma\eta(1-\gamma)^2}$. As a consequence, we reach

$$t_{\overline{1}}\big(\gamma^3 - 1/4\big) \le \max\big\{t', t_1(\gamma^2 - 1/4)\big\} + \frac{40}{c_{\mathrm{m}}\gamma\eta(1-\gamma)^2}. \tag{113}$$

Armed with the relation (113), the goal of upper bounding $t_{\overline{1}}(\gamma^3 - 1/4)$ can be accomplished by controlling t'. To this end, we claim for the moment that

$$t' \leq \frac{1121 t_1(\gamma^2 - 1/4)}{\gamma}. \tag{114}$$

If this claim holds, then combining it with (113) and (100) would result in the advertised bound (104):

$$t_{\overline{1}}(\gamma^3 - 1/4) \leq \frac{9972 c_{\mathrm{b},1}|\mathcal{S}|}{\gamma^4 c_{\mathrm{m}}\eta} + \frac{40}{c_{\mathrm{m}}\gamma\eta(1-\gamma)^2} \leq \frac{|\mathcal{S}|}{4\gamma^4\eta} \leq \frac{|\mathcal{S}|\log 3}{\eta(1+8c_{\mathrm{m}}/c_{\mathrm{b},2})},$$

where the penultimate inequality relies on the assumptions $\frac{c_{\mathrm{b},1}}{c_{\mathrm{m}}} \leq \frac{1}{79776}$ and $|\mathcal{S}| \geq \frac{320\gamma^3}{c_{\mathrm{m}}(1-\gamma)^2}$, and the last one holds as long as $\frac{1}{4\gamma^4} \leq \frac{\log 3}{1+8c_{\mathrm{m}}/c_{\mathrm{b},2}}$. To finish up, it suffices to establish the claim (114).

Proof of the claim (114). It is sufficient to consider the case where $t' > t_1(\gamma^2 - 1/4)$; otherwise the inequality (114) is trivially satisfied. Since Lemma 2 tells us that $d_\mu^\pi(\overline{1}) \geq c_{\mathrm{m}}\gamma(1-\gamma)^2$, we can see from (105a) that, for any t with $t_1(\gamma^2 - 1/4) \leq t < t'$,

$$\begin{aligned}\frac{\partial V^{(t)}(\mu)}{\partial\theta(\overline{1}, a_1)} &= \frac{1}{1-\gamma} d_\mu^{(t)}(\overline{1})\pi^{(t)}(a_1 \,|\, \overline{1})\pi^{(t)}(a_0 \,|\, \overline{1})\Big\{Q^{(t)}(\overline{1}, a_1) - Q^{(t)}(\overline{1}, a_0)\Big\} \\ &\geq 0.05 c_{\mathrm{m}}\gamma(1-\gamma)\pi^{(t)}(a_1 \,|\, \overline{1}) > 0,\end{aligned}$$

where the last line follows by combining (110) and the fact that $\pi^{(t)}(a_0 \,|\, \overline{1}) \geq 1/2$ for any $t < t'$ (see the definition (109) of t'). According to Lemma 14, we can demonstrate that

$$\pi^{(t+1)}\big(a_1 \,|\, \overline{1}\big) - \pi^{(t)}\big(a_1 \,|\, \overline{1}\big) \geq \eta\pi^{(t)}\big(a_1 \,|\, \overline{1}\big)\frac{\partial V^{(t)}(\mu)}{\partial\theta(\overline{1}, a_1)} \geq 0.05\eta c_{\mathrm{m}}\gamma(1-\gamma)\Big[\pi^{(t)}\big(a_1 \,|\, \overline{1}\big)\Big]^2$$

for any t obeying $t_1(\gamma^2 - 1/4) \leq t < t'$. Invoking Lemma 11, we then have

$$\begin{aligned}t' \leq \frac{1 + 0.025\eta c_{\mathrm{m}}\gamma(1-\gamma)}{0.05\eta c_{\mathrm{m}}\gamma(1-\gamma)\pi^{(t_1(\gamma^2-1/4))}\big(a_1 \,|\, \overline{1}\big)} + t_1(\gamma^2 - 1/4) &< \frac{40}{\eta c_{\mathrm{m}}\gamma(1-\gamma)\pi^{(t_1(\gamma^2-1/4))}\big(a_1 \,|\, \overline{1}\big)} + t_1(\gamma^2 - 1/4) \\ &\leq \frac{40\big(28 t_1(\tau_1) + 2\big)}{\gamma} + t_1(\gamma^2 - 1/4) \\ &\leq \frac{1121 t_1(\gamma^2 - 1/4)}{\gamma}\end{aligned}$$

as claimed, where the second line follows from (108).

C.3 Auxiliary facts

In this subsection, we collect some elementary facts that have been used multiple times in the proof of Lemma 4. Specifically, the lemma below makes clear an explicit link between the gradient $\nabla_\theta V^{(t)}(\mu)$ and the difference between two consecutive policy iterates.

Lemma 14. *Consider any s whose associated action space is $\{a_0, a_1\}$.*

- *If $\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)} \leq 0$, then one has*

$$\pi^{(t+1)}\big(a_1 \,|\, s\big) - \pi^{(t)}\big(a_1 \,|\, s\big) \geq 2\eta\pi^{(t)}\big(a_1 \,|\, s\big)\frac{\partial V^{(t)}(\mu)}{\partial\theta(s, a_1)}. \tag{115}$$

- *If $\pi^{(t+1)}\big(a_0 \,|\, s\big) \geq 1/2$ and $-1 \leq 2\eta\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)} \leq 0$, then we have*

$$\pi^{(t+1)}\big(a_1 \,|\, s\big) - \pi^{(t)}\big(a_1 \,|\, s\big) \leq \frac{\eta}{2}\pi^{(t)}\big(a_1 \,|\, s\big)\frac{\partial V^{(t)}(\mu)}{\partial\theta(s, a_1)}. \tag{116}$$

- *If* $\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)} \geq 0$ *and if* $\pi^{(t+1)}\big(a_0 \,|\, s\big) \geq 1/2$*, then one has*

$$\pi^{(t+1)}\big(a_1 \,|\, s\big) - \pi^{(t)}\big(a_1 \,|\, s\big) \geq \eta\pi^{(t)}\big(a_1 \,|\, s\big)\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)}. \tag{117}$$

Proof of Lemma 14. We make note of the following elementary identity

$$\begin{aligned}\frac{e^{\theta_1}}{e^{\theta_1}+e^{-\theta_1}} - \frac{e^{\theta_2}}{e^{\theta_2}+e^{-\theta_2}} &= \frac{e^{\theta_1-\theta_2}-e^{-\theta_1+\theta_2}}{\big(e^{\theta_1}+e^{-\theta_1}\big)\big(e^{\theta_2}+e^{-\theta_2}\big)} = \frac{e^{-\theta_1}}{e^{\theta_1}+e^{-\theta_1}}\frac{e^{\theta_2}}{e^{\theta_2}+e^{-\theta_2}}\left(e^{2(\theta_1-\theta_2)}-1\right)\\ &= \left(1-\frac{e^{\theta_1}}{e^{\theta_1}+e^{-\theta_1}}\right)\frac{e^{\theta_2}}{e^{\theta_2}+e^{-\theta_2}}\left(e^{2(\theta_1-\theta_2)}-1\right),\end{aligned}$$

which allows us to write

$$\begin{aligned}\pi^{(t+1)}\big(a_1 \,|\, s\big) - \pi^{(t)}\big(a_1 \,|\, s\big) &= \pi^{(t+1)}\big(a_0 \,|\, s\big)\pi^{(t)}\big(a_1 \,|\, s\big)\left\{\exp\Big[2\theta_{t+1}(s,a_1)-2\theta_t(s,a_1)\Big]-1\right\}\\ &= \pi^{(t+1)}\big(a_0 \,|\, s\big)\pi^{(t)}\big(a_1 \,|\, s\big)\left\{\exp\left[2\eta\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)}\right]-1\right\}.\end{aligned} \tag{118}$$

- If $\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)} \leq 0$, then one can deduce that

$$(118) \geq 2\eta\pi^{(t+1)}\big(a_0 \,|\, s\big)\pi^{(t)}\big(a_1 \,|\, s\big)\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)} \geq 2\eta\pi^{(t)}\big(a_1 \,|\, s\big)\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)},$$

  where the first inequality relies on the elementary fact $e^x - 1 \geq x$ for all $x \in \mathbb{R}$, and the second one holds since $\pi^{(t+1)}\big(a_0 \,|\, s\big) \leq 1$ and $\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)} \leq 0$.

- If $-1 \leq 2\eta\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)} \leq 0$ and $\pi^{(t+1)}\big(a_0 \,|\, s\big) \geq 1/2$, then one has

$$(118) \leq \eta\pi^{(t+1)}\big(a_0 \,|\, s\big)\pi^{(t)}\big(a_1 \,|\, s\big)\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)} \leq \frac{\eta}{2}\pi^{(t)}\big(a_1 \,|\, s\big)\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)},$$

  where the first inequality comes from the elementary inequality $e^x - 1 \leq 0.5x$ for any $-1 \leq x \leq 0$, and the last inequality is valid since $\pi^{(t+1)}\big(a_0 \,|\, s\big) \geq 1/2$ and $\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)} \leq 0$.

- If $\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)} \geq 0$ and if $\pi^{(t+1)}\big(a_0 \,|\, s\big) \geq 1/2$, then it follows that

$$(118) \geq 2\eta\pi^{(t+1)}\big(a_0 \,|\, s\big)\pi^{(t)}\big(a_1 \,|\, s\big)\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)} \geq \eta\pi^{(t)}\big(a_1 \,|\, s\big)\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)},$$

  as claimed in (117).

□

D Analysis for the initial stage (Lemma 5)

This section establishes Lemma 5, which investigates the dynamics of $\theta^{(t)}(s,a)$ prior to the threshold $t_{s-2}(\tau_{s-2})$. Before proceeding, let us introduce a rescaled version of $\pi^{(t)}(s,a)$ that is sometimes convenient to work with:

$$\widehat{\pi}^{(t)}(s,a) \coloneqq \exp\Big(\theta^{(t)}(s,a) - \max_{a'\in\mathcal{A}_s}\theta^{(t)}(s,a')\Big) \tag{119}$$

for any state-action pair (s,a). This is orderwise equivalent to $\pi^{(t)}(s,a)$ since

$$\widehat{\pi}^{(t)}(s,a) = \frac{\exp\big(\theta^{(t)}(s,a)\big)}{\max_{a'\in\mathcal{A}_s}\exp\big(\theta^{(t)}(s,a')\big)} \geq \frac{\exp\big(\theta^{(t)}(s,a)\big)}{\sum_{a'\in\mathcal{A}_s}\exp\big(\theta^{(t)}(s,a')\big)} = \pi^{(t)}(s,a); \tag{120a}$$

$$\widehat{\pi}^{(t)}(s,a) = \frac{\exp\big(\theta^{(t)}(s,a)\big)}{\max_{a'\in\mathcal{A}_s}\exp\big(\theta^{(t)}(s,a')\big)} \leq \frac{\exp\big(\theta^{(t)}(s,a)\big)}{\frac{1}{3}\sum_{a'\in\mathcal{A}_s}\exp\big(\theta^{(t)}(s,a')\big)} = 3\pi^{(t)}(s,a). \tag{120b}$$

D.1 Two key properties

Our proof is based on the following claim: in order to establish the advertised results of Lemma 5, it suffices to justify the following two properties

$$\frac{1}{1+56c_{\mathrm{m}}\eta(1-\gamma)t} \le \widehat{\pi}^{(t)}(a_1 \,|\, s) \le \frac{1}{1+\frac{c_{\mathrm{m}}\gamma}{35}\eta(1-\gamma)^2 t} \tag{121}$$

$$\text{and} \qquad Q^{(t)}(s,a_2) - V^{(t)}(s) \ge 0 \tag{122}$$

hold for any $t \le t_{s-2}(\tau_{s-2})$. In light of this claim, our subsequent analysis consists of validating these two inequalities separately, which forms the main content of Section D.2.

We now move on to justify the above claim, namely, Lemma 5 is valid as long as the two key properties (121) and (122) hold true. First, recall that Lemma 12 together with (27) and Lemma 4 tells us that

$$\theta^{(t)}(s,a_0) \ge \theta^{(t)}(s,a_2) \ge \theta^{(t)}(s,a_1) \qquad \text{for all } t < t_{s-2}(\tau_{s-2}) \le t_{s-1}(\tau_{s-1}) \tag{123}$$

for any $3 \le s \le H$. Next, note that the gradient takes the following form (cf. (9))

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} d_{\mu}^{(t)}(s)\pi^{(t)}(a|s)\big(Q^{(t)}(s,a) - V^{(t)}(s)\big), \qquad a \in \{a_0, a_1, a_2\} \tag{124}$$

which together with the assumption $Q^{(t)}(s,a_2) - V^{(t)}(s) \ge 0$ (cf. (122)) implies that

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a_2)} \ge 0 \qquad \text{for all } t < t_{s-2}(\tau_{s-2}).$$

Consequently, $\theta^{(t)}(s,a_2)$ keeps increasing before t exceeds $t_{s-2}(\tau_{s-2})$. This combined with the relation (123), the initialization $\theta^{(0)}(s,a_0) = \theta^{(0)}(s,a_2) = 0$ and the constraint $\sum_a \theta^{(t)}(s,a) = 0$ (see Part (vii) of Lemma 8) reveals that

$$\theta^{(t)}(s,a_0) \ge \theta^{(t)}(s,a_2) \ge 0 \ge \theta^{(t)}(s,a_1) \qquad \text{for all } t < t_{s-2}(\tau_{s-2}), \tag{125}$$

thereby confirming the desired property (34).

Further, given the non-negativity of $\theta^{(t)}(s,a_2)$ stated in (125), one can readily derive

$$\begin{aligned}
\widehat{\pi}^{(t)}(a_1 \,|\, s) &= \exp\Big(\theta^{(t)}(s,a_1) - \max_{a'} \theta^{(t)}(s,a')\Big) = \exp\Big(\theta^{(t)}(s,a_1) - \theta^{(t)}(s,a_0)\Big) \\
&= \exp\Big(2\theta^{(t)}(s,a_1) + \theta^{(t)}(s,a_2)\Big) \ge \exp\Big(2\theta^{(t)}(s,a_1)\Big),
\end{aligned}$$

where the last line also makes use of the identity $\theta^{(t)}(s,a_0) = -\theta^{(t)}(s,a_1) - \theta^{(t)}(s,a_2)$ (see Part (vii) of Lemma 8). With this observation in mind, the assumed property (121) directly leads to the advertised result (33).

D.2 Proof of the properties (121) and (122)

This subsection presents the proofs of the two key properties, which are somewhat intertwined and require a bit of induction. Before proceeding, we make note of the initialization $\widehat{\pi}^{(0)}(a_1 \,|\, s) = 1$, which clearly satisfies the property (121) for this base case. Our proof consists of two steps to be detailed below. As can be easily seen, combining these two steps in an inductive manner immediately establishes both properties (121) and (122) for any $t \le t_{s-2}(\tau_{s-2})$.

Step 1: justifying (122) for the t-th iteration if (121) holds for the t-th iteration

We first turn to the proof of the inequality (122), assuming that (121) holds for the t-th iteration. According to (120) and (121), we have

$$\pi^{(t)}(a_1 \,|\, s) \ge \frac{1}{3}\widehat{\pi}^{(t)}(a_1 \,|\, s) \ge \frac{1}{3+168c_{\mathrm{m}}\eta(1-\gamma)t}. \tag{126}$$

By virtue of the auxiliary fact (134c) in Lemma 15 (see Section D.3), one has

$$Q^{(t)}(s,a_0) - Q^{(t)}(s,a_2) \leq \frac{\gamma p}{\frac{c_{\mathrm{m}}\gamma}{2}\eta(1-\gamma)t + \frac{1}{\gamma\tau_{s-2}}}. \tag{127}$$

Given that $p \coloneqq c_{\mathrm{p}}(1-\gamma)$ for some small constant $0 < c_{\mathrm{p}} < \frac{1}{2016}$, the above two inequalities allow one to ensure that

$$Q^{(t)}(s,a_0) - Q^{(t)}(s,a_2) < (\gamma^{3/2} - \gamma^2)\tau_{s-1}\pi^{(t)}(a_1 \,|\, s). \tag{128}$$

With the above relation in mind, we are ready to control $Q^{(t)}(s,a_2) - V^{(t)}(s)$ as follows

$$\begin{aligned}
Q^{(t)}(s,a_2) - V^{(t)}(s) &= \Big(\sum_a \pi^{(t)}(a \,|\, s)\Big)Q^{(t)}(s,a_2) - \sum_a \pi^{(t)}(a \,|\, s)Q^{(t)}(s,a) \\
&= \pi^{(t)}(a_1 \,|\, s)\Big(Q^{(t)}(s,a_2) - Q^{(t)}(s,a_1)\Big) - \pi^{(t)}(a_0 \,|\, s)\Big(Q^{(t)}(s,a_0) - Q^{(t)}(s,a_2)\Big) \\
&\geq \pi^{(t)}(a_1 \,|\, s)(\gamma^{3/2} - \gamma^2)\tau_{s-1} - \Big(Q^{(t)}(s,a_0) - Q^{(t)}(s,a_2)\Big) \\
&> 0.
\end{aligned}$$

Here, the second lines arise from the auxiliary facts in Lemma 15, while the last inequality is a consequence of (128). Then we complete the proof of the inequality (122).

Step 2: justifying (121) for the $(t+1)$-th iteration if (122) holds up to the t-th iteration

Suppose that the inequality (122) holds up to the t-th iteration. To validate (121) for the $(t+1)$-th iteration, we claim for the moment that

$$-14c_{\mathrm{m}}(1-\gamma)\widehat{\pi}^{(t)}(a_1 \,|\, s) \leq \frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a_1)} \leq -\frac{c_{\mathrm{m}}\gamma}{24}(1-\gamma)^2\widehat{\pi}^{(t)}(a_1 \,|\, s) < 0 \tag{129}$$

as long as $t \leq t_s(\tau_s)$. Let us take this claim as given, and return to prove it shortly.

Recall from (123) that

$$\theta^{(t)}(s,a_0) \geq \theta^{(t)}(s,a_2) \geq \theta^{(t)}(s,a_1)$$

and hence $\theta^{(t)}_{\mathsf{max}}(s) = \theta^{(t)}(s,a_0)$ is increasing with t during this stage, and as a result,

$$\theta^{(t+1)}(s,a_1) - \theta^{(t)}(s,a_1) + \theta^{(t)}_{\mathsf{max}}(s) - \theta^{(t+1)}_{\mathsf{max}}(s) \leq \theta^{(t+1)}(s,a_1) - \theta^{(t)}(s,a_1) \leq 0.$$

The gradient expression (124) combined with the satisfaction of (122) up to the t-th iteration implies that $\theta^{(t)}(s,a_2)$ is increasing up to the t-th iteration. Given that $\sum_a \theta^{(t)}(s,a) = 0$ (see Part (vii) of Lemma 8), we can derive

$$\begin{aligned}
\theta^{(t)}_{\mathsf{max}}(s) - \theta^{(t+1)}_{\mathsf{max}}(s) = \theta^{(t)}(s,a_0) - \theta^{(t+1)}(s,a_0) &= \theta^{(t+1)}(s,a_1) - \theta^{(t)}(s,a_1) + \theta^{(t+1)}(s,a_2) - \theta^{(t)}(s,a_2) \\
&\geq \theta^{(t+1)}(s,a_1) - \theta^{(t)}(s,a_1),
\end{aligned}$$

thus indicating that

$$\theta^{(t+1)}(s,a_1) - \theta^{(t)}(s,a_1) + \theta^{(t)}_{\mathsf{max}}(s) - \theta^{(t+1)}_{\mathsf{max}}(s) \geq 2\Big(\theta^{(t+1)}(s,a_1) - \theta^{(t)}(s,a_1)\Big).$$

These combined with Lemma 16 in Section D.3 guarantee that

$$\begin{aligned}
\widehat{\pi}^{(t+1)}(a_1 \,|\, s) - \widehat{\pi}^{(t)}(a_1 \,|\, s) &\geq 2\widehat{\pi}^{(t)}(a_1 \,|\, s)\left(\theta^{(t+1)}(s,a_1) - \theta^{(t)}(s,a_1)\right), \\
\widehat{\pi}^{(t+1)}(a_1 \,|\, s) - \widehat{\pi}^{(t)}(a_1 \,|\, s) &\leq 0.7\widehat{\pi}^{(t)}(a_1 \,|\, s)\left(\theta^{(t+1)}(s,a_1) - \theta^{(t)}(s,a_1)\right),
\end{aligned}$$

and as a consequence,

$$2\widehat{\pi}^{(t)}(a_1 \,|\, s) \cdot \eta \frac{\partial V^{(t)}(\mu)}{\partial \theta(s, a_1)} \le \widehat{\pi}^{(t+1)}(a_1 \,|\, s) - \widehat{\pi}^{(t)}(a_1 \,|\, s) \le 0.7 \widehat{\pi}^{(t)}(a_1 \,|\, s) \cdot \eta \frac{\partial V^{(t)}(\mu)}{\partial \theta(s, a_1)}.$$

Taking this collectively with (129), we reach

$$-28 c_{\mathrm{m}} \eta (1-\gamma) \Big[\widehat{\pi}^{(t)}(a_1 \,|\, s)\Big]^2 \le \widehat{\pi}^{(t+1)}(a_1 \,|\, s) - \widehat{\pi}^{(t)}(a_1 \,|\, s) \le -\frac{c_{\mathrm{m}}\gamma}{35} \eta (1-\gamma)^2 \Big[\widehat{\pi}^{(t)}(a_1 \,|\, s)\Big]^2. \tag{130}$$

Apply Lemma 11 together with the initialization $\widehat{\pi}^{(0)}(a_1 \,|\, s) = 1$ to arrive at

$$\frac{1}{1 + 56 c_{\mathrm{m}} \eta (1-\gamma)(t+1)} \le \widehat{\pi}^{(t+1)}(a_1 \,|\, s) \le \frac{1}{1 + \frac{c_{\mathrm{m}}\gamma}{35} \eta (1-\gamma)^2 (t+1)}. \tag{131}$$

Proof of the inequality (129)**.** Recall the gradient expression (124):

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(s, a_1)} = \frac{1}{1-\gamma} d_\mu^{(t)}(s) \pi^{(t)}(a_1 \,|\, s) \Big(Q^{(t)}(s, a_1) - V^{(t)}(s) \Big), \tag{132}$$

each term of which will be bounded separately.

The first step is to control $Q^{(t)}(s, a_1) - V^{(t)}(s)$, towards which we start with the following decomposition

$$\begin{aligned} Q^{(t)}(s, a_1) - V^{(t)}(s) &= Q^{(t)}(s, a_1) - \sum\nolimits_{a \in \{a_0, a_1, a_2\}} \pi^{(t)}(a \,|\, s) Q^{(t)}(s, a) \\ &= -\pi^{(t)}(a_0 \,|\, s) \Big(Q^{(t)}(s, a_0) - Q^{(t)}(s, a_1) \Big) - \pi^{(t)}(a_2 \,|\, s) \Big(Q^{(t)}(s, a_2) - Q^{(t)}(s, a_1) \Big). \end{aligned} \tag{133}$$

The auxiliary facts stated in Lemma 15 (see Appendix D.3) imply that

$$Q^{(t)}(s, a_0) - Q^{(t)}(s, a_1) \ge Q^{(t)}(s, a_2) - Q^{(t)}(s, a_1) \ge (\gamma^{3/2} - \gamma^2) \tau_{s-1},$$

while Lemma 1 and Lemma 10 tell us that

$$Q^{(t)}(s, a_0) - Q^{(t)}(s, a_1) \le V^\star(s) - 0 = \gamma^{2s}.$$

At the same time, the auxiliary fact (134a) in Lemma 15 (see Appendix D.3) taken together with the gradient expression (9b) guarantees that

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(s, a_0)} \ge \frac{\partial V^{(t)}(\mu)}{\partial \theta(s, a_2)} \ge \frac{\partial V^{(t)}(\mu)}{\partial \theta(s, a_1)}$$

and hence $\theta^{(t)}(s, a_1) \le \theta^{(t)}(s, a_2) \le \theta^{(t)}(s, a_0)$ (or equivalently $\pi^{(t)}(a_1 \,|\, s) \le \pi^{(t)}(a_2 \,|\, s) \le \pi^{(t)}(a_0 \,|\, s)$) during this stage. As a result,

$$\pi^{(t)}(a_1 \,|\, s) \le 1/3 \qquad \text{and} \qquad 1 \ge \pi^{(t)}(a_0 \,|\, s) + \pi^{(t)}(a_2 \,|\, s) \ge 2/3.$$

Substituting the preceding bounds into the decomposition (133), we arrive at

$$\begin{aligned} Q^{(t)}(s, a_1) - V^{(t)}(s) &\le -\Big(\pi^{(t)}(a_0 \,|\, s) + \pi^{(t)}(a_2 \,|\, s)\Big) \min\Big\{ Q^{(t)}(s, a_0) - Q^{(t)}(s, a_1), Q^{(t)}(s, a_2) - Q^{(t)}(s, a_1) \Big\} \\ &\le -\frac{2}{3}(\gamma^{\frac{3}{2}} - \gamma^2)\tau_{s-1} = -\frac{2}{3} \frac{\gamma^{\frac{3}{2}} \tau_{s-1}}{1 + \sqrt{\gamma}} (1-\gamma) \le -\frac{1-\gamma}{8}, \end{aligned}$$

provided that $\gamma \ge 0.85$. Meanwhile, it follows from Lemma 1 and Lemma 10 that

$$Q^{(t)}(s, a_1) - V^{(t)}(s) \ge 0 - V^\star(s) \ge -1.$$

Further, from Lemma 3, we have learned that $c_{\mathrm{m}}\gamma(1-\gamma)^2 \le d_\mu^{(t)}(s) \le 14 c_{\mathrm{m}} (1-\gamma)^2$ for any $t \le t_s(\tau_s)$. Substituting the above bounds into (132) and invoking (120), we establish the desired inequality (129).

D.3 Auxiliary facts

We now gather a few basic facts that are useful throughout this section. The first lemma presents some preliminary facts regarding the difference of Q-function estimates across different actions in the current setting; the proof is deferred to Appendix D.3.1.

Lemma 15. *Consider any $t < t_{s-2}(\tau_{s-2})$. Under the assumption (28), the following are satisfied*

$$Q^{(t)}(s, a_0) > Q^{(t)}(s, a_2) > Q^{(t)}(s, a_1), \tag{134a}$$

$$Q^{(t)}(s, a_2) - Q^{(t)}(s, a_1) \geq (\gamma^{3/2} - \gamma^2)\tau_{s-1}, \tag{134b}$$

$$Q^{(t)}(s, a_0) - Q^{(t)}(s, a_2) \leq \frac{\gamma p}{\frac{c_{\mathrm{m}}\gamma}{2}\eta(1-\gamma)t + \frac{1}{\gamma\tau_{s-2}}}. \tag{134c}$$

Remark 7. Lemma 15 makes clear that — before t exceeds $t_{s-2}(\tau_{s-2})$ — action a_0 is perceived as the best choice, with a_1 being the least favorable one. In the meantime, it also reveals that (i) $Q^{(t)}(s, a_2)$ is considerably larger than $Q^{(t)}(s, a_1)$, while (ii) the gap between $Q^{(t)}(s, a_0)$ and $Q^{(t)}(s, a_2)$ decays at least as rapidly as $O(1/t)$ in this stage.

The second lemma is concerned with the consecutive difference between two rescaled policy iterates. The proof can be found in Appendix D.3.2.

Lemma 16. *Suppose that $0 < \eta \leq (1-\gamma)/6$. For any $t \geq 0$ and any $3 \leq s \leq H$, define $\theta^{(t)}_{\mathsf{max}}(s) \coloneqq \max_a \theta^{(t)}(s, a)$. If we write*

$$\widehat{\pi}^{(t+1)}(a_1 \mid s) - \widehat{\pi}^{(t)}(a_1 \mid s) = c\widehat{\pi}^{(t)}(a_1 \mid s)\Big(\theta^{(t+1)}(s, a_1) - \theta^{(t)}(s, a_1) + \theta^{(t)}_{\mathsf{max}}(s) - \theta^{(t+1)}_{\mathsf{max}}(s)\Big) \tag{135}$$

for some $c \in \mathbb{R}$, then we necessarily have

$$\begin{aligned} c &\in [1, 1.5) \qquad \textit{if } \theta^{(t+1)}(s, a_1) \geq \theta^{(t)}(s, a_1) \quad \textit{and} \quad \theta^{(t)}_{\mathsf{max}}(s) \geq \theta^{(t+1)}_{\mathsf{max}}(s);\\ c &\in (0.72, 1] \qquad \textit{if } \theta^{(t+1)}(s, a_1) \leq \theta^{(t)}(s, a_1) \quad \textit{and} \quad \theta^{(t)}_{\mathsf{max}}(s) \leq \theta^{(t+1)}_{\mathsf{max}}(s). \end{aligned}$$

D.3.1 Proof of Lemma 15

In view of Lemma 10, one has $V^{(t)}(\overline{s-2}) \geq 0$ for all $t \geq 0$. Therefore, the relation (47) yields

$$Q^{(t)}(s, a_2) = r_s + \gamma p V^{(t)}(\overline{s-2}) \geq r_s = \gamma^{3/2}\tau_{s-1}.$$

In addition, for any $t < t_{s-2}(\tau_{s-2}) \leq t_{s-1}(\tau_{s-1}) \leq t_{\overline{s-1}}(\gamma\tau_{s-1})$ (see Lemma 8 and Lemma 4), we have $V^{(t)}(\overline{s-1}) < \gamma\tau_{s-1}$, and hence it is seen from the relation (47) that

$$Q^{(t)}(s, a_2) - Q^{(t)}(s, a_1) = Q^{(t)}(s, a_2) - \gamma V^{(t)}(\overline{s-1}) \geq (\gamma^{3/2} - \gamma^2)\tau_{s-1} > 0,$$

as claimed in (134b). Also, Part (i) of Lemma 8 tells us that

$$Q^{(t)}(s, a_0) - Q^{(t)}(s, a_2) = r_s + \gamma^2 p\tau_{s-2} - r_s - \gamma p V^{(t)}(\overline{s-2}) = \gamma p\Big(\gamma\tau_{s-2} - V^{(t)}(\overline{s-2})\Big) \geq 0,$$

where the last inequality holds for any $t < t_{s-2}(\tau_{s-2})$ (see Part (iii) of Lemma 8). These taken together validate (134a).

It remains to justify (134c), which is the content of the rest of this proof. The main step lies in demonstrating that, for any $t < t_s(\tau_s)$ and any $1 \leq s \leq H$,

$$\gamma\tau_s - V^{(t)}(\overline{s}) \leq \frac{1}{\frac{c_{\mathrm{m}}\gamma}{2}\eta(1-\gamma)t + \frac{1}{\gamma\tau_s}}. \tag{136}$$

If this were true, than taking it together with the following property (which is a consequence of (47))

$$Q^{(t)}(s, a_0) - Q^{(t)}(s, a_2) = \gamma p\left(\gamma\tau_{s-2} - V^{(t)}(\overline{s-2})\right), \tag{137}$$

would establish the inequality (134c). It then boils down to justifying (136). Towards this, we first make the observation that

$$\begin{aligned}
V^{(t)}(\overline{s}) - \gamma\tau_s &= \pi^{(t)}(a_0 \,|\, \overline{s})Q^{(t)}(\overline{s}, a_0) + \pi^{(t)}(a_1 \,|\, \overline{s})Q^{(t)}(\overline{s}, a_1) - \gamma\tau_s \\
&= \pi^{(t)}(a_0 \,|\, \overline{s})\gamma\tau_s + \pi^{(t)}(a_1 \,|\, \overline{s})Q^{(t)}(\overline{s}, a_1) - \gamma\tau_s \\
&= \pi^{(t)}(a_1 \,|\, \overline{s})\Big(Q^{(t)}(\overline{s}, a_1) - \gamma\tau_s\Big),
\end{aligned} \tag{138}$$

where the second line holds since $Q^{(t)}(\overline{s}, a_0) = \gamma\tau_s$ (see (49)). Additionally, recall from the definition that for any $t < t_s(\tau_s)$, one has $V^{(t)}(s) < \tau_s$ and hence

$$\begin{aligned}
\frac{\partial V^{(t)}(\mu)}{\partial\theta(\overline{s}, a_1)} &= \frac{1}{1-\gamma} d_\mu^{(t)}(\overline{s})\pi^{(t)}(a_1 \,|\, \overline{s})\Big(Q^{(t)}(\overline{s}, a_1) - \pi^{(t)}(a_0 \,|\, \overline{s})Q^{(t)}(\overline{s}, a_0) - \pi^{(t)}(a_1 \,|\, \overline{s})Q^{(t)}(\overline{s}, a_1)\Big) \\
&= \frac{1}{1-\gamma} d_\mu^{(t)}(\overline{s})\pi^{(t)}(a_0 \,|\, \overline{s})\pi^{(t)}(a_1 \,|\, \overline{s})\Big(Q^{(t)}(\overline{s}, a_1) - Q^{(t)}(\overline{s}, a_0)\Big) \\
&= \frac{1}{1-\gamma} d_\mu^{(t)}(\overline{s})\pi^{(t)}(a_0 \,|\, \overline{s})\pi^{(t)}(a_1 \,|\, \overline{s})\Big(\gamma V^{(t)}(s) - \gamma\tau_s\Big) < 0,
\end{aligned} \tag{139}$$

where the last line makes use of the identities in (49). This means that $\theta^{(t)}(\overline{s}, a_1)$ keeps decreasing, and hence $\theta^{(t)}(\overline{s}, a_1) \le 0$ given the initialization $\theta^{(0)}(\overline{s}, a_1) = 0$. As an immediate consequence, one has $\theta^{(t)}(\overline{s}, a_0) = -\theta^{(t)}(\overline{s}, a_1) \ge 0$ and $\pi^{(t)}(a_0 \,|\, \overline{s}) \ge 1/2$. Taking this observation together with (139) and Lemma 2 gives

$$\begin{aligned}
\frac{\partial V^{(t)}(\mu)}{\partial\theta(\overline{s}, a_1)} &= \frac{1}{1-\gamma} d_\mu^{(t)}(\overline{s})\pi^{(t)}(a_0 \,|\, \overline{s})\pi^{(t)}(a_1 \,|\, \overline{s})\Big(Q^{(t)}(\overline{s}, a_1) - Q^{(t)}(\overline{s}, a_0)\Big) \\
&= \frac{1}{1-\gamma} d_\mu^{(t)}(\overline{s})\pi^{(t)}(a_0 \,|\, \overline{s})\pi^{(t)}(a_1 \,|\, \overline{s})\Big(Q^{(t)}(\overline{s}, a_1) - \gamma\tau_s\Big) \\
&\le \frac{c_{\mathrm{m}}\gamma}{2}(1-\gamma)\pi^{(t)}(a_1 \,|\, \overline{s})\Big(Q^{(t)}(\overline{s}, a_1) - \gamma\tau_s\Big) < 0.
\end{aligned}$$

Moreover, combine (139) with Lemma 3 and Lemma 1 to yield

$$\left|\frac{\partial V^{(t)}(\mu)}{\partial\theta(\overline{s}, a_1)}\right| \le \frac{1}{1-\gamma} d_\mu^{(t)}(\overline{s})\Big|Q^{(t)}(\overline{s}, a_1) - \gamma\tau_s\Big| \le \frac{1}{1-\gamma} 14c_{\mathrm{m}}(1-\gamma)^2\Big(\big|Q^{(t)}(\overline{s}, a_1)\big| + \gamma\tau_s\Big) \le 28c_{\mathrm{m}}(1-\gamma),$$

If $28c_{\mathrm{m}}\eta(1-\gamma) < 1/2$, then the above inequalities taken together with Lemma 14 give

$$\pi^{(t+1)}(a_1 \,|\, \overline{s}) - \pi^{(t)}(a_1 \,|\, \overline{s}) \le \frac{c_{\mathrm{m}}\gamma}{2}\eta(1-\gamma)\Big[\pi^{(t)}(a_1 \,|\, \overline{s})\Big]^2\Big(Q^{(t)}(\overline{s}, a_1) - \gamma\tau_s\Big) \tag{140}$$

for all $t < t_s(\tau_s)$. This combined with (138) and the monotonicity of $Q^{(t)}(\overline{s}, a_1)$ (see Lemma 9) gives

$$\begin{aligned}
\gamma\tau_s - V^{(t+1)}(\overline{s}) &= \pi^{(t+1)}(a_1 \,|\, \overline{s})\Big(\gamma\tau_s - Q^{(t+1)}(\overline{s}, a_1)\Big) \le \pi^{(t+1)}(a_1 \,|\, \overline{s})\Big(\gamma\tau_s - Q^{(t)}(\overline{s}, a_1)\Big) \\
&\le \Big\{\pi^{(t)}(a_1 \,|\, \overline{s}) - \frac{c_{\mathrm{m}}\gamma}{2}\eta(1-\gamma)\Big[\pi^{(t)}(a_1 \,|\, \overline{s})\Big]^2\Big(\gamma\tau_s - Q^{(t)}(\overline{s}, a_1)\Big)\Big\}\Big(\gamma\tau_s - Q^{(t)}(\overline{s}, a_1)\Big) \\
&= \Big\{1 - \frac{\eta c_{\mathrm{m}}\gamma(1-\gamma)}{2}\Big(\gamma\tau_s - V^{(t)}(\overline{s})\Big)\Big\}\Big(\gamma\tau_s - V^{(t)}(\overline{s})\Big),
\end{aligned}$$

where the penultimate line follows from the inequality (140) for the iteration $t-1$, and the last identity makes use of (138). In conclusion, we have arrived at the following inductive relation

$$\gamma\tau_s - V^{(t+1)}(\overline{s}) \le \gamma\tau_s - V^{(t)}(\overline{s}) - \frac{\eta c_{\mathrm{m}}\gamma(1-\gamma)}{2}\Big(\gamma\tau_s - V^{(t)}(\overline{s})\Big)^2,$$

which bears resemblance to the recursive relations studied in Lemma 11. Recognizing that $\gamma\tau_s - V^{(0)}(\overline{s}) \le \gamma\tau_{s-2}$ (since $V^{(0)}(\overline{s}) \ge 0$ according to Lemma 10), we can invoke Lemma 11 to derive

$$\gamma\tau_s - V^{(t)}(\overline{s}) \le \frac{1}{\frac{c_{\mathrm{m}}\gamma}{2}\eta(1-\gamma)t + \frac{1}{\gamma\tau_s}}.$$

Putting the above pieces together concludes the proof of (134c).

D.3.2 Proof of Lemma 16

From the definition (119), direct calculations lead to

$$\begin{aligned}\widehat{\pi}^{(t+1)}(a_1 \,|\, s) - \widehat{\pi}^{(t)}(a_1 \,|\, s) &= \exp\Big(\theta^{(t+1)}(s,a_1) - \theta^{(t+1)}_{\mathsf{max}}(s)\Big) - \exp\Big(\theta^{(t)}(s,a_1) - \theta^{(t)}_{\mathsf{max}}(s)\Big) \\ &= \widehat{\pi}^{(t)}(a_1 \,|\, s)\Big\{\exp\Big(\theta^{(t+1)}(s,a_1) - \theta^{(t)}(s,a_1) + \theta^{(t)}_{\mathsf{max}}(s) - \theta^{(t+1)}_{\mathsf{max}}(s)\Big) - 1\Big\}.\end{aligned}$$

According to Lemma 1, we have $|Q^{(t)}(s,a)| \leq 1$ and $|V^{(t)}(s)| \leq 1$, which indicates — for any action $a \in \{a_0, a_1, a_2\}$ — that

$$\Big|\theta^{(t+1)}(s,a) - \theta^{(t)}(s,a)\Big| = \Big|\eta \frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a)}\Big| = \frac{\eta}{1-\gamma} d_{\mu}^{(t)}(s)\pi^{(t)}(a \,|\, s)\Big|Q^{(t)}(s,a) - V^{(t)}(s)\Big| \leq \frac{1}{3},$$

provided that $\eta \leq (1-\gamma)/6$. An immediate consequence is that $|\theta^{(t+1)}_{\mathsf{max}}(s) - \theta^{(t)}_{\mathsf{max}}(s)| \leq 1/3$ and hence

$$\Big|\theta^{(t+1)}(s,a_1) - \theta^{(t)}(s,a_1) + \theta^{(t)}_{\mathsf{max}}(s) - \theta^{(t+1)}_{\mathsf{max}}(s)\Big| \leq 2/3.$$

This taken together with the following elementary facts

$$(e^x - 1)/x \in [1, 1.5) \quad \text{for } 0 \leq x < 2/3, \qquad \text{and} \qquad (e^x - 1)/x \in (0.72, 1] \quad \text{for } -2/3 < x \leq 0$$

establishes the claim (135).

E Analysis for the intermediate stage (Lemma 6)

We now turn attention to Lemma 6, which studies the dynamics during an intermediate stage between $t_{s-2}(\tau_{s-2})$ and $t_{\overline{s-1}}(\tau_s)$.

E.1 Main steps

Key facts regarding crossing times. Our proof for Lemma 6 relies on several crucial properties regarding the crossing times for both the key primary states and the adjoint states, as stated in the following two lemmas.

Lemma 17. *Suppose that (28) holds. There exists some constant $0 < c_0 \leq \frac{1222}{c_{\mathrm{m}}\gamma}$ such that:*

$$t_s\big(\gamma^{2s} - 1/4\big) - \max\Big\{t_{\overline{s-1}}\big(\gamma^{2s-1} - 1/4\big),\, t_s(\tau_s)\Big\} \leq \frac{c_0}{\eta(1-\gamma)^2} \tag{141}$$

holds for every $3 \leq s \leq H$, and

$$t_{\overline{s}}\big(\gamma^{2s+1} - 1/4\big) - \max\Big\{t_s\big(\gamma^{2s} - 1/4\big),\, t_{\overline{s}}(\tau_{s+1})\Big\} \leq \frac{c_0}{\eta(1-\gamma)^2} \tag{142}$$

holds for every $1 \leq s \leq H$.

Lemma 18. *Suppose that (28) holds and*

$$t_3(\tau_3) > t_2(\gamma^4 - 1/4). \tag{143}$$

Then for every $3 \leq s \leq H$, we have

$$t_s\big(\gamma^{2s} - 1/4\big) - t_s\big(\tau_s\big) \leq \frac{2sc_0}{\eta(1-\gamma)^2}, \tag{144a}$$

$$t_{\overline{s-1}}\big(\gamma^{2s-1} - 1/4\big) - t_{\overline{s-1}}(\tau_s) \leq \frac{2sc_0}{\eta(1-\gamma)^2}. \tag{144b}$$

In addition, if we further have $t_{s-1}(\tau_{s-1}) > t_{\overline{s-2}}(\tau_{s-1}) + \frac{2sc_0}{\eta(1-\gamma)^2}$, then

$$t_{\overline{s-2}}\big(\gamma^{2s-3} - 1/4\big) \leq t_{\overline{s-1}}(\tau_s). \tag{144c}$$

Furthermore, (144c) still holds for $s = 3$ without requiring the assumption (143).

The proofs of the above two lemmas are postponed to Appendix E.2 and Appendix E.3, respectively. Let us take a moment to explain these two lemmas; to provide some intuitions, let us treat $\gamma^{2s} \approx 1$. Lemma 17 makes clear that: once the value function estimates for states $\overline{s-1}$ and s are both sufficiently large (i.e., $V^{(t)}(\overline{s-1}) \gtrapprox 0.75$ and $V^{(t)}(s) \gtrapprox 0.5$), then it does not take long for $V^{(t)}(s)$ to (approximately) exceed 0.75. A similar message holds true if we replace s (resp. $\overline{s-1}$) with $\overline{s}$ (resp. s). Built upon this observation, Lemma 18 further reveals that: the time taken for $V^{(t)}(s)$ (resp. $V^{(t)}(\overline{s-1})$) to rise from 0.5 to 0.75 is fairly short.

Proof of Lemma 6. We are now in a position to present the proof of Lemma 6. To begin with, recall from Lemma 8 that: for any $t \leq t_{\overline{s-1}}(\tau_s) \leq t_{\overline{s-1}}(\tau_{s-1})$, one has

$$Q^{(t)}(s, a_1) = \gamma V^{(t)}(\overline{s-1}) \leq \gamma \tau_s \leq \min\Big\{Q^{(t)}(s, a_0), Q^{(t)}(s, a_2)\Big\}. \tag{145}$$

Given that $V^{(t)}(s)$ is a convex combination of $\{Q^{(t)}(s,a)\}_{a\in\{a_0,a_1,a_2\}}$, one has $V^{(t)}(s) - Q^{(t)}(s, a_1) \geq 0$, which together with the gradient expression (124) indicates that

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(s, a_1)} \leq 0 \tag{146}$$

and hence $\theta^{(t)}(s, a_1)$ is non-increasing with t for any $t < t_{\overline{s-1}}(\tau_s)$. Additionally, we have learned from Lemma 18 that

$$t_{\overline{s-1}}(\tau_s) \geq t_{\overline{s-2}}\big(\gamma^{2s-3} - 1/4\big) \geq t_{\overline{s-2}}\big(\gamma\tau_{s-2}\big) = t_{s-2}\big(\tau_{s-2}\big),$$

where the second inequality holds since $\gamma^{2s-3} - 1/4 \geq \gamma\tau_{s-2}$, and the last identity results from Part (iii) of Lemma 8. This combined with the non-increasing nature of $\theta^{(t)}(s, a_1)$ readily establishes the advertised inequality $\theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1) \leq \theta^{(t_{s-2}(\tau_{s-2}))}(s, a_1)$.

The next step is to justify $\theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_2) \geq 0$. Notice that for $t > t_{s-2}(\tau_{s-2})$, we have $V^{(t)}(s-2) > \tau_{s-2}$, and then $V^{(t)}(\overline{s-2}) > \gamma\tau_{s-2}$ by (50), which leads to $Q^{(t)}(s, a_2) > Q^{(t)}(s, a_0)$ by (47) in Lemma 8. Recall (145) that $Q^{(t)}(s, a_1) \leq \gamma\tau_s \leq \min\Big\{Q^{(t)}(s, a_0), Q^{(t)}(s, a_2)\Big\}$. Then, one has $Q^{(t)}(s, a_2) - V^{(t)}(s) \geq 0$, which together with the gradient expression (124) indicates that

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(s, a_2)} \geq 0 \tag{147}$$

and hence $\theta^{(t)}(s, a_2)$ is non-decreasing with t for any $t < t_{\overline{s-1}}(\tau_s)$. This establishes $\theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_2) \geq 0$.

E.2 Proof of Lemma 17

For every $t \geq \max\big\{t_{\overline{s-1}}(\gamma^{2s-1} - 1/4), t_s(\tau_s)\big\}$, we isolate the following properties that will prove useful.

- The definition (24) of $t_{\overline{s-1}}(\cdot)$ together with the monotonicity property in Lemma 9 requires that $V^{(t)}(\overline{s-1}) \geq \gamma^{2s-1} - 1/4$, and hence it is seen from (47) that

$$Q^{(t)}(s, a_1) = \gamma V^{(t)}(\overline{s-1}) \geq \gamma^{2s} - \gamma/4. \tag{148}$$

- In the meantime, since $t \geq t_s(\tau_s)$, Lemma 8 (cf. (48)) guarantees that

$$\pi^{(t)}(a_1 \,|\, s) \geq (1-\gamma)/2. \tag{149}$$

- Given that $t_s(\tau_s) \geq t_{s-2}(\tau_{s-2})$ (see (27)) and the monotonicity property in Lemma 9, one has $V^{(t)}(\overline{s-2}) \geq \tau_{s-2}$, and thus we can see from (47) that

$$Q^{(t)}(s, a_2) - Q^{(t)}(s, a_0) = \gamma p\big(V^{(t)}(\overline{s-2}) - \tau_{s-2}\big) \geq 0. \tag{150}$$

- In addition, Lemma 8 ensures that both $Q^{(t)}(s,a_2)$ and $Q^{(t)}(s,a_0)$ are bounded above by $\gamma^{1/2}\tau_s$. Therefore, it is easily seen that

$$Q^{(t)}(s,a_1) \geq \gamma^{2s} - \gamma/4 > \gamma^{1/2}\tau_s \geq Q^{(t)}(s,a_2) \geq Q^{(t)}(s,a_0), \tag{151}$$

  where the first inequality comes from (148), the second one holds when $\gamma^{2s} > 0.75$, and the last inequality has been justified in (150).

- Moreover, given that $V^{(t)}(s) \geq \tau_s$ (since $t \geq t_s(\tau_s)$), one further has

$$\begin{aligned} Q^{(t)}(s,a_1) - \max\left\{Q^{(t)}(s,a_2), Q^{(t)}(s,a_0)\right\} &> V^{(t)}(s) - \max\left\{Q^{(t)}(s,a_2), Q^{(t)}(s,a_0)\right\} \\ &> \tau_s - \gamma^{1/2}\tau_s > 0. \end{aligned} \tag{152}$$

  Here, the first inequality comes from (151), while the penultimate inequality is a consequence of (151).

- We have seen from the above bullet points that

$$Q^{(t)}(s,a_1) > V^{(t)}(s) > \max\left\{Q^{(t)}(s,a_2), Q^{(t)}(s,a_0)\right\}, \tag{153}$$

  which combined with the gradient expression (124) reveals that

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a_1)} > 0 > \max\left\{\frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a_0)}, \frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a_2)}\right\}. \tag{154}$$

With the above properties in place, we are now ready to prove our lemma, for which we shall look at the key primary states $3 \leq s \leq H$ and the adjoint states separately.

Analysis for the key primary states. Let us start with any state $3 \leq s \leq H$, and control $t_s(\gamma^{2s} - 1/4)$ as claimed in (141). As before, define

$$\theta^{(t)}_{\mathsf{max}}(s) \coloneqq \max_a \theta^{(t)}(s,a).$$

From the above fact (154), we know that $\theta^{(t)}(s,a_1)$ keeps increasing with t while $\theta^{(t)}(s,a_0), \theta^{(t)}(s,a_2)$ are both decreasing with t. As a result, once $\theta^{(t)}(s,a_1) = \theta^{(t)}_{\mathsf{max}}(s)$, then $\theta^{(t)}(s,a_1)$ will remain equal to $\theta^{(t)}_{\mathsf{max}}(s)$ for the subsequent iterations. This allows us to divide into two stages as follows.

- **Stage 1: the duration when $\theta^{(t)}(s,a_1) < \theta^{(t)}_{\mathsf{max}}(s)$.** Our aim is to show that this stage contains at most $O\big(\frac{1}{\eta(1-\gamma)^2}\big)$ iterations. In order to prove this, the starting point is again the gradient expression (124):

$$\begin{aligned} \frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a_1)} &= \frac{1}{1-\gamma} d^{(t)}_{\mu}(s)\pi^{(t)}(a_1 \,|\, s)\big(Q^{(t)}(s,a_1) - V^{(t)}(s)\big) \\ &\geq c_{\mathrm{m}}\gamma(1-\gamma)\pi^{(t)}(a_1 \,|\, s)\big(Q^{(t)}(s,a_1) - V^{(t)}(s)\big), \end{aligned} \tag{155}$$

  where the last line relies on Lemma 2 and the fact $Q^{(t)}(s,a_1) > V^{(t)}(s)$ (cf. (153)). Regarding the size of $Q^{(t)}(s,a_1) - V^{(t)}(s)$, we make the observation that

$$\begin{aligned} Q^{(t)}(s,a_1) - V^{(t)}(s) &= \pi^{(t)}(a_0 \,|\, s)\Big(Q^{(t)}(s,a_1) - Q^{(t)}(s,a_0)\Big) + \pi^{(t)}(a_2 \,|\, s)\Big(Q^{(t)}(s,a_1) - Q^{(t)}(s,a_2)\Big) \\ &\geq \big(\pi^{(t)}(a_0 \,|\, s) + \pi^{(t)}(a_2 \,|\, s)\big)\Big(Q^{(t)}(s,a_1) - \max_{a\in\{a_0,a_2\}} Q^{(t)}(s,a)\Big) \\ &\overset{\text{(i)}}{\geq} \frac{1}{2}\Big(Q^{(t)}(s,a_1) - \max_{a\in\{a_0,a_2\}} Q^{(t)}(s,a)\Big) \overset{\text{(ii)}}{\geq} \frac{1}{2}\big(\gamma^{2s} - \gamma/4 - \gamma^{1/2}\tau_s\big) \overset{\text{(iii)}}{\geq} \frac{1}{16}. \end{aligned}$$

  Here, (i) follows since $\theta^{(t)}(s,a_1) < \theta^{(t)}_{\mathsf{max}}(s)$ during this stage and, therefore, $\pi^{(t)}(a_1 \,|\, s) \leq 1/2$; (ii) arises from the relation (151); and (iii) holds whenever $\gamma^{2s} - \gamma/4 > 5/8$. Substitution into (155) yields

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a_1)} \geq \frac{1}{16} c_{\mathrm{m}}\gamma(1-\gamma)\pi^{(t)}(a_1 \,|\, s) \geq \frac{1}{48} c_{\mathrm{m}}\gamma(1-\gamma)\widehat{\pi}^{(t)}(a_1 \,|\, s), \tag{156}$$

where the last inequality comes from (120). In addition, recall that $\theta^{(t)}(s,a_1)$ is increasing with t, while $\theta^{(t)}(s,a_0)$ and $\theta^{(t)}(s,a_2)$ are both decreasing (and hence $\theta^{(t)}_{\mathsf{max}}(s)$ is also decreasing). Invoking Lemma 16 then yields

$$\begin{aligned}
\widehat{\pi}^{(t+1)}(a_1\,|\,s)-\widehat{\pi}^{(t)}(a_1\,|\,s) &\geq \widehat{\pi}^{(t)}(a_1\,|\,s)\Big(\theta^{(t+1)}(s,a_1)-\theta^{(t)}(s,a_1)+\theta^{(t)}_{\mathsf{max}}(s)-\theta^{(t+1)}_{\mathsf{max}}(s)\Big)\\
&\geq \widehat{\pi}^{(t)}(a_1\,|\,s)\Big(\theta^{(t+1)}(s,a_1)-\theta^{(t)}(s,a_1)\Big)\\
&= \widehat{\pi}^{(t)}(a_1\,|\,s)\cdot\eta\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)} \geq \frac{1}{48}c_{\mathrm{m}}\eta\gamma(1-\gamma)\Big[\widehat{\pi}^{(t)}(a_1\,|\,s)\Big]^2,
\end{aligned}$$

where the last line arises from (156). Given this recursive relation, Lemma 11 implies that: if $\widehat{\pi}^{(t)}(a_1\,|\,s)<1$ (or equivalently, $\theta^{(t)}(s,a_1)<\theta^{(t)}_{\mathsf{max}}(s)$), then one necessarily has

$$t-t_{0,1}\leq\frac{1+\frac{1}{48}c_{\mathrm{m}}\eta\gamma(1-\gamma)}{\frac{1}{48}c_{\mathrm{m}}\eta\gamma(1-\gamma)\pi^{(t_0)}(a_1\,|\,s)}\leq\frac{2}{\frac{1}{48}c_{\mathrm{m}}\eta\gamma(1-\gamma)\pi^{(t_{0,1})}(a_1\,|\,s)}\leq\frac{240}{c_{\mathrm{m}}\eta\gamma(1-\gamma)^2},$$

with $t_{0,1}\coloneqq\max\big\{t_{\overline{s-1}}(\gamma^{2s-1}-1/4),t_s(\tau_s)\big\}$. Here, the last inequality relies on the property (149).

- **Stage 2: the duration when $\theta^{(t)}(s,a_1)=\theta^{(t)}_{\mathsf{max}}(s)$.** For this stage, we intend to demonstrate that it takes at most $O\big(\frac{1}{\eta(1-\gamma)^2}\big)$ iterations to achieve $\max\big\{\pi^{(t)}(a_0\,|\,s),\pi^{(t)}(a_2\,|\,s)\big\}\leq(1-\gamma)/8$. To this end, we again begin by studying the gradient as follows:

$$\begin{aligned}
\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_2)} &= \frac{1}{1-\gamma}d^{(t)}_{\mu}(s)\pi^{(t)}(a_2\,|\,s)\Big(Q^{(t)}(s,a_2)-V^{(t)}(s)\Big)\leq c_{\mathrm{m}}\gamma(1-\gamma)\pi^{(t)}(a_2\,|\,s)\Big(Q^{(t)}(s,a_2)-V^{(t)}(s)\Big)\\
&\leq\frac{1}{3}c_{\mathrm{m}}\gamma(1-\gamma)\widehat{\pi}^{(t)}(a_2\,|\,s)\Big(Q^{(t)}(s,a_2)-V^{(t)}(s)\Big).
\end{aligned}$$

Here, the first inequality comes from Lemma 2 and the fact $Q^{(t)}(s,a_2)<V^{(t)}(s)$ (see (153)), whereas the last inequality is a consequence of (120). In order to control $Q^{(t)}(s,a_2)-V^{(t)}(s)$, we observe that

$$\begin{aligned}
Q^{(t)}(s,a_2)-V^{(t)}(s) &= \pi^{(t)}(a_1\,|\,s)\Big(Q^{(t)}(s,a_2)-Q^{(t)}(s,a_1)\Big)+\pi^{(t)}(a_0\,|\,s)\Big(Q^{(t)}(s,a_2)-Q^{(t)}(s,a_0)\Big)\\
&\leq\pi^{(t)}(a_1\,|\,s)\Big(\gamma^{1/2}\tau_s-\gamma^{2s}+\gamma/4\Big)+\pi^{(t)}(a_2\,|\,s)\gamma p\left(V^{(t)}(\overline{s-2})-\tau_{s-2}\right)\\
&\leq\frac{1}{3}\Big(\gamma^{1/2}\tau_s-\gamma^{2s}+\gamma/4\Big)+\gamma p\leq-\frac{1}{24},
\end{aligned}$$

where the second line arises from (151) and (150), and the last line holds since $V^{(t)}(\overline{s-2})\leq 1$ as well as the fact $\pi^{(t)}(a_1\,|\,s)\geq 1/3$ during this stage (since $\theta^{(t)}(s,a_1)=\theta^{(t)}_{\mathsf{max}}(s)$). Putting the above two bounds together leads to

$$\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_2)}\leq-\frac{1}{72}c_{\mathrm{m}}\gamma(1-\gamma)\widehat{\pi}^{(t)}(a_2\,|\,s). \tag{157}$$

Next, Lemma 16 tells us that

$$\begin{aligned}
\widehat{\pi}^{(t+1)}(a_2\,|\,s)-\widehat{\pi}^{(t)}(a_2\,|\,s) &\leq 0.72\widehat{\pi}^{(t)}(a_2\,|\,s)\Big(\theta^{(t+1)}(s,a_2)-\theta^{(t)}(s,a_2)+\theta^{(t)}_{\mathsf{max}}(s)-\theta^{(t+1)}_{\mathsf{max}}(s)\Big)\\
&\leq 0.72\widehat{\pi}^{(t)}(a_2\,|\,s)\Big(\theta^{(t+1)}(s,a_2)-\theta^{(t)}(s,a_2)\Big)=0.72\widehat{\pi}^{(t)}(a_2\,|\,s)\cdot\eta\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_2)}\\
&\leq-0.01\eta c_{\mathrm{m}}\gamma(1-\gamma)\Big[\widehat{\pi}^{(t)}(a_2\,|\,s)\Big]^2,
\end{aligned}$$

where the first inequality makes use of the facts $\theta^{(t+1)}(s,a_2)\leq\theta^{(t)}(s,a_2)$ and $\theta^{(t)}_{\mathsf{max}}(s)=\theta^{(t)}(s,a_1)\leq\theta^{(t+1)}(s,a_1)=\theta^{(t+1)}_{\mathsf{max}}(s)$ (see (154)). Denoting by $t_{0,2}$ the first iteration in this stage, we can invoke Lemma 11 to reach

$$\widehat{\pi}^{(t-t_{0,2})}(a_2\,|\,s)\leq\frac{1}{0.01\eta c_{\mathrm{m}}\gamma(1-\gamma)(t-t_{0,2})+1}. \tag{158}$$

As a consequence, once $t - t_{0,2}$ exceeds

$$\frac{800}{\eta c_{\mathrm{m}} \gamma (1-\gamma)^2},$$

then one has $\pi^{(t)}(a_2 \,|\, s) \le (1-\gamma)/8$. The same conclusion holds for a_0 as well.

Combining the above analysis for the two stages, we see that: if

$$t - t_{0,1} \ge \frac{240}{\eta c_{\mathrm{m}} \gamma (1-\gamma)^2} + \frac{800}{\eta c_{\mathrm{m}} \gamma (1-\gamma)^2} = \frac{1040}{\eta c_{\mathrm{m}} \gamma (1-\gamma)^2}$$

with $t_{0,1} \coloneqq \max\big\{ t_{\overline{s-1}}(\gamma^{2s-1} - 1/4), t_s(\tau_s) \big\}$, then one has

$$\pi^{(t)}(a_1 \,|\, s) = 1 - \pi^{(t)}(a_0 \,|\, s) - \pi^{(t)}(a_2 \,|\, s) \ge 1 - (1-\gamma)/4,$$

which combined with (151) leads to

$$V^{(t)}(s) \ge \pi^{(t)}(a_1 \,|\, s) Q^{(t)}(s, a_1) \ge \big(1 - (1-\gamma)/4\big)\big(\gamma^{2s} - \gamma/4\big) \ge \gamma^{2s} - 1/4.$$

This means that one necessairly has $t \ge t_s(\gamma^{2s} - 1/4)$. It then follows that

$$t_s\big(\gamma^{2s} - 1/4\big) - \max\big\{ t_{\overline{s-1}}\big(\gamma^{2s-1} - 1/4\big), t_s(\tau_s) \big\} = t_s\big(\gamma^{2s} - 1/4\big) - t_{0,1} \le \frac{1040}{\eta c_{\mathrm{m}} \gamma (1-\gamma)^2},$$

thus concluding the proof of (141).

Analysis for the adjoint states. We then move forward to the adjoint states $\{\overline{1}, \cdots, \overline{H}\}$ and control $t_{\overline{s}}(\gamma^{2s+1} - 1/4)$ as desired in (142). The proof consists of studying the dynamic for any t obeying

$$\max\Big\{ t_s\big(\gamma^{2s} - 1/4\big),\, t_{\overline{s}}(\tau_{s+1}) \Big\} \le t \le t_{\overline{s}}\big(\gamma^{2s+1} - 1/4\big).$$

Once again, we divide into two stages and analyze each of them separately.

- **Stage 1: the duration where $\theta^{(t)}(\overline{s}, a_1) < \theta^{(t)}(\overline{s}, a_0)$.** We aim to demonstrate that it takes no more than $O\big(\frac{1}{\eta(1-\gamma)^2}\big)$ iterations for $\theta^{(t)}(\overline{s}, a_1)$ to surpass $\theta^{(t)}(\overline{s}, a_0)$. In order to do so, note that

$$\begin{aligned} \frac{\partial V^{(t)}(\mu)}{\partial \theta(\overline{s}, a_1)} &= \frac{1}{1-\gamma} d_{\mu}^{(t)}(\overline{s}) \pi^{(t)}(a_1 \,|\, \overline{s}) \pi^{(t)}(a_0 \,|\, \overline{s}) \Big(Q^{(t)}(\overline{s}, a_1) - Q^{(t)}(\overline{s}, a_0) \Big) \\ &\ge \frac{1}{16} c_{\mathrm{m}} \gamma (1-\gamma) \pi^{(t)}(a_1 \,|\, \overline{s}) > 0. \end{aligned} \tag{159}$$

  Here, the last line applies Lemma 2 and makes use of the fact

$$Q^{(t)}(\overline{s}, a_1) - Q^{(t)}(\overline{s}, a_0) = \gamma V^{(t)}(s) - \gamma \tau_s \ge \gamma(\gamma^{2s} - 1/4 - \tau_s) \ge 1/8. \tag{160}$$

  where the inequality comes from the assumption $t \ge t_s\big(\gamma^{2s} - 1/4\big)$ as well as the monotonicity property in Lemma 9. As a result, the PG update rule (9a) implies that $\theta^{(t)}(\overline{s}, a_1)$ is increasing in t, and hence $\theta^{(t)}(\overline{s}, a_0)$ is decreasing in t (since $\sum_a \theta^{(t)}(s, a) = 0$); these taken collectively mean that

$$\theta^{(t+1)}(\overline{s}, a_1) - \theta^{(t)}(\overline{s}, a_1) + \theta^{(t)}(\overline{s}, a_0) - \theta^{(t+1)}(\overline{s}, a_0) \ge \theta^{(t+1)}(\overline{s}, a_1) - \theta^{(t)}(\overline{s}, a_1) \ge 0.$$

  Invoking Lemma 16 then reveals that

$$\begin{aligned} \widehat{\pi}^{(t+1)}(a_1 \,|\, \overline{s}) - \widehat{\pi}^{(t)}(a_1 \,|\, \overline{s}) &\ge \widehat{\pi}^{(t)}(a_1 \,|\, \overline{s}) \Big(\theta^{(t+1)}(\overline{s}, a_1) - \theta^{(t)}(\overline{s}, a_1) + \theta^{(t)}(\overline{s}, a_0) - \theta^{(t+1)}(\overline{s}, a_0) \Big) \\ &\ge \widehat{\pi}^{(t)}(a_1 \,|\, \overline{s}) \Big(\theta^{(t+1)}(\overline{s}, a_1) - \theta^{(t)}(\overline{s}, a_1) \Big) = \eta \widehat{\pi}^{(t)}(a_1 \,|\, \overline{s}) \frac{\partial V^{(t)}(\mu)}{\partial \theta(\overline{s}, a_1)} \end{aligned}$$

$$\geq \frac{1}{48}\eta c_{\mathrm{m}}\gamma(1-\gamma)\Big[\widehat{\pi}^{(t)}(a_1\,|\,\overline{s})\Big]^2,$$

where the last inequality relies on (159) and (120). Given this recursive relation, Lemma 11 tells us that: one has $\widehat{\pi}^{(t)}(a_1\,|\,\overline{s}) \geq 1$ (which means a_1 becomes the favored action by (119)) as soon as $t - t_{0,3}$ exceeds

$$\frac{2}{\frac{1}{48}\eta c_{\mathrm{m}}\gamma(1-\gamma)\widehat{\pi}^{(t_{0,3})}(a_1\,|\,\overline{s})} \leq \frac{96}{\eta c_{\mathrm{m}}\gamma(1-\gamma)\pi^{(t_{0,3})}(a_1\,|\,\overline{s})} \leq \frac{1152}{\eta c_{\mathrm{m}}\gamma(1-\gamma)^2},$$

where $t_{0,3} \coloneqq \max\Big\{t_s\big(\gamma^{2s}-1/4\big),\, t_{\overline{s}}(\tau_{s+1})\Big\}$. Here, the last inequality is valid as long as

$$\pi^{(t_{0,3})}(a_1\,|\,\overline{s}) \geq (1-\gamma)/12 \tag{161}$$

holds. It thus remains to justify (161). Towards this end, observe that for any $t \geq t_{\overline{s}}(\tau_{s+1})$,

$$\begin{aligned}\tau_{s+1} \leq V^{(t)}(\overline{s}) &= \pi^{(t)}(a_0\,|\,\overline{s})Q^{(t)}(\overline{s},a_0) + \pi^{(t)}(a_1\,|\,\overline{s})Q^{(t)}(\overline{s},a_1) = \pi^{(t)}(a_0\,|\,\overline{s})\gamma\tau_s + \pi^{(t)}(a_1\,|\,\overline{s})\gamma V^{(t)}(s)\\ &= \gamma\tau_s + \pi^{(t)}(a_1\,|\,\overline{s})\gamma\Big(V^{(t)}(s)-\tau_s\Big) \leq \gamma\tau_s + \pi^{(t)}(a_1\,|\,\overline{s})\gamma,\end{aligned}$$

and, as a result,

$$\pi^{(t)}(a_1\,|\,\overline{s}) \geq \frac{\tau_{s+1}}{\gamma} - \tau_s = \frac{1}{2}\frac{\gamma^{\frac{2}{3}s+\frac{2}{3}} - \gamma^{\frac{2}{3}s+1}}{\gamma} = \frac{\gamma^{\frac{2}{3}s-1}}{2}\Big(\gamma^{\frac{2}{3}} - \gamma\Big) \geq \frac{1-\gamma}{12},$$

provided that $\gamma \geq 0.9$ (so that $\gamma^{\frac{2}{3}} - \gamma \geq 0.3(1-\gamma)$) and $\gamma^{\frac{2}{3}H} \geq 0.7$. This concludes the analysis of this stage.

- **Stage 2: the duration where $\theta^{(t)}(\overline{s},a_1) \geq \theta^{(t)}(\overline{s},a_0)$.** Similar to the above argument, we intend to show that it takes at most $O\big(\frac{1}{\eta(1-\gamma)^2}\big)$ iterations for $\pi^{(t)}(a_0\,|\,\overline{s}) \leq 1-\gamma$ to occur. From the gradient expression and the property (160), we obtain

$$\begin{aligned}\frac{\partial V^{(t)}(\mu)}{\partial\theta(\overline{s},a_0)} &= \frac{1}{1-\gamma}d_\mu^{(t)}(\overline{s})\pi^{(t)}(a_0\,|\,\overline{s})\pi^{(t)}(a_1\,|\,\overline{s})\Big(Q^{(t)}(\overline{s},a_0) - Q^{(t)}(\overline{s},a_1)\Big)\\ &\leq -\frac{1}{16}c_{\mathrm{m}}\gamma(1-\gamma)\pi^{(t)}(a_0\,|\,\overline{s}) \leq -\frac{1}{48}c_{\mathrm{m}}\gamma(1-\gamma)\widehat{\pi}^{(t)}(a_0\,|\,\overline{s}),\end{aligned}$$

where the first inequality uses Lemma 2 and the property $\pi^{(t)}(a_1\,|\,\overline{s}) \geq 1/2$ (since $\theta^{(t)}(\overline{s},a_1) \geq \theta^{(t)}(\overline{s},a_0)$), and the last inequality relies on (120). Repeating a similar argument as above, we can demonstrate that

$$\widehat{\pi}^{(t+1)}(a_0\,|\,\overline{s}) - \widehat{\pi}^{(t)}(a_0\,|\,\overline{s}) \leq -\frac{1}{70}\eta c_{\mathrm{m}}\gamma(1-\gamma)\Big[\widehat{\pi}^{(t)}(a_0\,|\,\overline{s})\Big]^2. \tag{162}$$

This combined with Lemma 11 implies that

$$\widehat{\pi}^{(t)}(a_0\,|\,\overline{s}) \leq \frac{1}{\frac{1}{70}\eta c_{\mathrm{m}}\gamma(1-\gamma)(t-t_{0,4})+1}, \tag{163}$$

with $t_{0,4}$ denoting the first iteration of this stage. Consequently, one has $\widehat{\pi}^{(t)}(a_0\,|\,\overline{s}) \leq 1-\gamma$ — and therefore $\pi^{(t)}(a_0\,|\,\overline{s}) \leq 1-\gamma$ according to (120) — as soon as $t - t_{0,4}$ exceeds

$$\frac{70}{\eta c_{\mathrm{m}}\gamma(1-\gamma)^2}.$$

Finally, if $\pi^{(t)}(a_0\,|\,\overline{s}) \leq 1-\gamma$, then one has

$$V^{(t)}(\overline{s}) = \pi^{(t)}(a_0\,|\,\overline{s})Q^{(t)}(\overline{s},a_0) + \pi^{(t)}(a_1\,|\,\overline{s})Q^{(t)}(\overline{s},a_1) = \pi^{(t)}(a_0\,|\,\overline{s})\gamma\tau_s + \pi^{(t)}(a_1\,|\,\overline{s})\gamma V^{(t)}(s)$$

$$
\begin{aligned}
&\geq \pi^{(t)}(a_0 \,|\, \overline{s})\gamma\tau_s + \Big(1-\pi^{(t)}(a_0 \,|\, \overline{s})\Big)\gamma\left(\gamma^{2s}-1/4\right) = \pi^{(t)}(a_0 \,|\, \overline{s})\left(\gamma\tau_s - \gamma^{2s+1} - \gamma/4\right) + \gamma^{2s+1} - \gamma/4 \\
&\geq (1-\gamma)\left(\gamma\tau_s - \gamma^{2s+1} - \gamma/4\right) + \gamma^{2s+1} - \gamma/4 \geq \gamma^{2s+1} - 1/4,
\end{aligned}
$$

where the first inequality holds by recalling that $t \geq t_s(\gamma^{2s}-1/4)$. Consequently, putting the above pieces (regarding the duration of the two stages) together allows us to conclude that

$$
t_{\overline{s}}\big(\gamma^{2s+1}-1/4\big) - \max\Big\{t_s\big(\gamma^{2s}-1/4\big),\, t_{\overline{s}}(\tau_{s+1})\Big\} \leq \frac{1152}{\eta c_{\mathrm{m}}\gamma(1-\gamma)^2} + \frac{70}{\eta c_{\mathrm{m}}\gamma(1-\gamma)^2} = \frac{1222}{\eta c_{\mathrm{m}}\gamma(1-\gamma)^2}
$$

as claimed.

E.3 Proof of Lemma 18

Before proceeding, we first single out two properties that play a crucial role in the proof of Lemma 18.

Lemma 19. *The following basic properties hold true for any $2 \leq s \leq H$:*

$$
t_s(\tau_s) \geq t_{\overline{s-1}}(\tau_s); \tag{164a}
$$
$$
t_{\overline{s-1}}(\tau_s) \geq t_{s-1}(\tau_{s-1}). \tag{164b}
$$

The proof of this auxiliary lemma is deferred to the end of this subsection. Equipped with this result, we are positioned to present the proof of Lemma 18. To begin with, we seek to bound the quantity $t_s(\gamma^{2s} - 1/4) - t_s(\tau_s)$. Apply Lemma 17 with a little algebra to yield

$$
\begin{aligned}
t_s(\gamma^{2s}-1/4) - t_s(\tau_s) &\leq \max\Big\{t_{\overline{s-1}}(\gamma^{2s-1}-1/4), t_s(\tau_s)\Big\} + \frac{c_0}{\eta(1-\gamma)^2} - t_s(\tau_s) \\
&= \max\Big\{t_{\overline{s-1}}(\gamma^{2s-1}-1/4) - t_s(\tau_s), 0\Big\} + \frac{c_0}{\eta(1-\gamma)^2}.
\end{aligned} \tag{165}
$$

With the assistance of the bound (164a) in Lemma 19, we can continue the bound in (165) to derive

$$
\begin{aligned}
t_s(\gamma^{2s}-1/4) - t_s(\tau_s) &\leq \max\Big\{t_{\overline{s-1}}(\gamma^{2s-1}-1/4) - t_{\overline{s-1}}(\tau_s), 0\Big\} + \frac{c_0}{\eta(1-\gamma)^2} \\
&= t_{\overline{s-1}}(\gamma^{2s-1}-1/4) - t_{\overline{s-1}}(\tau_s) + \frac{c_0}{\eta(1-\gamma)^2}.
\end{aligned} \tag{166}
$$

To continue, we shall bound the quantity $t_{\overline{s-1}}(\gamma^{2s-1}-1/4) - t_{\overline{s-1}}(\tau_s)$. Similar to the derivation of the inequality (165), we can apply Lemma 17 to show that

$$
\begin{aligned}
t_{\overline{s-1}}(\gamma^{2s-1}-1/4) - t_{\overline{s-1}}(\tau_s) &\leq \max\Big\{t_{s-1}(\gamma^{2s-2}-1/4), t_{\overline{s-1}}(\tau_s)\Big\} + \frac{c_0}{\eta(1-\gamma)^2} - t_{\overline{s-1}}(\tau_s) \\
&= \max\Big\{t_{s-1}(\gamma^{2s-2}-1/4) - t_{\overline{s-1}}(\tau_s), 0\Big\} + \frac{c_0}{\eta(1-\gamma)^2} \\
&\leq \max\Big\{t_{s-1}(\gamma^{2s-2}-1/4) - t_{s-1}(\tau_{s-1}), 0\Big\} + \frac{c_0}{\eta(1-\gamma)^2} \\
&= t_{s-1}(\gamma^{2s-2}-1/4) - t_{s-1}(\tau_{s-1}) + \frac{c_0}{\eta(1-\gamma)^2},
\end{aligned} \tag{167}
$$

where the third line makes use of (164b) in Lemma 19.

Applying the inequalities (166) and (167) recursively, one arrives at

$$
\begin{aligned}
t_s(\gamma^{2s}-1/4) - t_s(\tau_s) &\leq t_{s-1}(\gamma^{2s-2}-1/4) - t_{s-1}(\tau_{s-1}) + \frac{2c_0}{\eta(1-\gamma)^2} \leq \cdots \\
&\leq t_3(\gamma^6-1/4) - t_3(\tau_3) + \frac{2(s-3)c_0}{\eta(1-\gamma)^2}.
\end{aligned} \tag{168}
$$

To continue, note that Lemma 17 and the bound (164a) give

$$t_3(\gamma^6 - 1/4) \leq \max\left\{t_{\overline{2}}(\gamma^5 - 1/4), t_3(\tau_3)\right\} + \frac{c_0}{\eta(1-\gamma)^2},$$
$$t_{\overline{2}}(\gamma^5 - 1/4) \leq \max\left\{t_2(\gamma^4 - 1/4), t_{\overline{2}}(\tau_3)\right\} + \frac{c_0}{\eta(1-\gamma)^2} \leq \max\left\{t_2(\gamma^4 - 1/4), t_3(\tau_3)\right\} + \frac{c_0}{\eta(1-\gamma)^2},$$

which together leads to

$$t_3(\gamma^6 - 1/4) \leq \max\left\{t_2(\gamma^4 - 1/4), t_3(\tau_3)\right\} + \frac{2c_0}{\eta(1-\gamma)^2}. \tag{169}$$

Plugging back to (168) leads to

$$\begin{aligned} t_s(\gamma^{2s} - 1/4) - t_s(\tau_s) &\leq \max\left\{t_2(\gamma^4 - 1/4), t_3(\tau_3)\right\} - t_3(\tau_3) + \frac{2c_0}{\eta(1-\gamma)^2} + \frac{2(s-3)c_0}{\eta(1-\gamma)^2} \\ &\leq \frac{(2s-4)c_0}{\eta(1-\gamma)^2}, \end{aligned} \tag{170}$$

where the last step arises from the assumption (143), that is, $t_2(\gamma^4 - 1/4) < t_3(\tau_3)$.

Further, the above inequality taken together with (167) yields

$$t_{\overline{s-1}}(\gamma^{2s-1} - 1/4) - t_{\overline{s-1}}(\tau_s) \leq \frac{(2s-4)c_0}{\eta(1-\gamma)^2} + \frac{c_0}{\eta(1-\gamma)^2} = \frac{(2s-3)c_0}{\eta(1-\gamma)^2}. \tag{171}$$

We have thus established (144a) and (144b).

Finally, we turn to the proof of (144c). In view of (144b), one has

$$t_{\overline{s-2}}(\gamma^{2s-3} - 1/4) - t_{\overline{s-2}}(\tau_{s-1}) \leq \frac{2sc_0}{\eta(1-\gamma)^2}.$$

In addition,

$$\begin{aligned} t_{\overline{s-1}}(\tau_s) - t_{\overline{s-2}}(\tau_{s-1}) &\geq t_{\overline{s-1}}(\gamma\tau_{s-1}) - t_{\overline{s-2}}(\tau_{s-1}) = t_{s-1}(\tau_{s-1}) - t_{\overline{s-2}}(\tau_{s-1}) \\ &\geq t_{s-1}(\tau_{s-1}) - t_{\overline{s-2}}(\tau_{s-1}) > \frac{2sc_0}{\eta(1-\gamma)^2}, \end{aligned}$$

where the identity in the first line comes from Part (iii) of Lemma 8, and the last inequality uses the assumption $t_{s-1}(\tau_{s-1}) > t_{\overline{s-2}}(\tau_{s-1}) + \frac{2sc_0}{\eta(1-\gamma)^2}$. Combining the above two inequalities justifies the validity of the advertised inequality (144c). Then, we establish (144c) for $s = 3$ through Lemma 4, which gives

$$t_{\overline{1}}\big(\gamma^3 - 1/4\big) \leq t_2(\tau_2) \leq t_{\overline{2}}(\tau_3), \tag{172}$$

where the last inequality comes from (164b).

Proof of Lemma 19. To begin with, the claim (164a) holds when $s = 2$ as a result of the inequality (31b) in Lemma 4. We now turn to the case with $3 \leq s \leq H$. In view of the property (47) in Lemma 8, we have

$$\max\left\{Q^{(t)}(s, a_0), Q^{(t)}(s, a_2)\right\} \leq \gamma^{\frac{1}{2}}\tau_s < \tau_s \qquad \text{and} \qquad Q^{(t)}(s, a_1) = \gamma V^{(t)}(\overline{s-1}).$$

Recognizing that $V^{(t)}(s)$ is a convex combination of $\left\{Q^{(t)}(s, a)\right\}_{a \in \{a_0, a_1, a_2\}}$, we know that if $V^{(t)}(s) \geq \tau_s$, then one necessarily has $Q^{(t)}(s, a_1) > \tau_s$, or equivalently, $V^{(t)}(\overline{s-1}) > \tau_s/\gamma \geq \tau_s$. This essentially means that $t_s(\tau_s) \geq t_{\overline{s-1}}(\tau_s)$, thus establishing the claim (164a).

Similarly, Lemma 8 (cf. (49)) also tells us that

$$Q^{(t)}(\overline{s}, a_0) = \gamma\tau_s \qquad \text{and} \qquad Q^{(t)}(\overline{s}, a_1) = \gamma V^{(t)}(s).$$

This means that if $V^{(t)}(s-1) \leq \tau_{s-1}$, then

$$V^{(t)}(\overline{s-1}) \leq \max\left\{Q^{(t)}(\overline{s-1}, a_0), Q^{(t)}(\overline{s-1}, a_1)\right\} \leq \gamma\tau_{s-1} \leq \tau_s.$$

Consequently, we conclude that $t_{\overline{s-1}}(\tau_s) \geq t_{s-1}(\tau_{s-1})$, as claimed in (164b). $\square$

F Analysis for the blowing-up lemma (Lemma 7)

In this section, we establish the blowing-up phenomenon as asserted in Lemma 7.

F.1 Which reference point t_{ref} shall we choose?

Let us specify the time instance t_{ref} as required in Lemma 7 as follows

$$t_{\mathsf{ref}} \coloneqq \min \Big\{ t \in \big[t_{\overline{s-1}}(\tau_s), t_s(\tau_s) \big) \;\mid\; c_{\mathsf{ref}}(1-\gamma)\pi^{(t)}(a_0 \,|\, s) \le \pi^{(t)}(a_1 \,|\, s) \Big\}, \tag{173}$$

where $c_{\mathsf{ref}} \in (0, 1/3)$ is some constant to be specified shortly.

Existence. An important step is to justify that (173) is well-defined, namely, there exists at least one time instance within $\big[t_{\overline{s-1}}(\tau_s), t_s(\tau_s) \big)$ that satisfies $c_{\mathsf{ref}}(1-\gamma)\pi^{(t)}(a_0 \,|\, s) \le \pi^{(t)}(a_1 \,|\, s)$. Towards this, we note that if the time instance $t_{\overline{s-1}}(\tau_s)$ obeys

$$c_{\mathsf{ref}}(1-\gamma)\pi^{(t)}(a_0 \,|\, s) \le \pi^{(t)}(a_1 \,|\, s) \qquad \text{when } t = t_{\overline{s-1}}(\tau_s),$$

then we simply have $t_{\mathsf{ref}} = t_{\overline{s-1}}(\tau_s)$. We then move on to the complement case where

$$\begin{aligned} & c_{\mathsf{ref}}(1-\gamma)\pi^{(t_{\overline{s-1}}(\tau_s))}(a_0 \,|\, s) > \pi^{(t_{\overline{s-1}}(\tau_s))}(a_1 \,|\, s), \\ \text{or equivalently,} \qquad & \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_0) > \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1) - \log(c_{\mathsf{ref}}(1-\gamma)). \end{aligned} \tag{174}$$

To justify that the construction (173) makes sense, it suffices to show that the endpoint $t_s(\tau_s)$ obeys

$$c_{\mathsf{ref}}(1-\gamma)\pi^{(t_s(\tau_s))}(a_0 \,|\, s) < \pi^{(t_s(\tau_s))}(a_1 \,|\, s). \tag{175}$$

In order to validate (175), recall that the inequality (48) in Lemma 8 ensures that

$$\pi^{(t_s(\tau_s))}(a_1 \,|\, s) \ge \frac{1-\gamma}{2},$$

given that $V^{(t_s(\tau_s))}(s) \ge \tau_s$. Therefore, the inequality (175) must be satisfied when $c_{\mathsf{ref}} < 1/2$, given that the left-hand side of (175) obeys

$$c_{\mathsf{ref}}(1-\gamma)\pi^{(t_s(\tau_s))}(a_0 \,|\, s) \le c_{\mathsf{ref}}(1-\gamma) < \frac{1-\gamma}{2}.$$

This in turn validates the existence of (175) for this case.

Several immediate properties about t_{ref} and $t_{\overline{s-1}}(\tau_s)$. We pause to single out a couple of immediate properties about the t_{ref} constructed above as well as $t_{\overline{s-1}}(\tau_s)$.

Consider the case where $t_{\overline{s-1}}(\tau_s)$ obeys

$$\begin{aligned} & c_{\mathsf{ref}}(1-\gamma)\pi^{(t_{\overline{s-1}}(\tau_s))}(a_0 \,|\, s) \le \pi^{(t_{\overline{s-1}}(\tau_s))}(a_1 \,|\, s), \\ \text{or equivalently,} \qquad & \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_0) \le \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1) - \log\big(c_{\mathsf{ref}}(1-\gamma)\big), \end{aligned}$$

then one has $t_{\mathsf{ref}} = t_{\overline{s-1}}(\tau_s)$ (as discussed above). As can be clearly seen, $t_{\overline{s-1}}(\tau_s)$ satisfies the advertised inequality (38a) by taking $c_{\mathsf{ref}} \ge c_{\mathrm{p}}/8064$. Additionally, let us first recall from (144c) in Lemma 18 that

$$t_{\overline{s-1}}(\tau_s) = \max\Big\{ t_{\overline{s-2}}(\gamma^{2s-3} - 1/4),\; t_{s-1}(\tau_s) \Big\}.$$

This combined with Lemma 6 (see (36)) tells us that

$$\theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1) = \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1) \le \theta^{(t_{s-2}(\tau_{s-2}))}(s, a_1) \le -\frac{1}{2}\log\Big(1 + \frac{c_{\mathrm{m}}\gamma}{35}\eta(1-\gamma)^2 t_{s-2}(\tau_{s-2})\Big), \tag{176}$$

where the last relation utilizes the bound (33) in Lemma 5. This leads to the advertised bound (38b).

As a result, the claims (38a)-(38b) only need to be justified under the assumption (174).

Organization of the proof. In light of the above basic facts, the subsequent proof focuses on the scenario where (174) is satisfied, namely, the case where

$$t_{\overline{s-1}}(\tau_s) < t_{\mathsf{ref}}.$$

We shall start by justifying that $\theta^{(t)}(s, a_1)$ has not increased much during $[t_{\overline{s-1}}(\tau_s), t_{\mathsf{ref}}]$, as detailed in Appendix F.2 and F.3 (focusing on two separate stages respectively). This feature will then be used in Appendix F.4 to establish the claims (38a)-(38b), and in Appendix F.5 to establish the claim (38c).

F.2 Stage I: the duration where $\theta^{(t)}(s, a_2) < \theta^{(t)}(s, a_0)$

Suppose that at the starting point we have $\theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_2) < \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_0)$; otherwise the reader can proceed directly to Stage II in Appendix F.3. The goal is to control the number of iterations taken to achieve $\theta^{(t)}(s, a_2) \geq \theta^{(t)}(s, a_0)$. More specifically, let us define the transition point

$$t_{\mathsf{tran}} := \min \Big\{ t \,|\, \theta^{(t)}(s, a_2) \geq \theta^{(t)}(s, a_0),\, t \geq t_{\overline{s-1}}(\tau_s) \Big\}. \tag{177}$$

In this subsection, we seek to develop an upper bound on $t_{\mathsf{tran}} - t_{\overline{s-1}}(\tau_s)$, and to show that $\theta^{(t)}(s, a_1) - \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1) \leq 1/2$ holds throughout this stage.

Preparation: basic facts and rescaled policies. Before moving forward, we first gather some basic facts. To begin with, from the definition (173) of t_{ref}, we know that the inequality $c_{\mathsf{ref}}(1-\gamma)\pi^{(t)}(a_0 \,|\, s) > \pi^{(t)}(a_1 \,|\, s)$ holds true for every $t \in [t_{\overline{s-1}}(\tau_s),\ t_{\mathsf{ref}})$, or equivalently,

$$\theta^{(t)}(s, a_0) > \theta^{(t)}(s, a_1) - \log\big(c_{\mathsf{ref}}(1-\gamma)\big) \qquad \text{for all } t \in [t_{\overline{s-1}}(\tau_s),\ t_{\mathsf{ref}}). \tag{178}$$

In the case considered here, we have — according to (178) and (177) — that

$$\theta^{(t)}(s, a_0) > \theta^{(t)}(s, a_1) - \log\big(c_{\mathsf{ref}}(1-\gamma)\big) \qquad \text{and} \qquad \theta^{(t)}(s, a_0) > \theta^{(t)}(s, a_2) \tag{179}$$

for any t obeying $t_{\overline{s-1}}(\tau_s) \leq t < \min\{t_{\mathsf{tran}}, t_{\mathsf{ref}}\}$. This means that

$$\theta^{(t)}(s, a_0) = \max_a \theta^{(t)}(s, a) \qquad \text{and hence} \qquad \pi^{(t)}(a_0 \,|\, s) > 1/3 \tag{180}$$

holds for any t obeying $t_{\overline{s-1}}(\tau_s) \leq t < \min\{t_{\mathsf{tran}}, t_{\mathsf{ref}}\}$, provided that $0 < c_{\mathsf{ref}} < 1$.

Moreover, let us introduce the rescaled policy $\widehat{\pi}^{(t)}(a \,|\, s)$ as before

$$\widehat{\pi}^{(t)}(a \,|\, s) := \exp\Big(\theta^{(t)}(s, a) - \max_{a' \in \mathcal{A}_s} \theta^{(t)}(s, a')\Big).$$

In view of (180), the rescaled policy can therefore be written as

$$\begin{aligned} \widehat{\pi}^{(t)}(a_2 \,|\, s) &= \exp\big(\theta^{(t)}(s, a_2) - \theta^{(t)}(s, a_0)\big) = \exp\big(2\theta^{(t)}(s, a_2) + \theta^{(t)}(s, a_1)\big) \\ \widehat{\pi}^{(t)}(a_1 \,|\, s) &= \exp\big(\theta^{(t)}(s, a_1) - \theta^{(t)}(s, a_0)\big) = \exp\big(2\theta^{(t)}(s, a_1) + \theta^{(t)}(s, a_2)\big) \end{aligned} \tag{181}$$

for any t with $t_{\overline{s-1}}(\tau_s) \leq t < \min\{t_{\mathsf{tran}}, t_{\mathsf{ref}}\}$, where we have used the constraint $\sum_a \theta^{(t)}(s, a) = 0$ (see Part (vii) of Lemma 8).

Showing $\theta^{(t)}(s, a_1) - \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1) \leq 1/2$ by induction. In the following, we seek to prove by induction the following key property

$$\theta^{(t)}(s, a_1) - \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1) \leq 1/2 \tag{182}$$

for any t that obeys $t_{\overline{s-1}}(\tau_s) \leq t \leq \min\{t_{\mathsf{tran}}, t_{\mathsf{ref}}\}$ and

$$t - t_{\overline{s-1}}(\tau_s) \leq \frac{225}{c_{\mathrm{p}} c_{\mathrm{m}} \eta (1-\gamma)^2 \exp\big(\theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1)\big)} =: \widetilde{t}. \tag{183}$$

We shall return to justify (183) for all t within this stage later on. In words, the claim (182) essentially means that $\theta^{(t)}(s,a_1)$ does not deviate much from $\theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_1)$ during this stage. With regards to the base case where $t = t_{\overline{s-1}}(\tau_s)$, the hypothesis (182) holds true trivially. Next, assuming that (182) is satisfied for every integer less than or equal to $t-1$, we intend to establish this hypothesis for the t-th iteration, which is accomplished as follows.

First, Lemma 1 and Lemma 10 tell us that $Q^{(t)}(s,a_1) - V^{(t)}(s) \le 1$. It then follows that

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a_1)} = \frac{1}{1-\gamma} d_\mu^{(t)}(s)\pi^{(t)}(a_1 \,|\, s)\Big\{Q^{(t)}(s,a_1) - V^{(t)}(s)\Big\} \le 14c_{\mathrm{m}}\eta(1-\gamma)\pi^{(t)}(a_1 \,|\, s),$$

which relies on the bound $d_\mu^{(t)}(s) \le 14c_{\mathrm{m}}(1-\gamma)^2$ stated in Lemma 3. As a result, it can be derived from the PG update rule (9a) that

$$\begin{aligned}\theta^{(t)}(s,a_1) - \theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_1) &= \sum_{j=t_{\overline{s-1}}(\tau_s)}^{t-1} \eta \frac{\partial V^{(j)}(\mu)}{\partial \theta(s,a_1)} \le \sum_{j=t_{\overline{s-1}}(\tau_s)}^{t-1} 14c_{\mathrm{m}}\eta(1-\gamma)\pi^{(j)}(a_1 \,|\, s) \\ &\le 14c_{\mathrm{m}}\eta(1-\gamma)(t - t_{\overline{s-1}}(\tau_s)) \max_{t_{\overline{s-1}}(\tau_s)\le j<t} \pi^{(j)}(a_1 \,|\, s). \end{aligned} \tag{184}$$

Regarding the term involving $\pi^{(j)}(a_1 \,|\, s)$, we observe that for any $t_{\overline{s-1}}(\tau_s) \le j < t$,

$$\pi^{(j)}(a_1 \,|\, s) \overset{\text{(i)}}{\le} \widehat{\pi}^{(j)}(a_1 \,|\, s) \overset{\text{(ii)}}{\le} \exp\Big(\frac{3}{2}\theta^{(j)}(s,a_1)\Big) \tag{185}$$

$$\overset{\text{(iii)}}{\le} \exp\Big(\frac{3}{2}\big(\theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_1) + 1/2\big)\Big). \tag{186}$$

Here, (i) is a consequence of (120), (ii) holds since (in view of (181), $\theta^{(j)}(s,a_0) \ge 0$, and $\sum_a \theta^{(j)}(s,a) = 0$)

$$\begin{aligned}\widehat{\pi}^{(j)}(a_1 \,|\, s) &= \exp\Big(2\theta^{(j)}(s,a_1) + \theta^{(j)}(s,a_2)\Big) \le \exp\Big(2\theta^{(j)}(s,a_1) + 0.5\theta^{(j)}(s,a_2) + 0.5\theta^{(j)}(s,a_0)\Big) \\ &= \exp\Big(1.5\theta^{(j)}(s,a_1)\Big),\end{aligned}$$

whereas (iii) follows from the induction hypothesis (182) for any $t_{\overline{s-1}}(\tau_s) \le j < t$. Combine the inequalities (184) and (186) to reach

$$\theta^{(t)}(s,a_1) - \theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_1) \le 14c_{\mathrm{m}}\eta(1-\gamma)\big(t - t_{\overline{s-1}}(\tau_s)\big)\exp\Big(\frac{3}{2}\big(\theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_1) + 1/2\big)\Big).$$

Consequently, under the constraint (183), the preceding inequality implies that

$$\begin{aligned}\theta^{(t)}(s,a_1) - \theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_1) \le\ & 14c_{\mathrm{m}}\eta(1-\gamma)\frac{225}{c_{\mathrm{p}}c_{\mathrm{m}}\eta(1-\gamma)^2\exp(\theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_1))}\exp\Big(\frac{3}{2}\big(\theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_1) + 1/2\big)\Big) \\ =\ & \frac{3150e\exp\big(\frac{1}{2}\theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_1)\big)}{c_{\mathrm{p}}(1-\gamma)} \le \frac{1}{2},\end{aligned} \tag{187}$$

where the last inequality makes use of (176) and the assumption (37). These allow us to establish the induction hypothesis for the t-th iteration, namely,

$$\theta^{(t)}(s,a_1) - \theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_1) \le 1/2. \tag{188}$$

Validating the constraint (183) and upper bounding $\min\{t_{\mathsf{tran}}, t_{\mathsf{ref}}\} - t_{\overline{s-1}}(\tau_s)$. It remains to justify the assumed condition (183) for all iteration within this stage. To this end, suppose instead that

$$t_{\overline{s-1}}(\tau_s) + \widetilde{t} \le \min\{t_{\mathsf{tran}}, t_{\mathsf{ref}}\}, \tag{189}$$

where $\widetilde{t}$ is defined in (183). We claim that the following relation is satisfied

$$\widehat{\pi}^{(t)}(a_2 \,|\, s) - \widehat{\pi}^{(t-1)}(a_2 \,|\, s) \geq \frac{c_{\mathrm{p}} c_{\mathrm{m}}}{150} \eta (1-\gamma)^2 \Big[\widehat{\pi}^{(t-1)}(a_2 \,|\, s)\Big]^2 \tag{190}$$

for any t obeying $t_{\overline{s-1}}(\tau_s) \leq t \leq t_{\overline{s-1}}(\tau_s) + \widetilde{t} \leq \min\{t_{\mathsf{tran}}, t_{\mathsf{ref}}\}$. Equipped with this recursive relation, we can invoke Lemma 11 to develop a lower bound on $\widehat{\pi}^{(t)}(a_2 \,|\, s)$, provided that an initial lower bound is available. In order to do so, in view of the expression (181), we can deduce that

$$\widehat{\pi}^{(t_{\overline{s-1}}(\tau_s))}(a_2 \,|\, s) = \exp\Big(2\theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_2) + \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1)\Big) \geq \exp\Big(\theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1)\Big),$$

where the last relation is due to the bound $\theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_2) \geq 0$ (see (36) in Lemma 6). Combining the above two inequalities and applying Lemma 11 (see (57b)), we arrive at $\pi^{(t)}(s, a_2) \geq 1/2$ — and hence $\pi^{(t)}(s, a_2) \geq \pi^{(t)}(s, a_0)$ — as soon as $t - t_{\overline{s-1}}(\tau_s)$ exceeds

$$\frac{1 + \frac{c_{\mathrm{p}} c_{\mathrm{m}}}{100} \eta (1-\gamma)^2}{\frac{c_{\mathrm{p}} c_{\mathrm{m}}}{150} \eta (1-\gamma)^2 \widehat{\pi}^{(t_{\overline{s-1}}(\tau_s))}(a_2 \,|\, s)}.$$

This together with the definition of t_{tran} thus indicates that

$$\begin{aligned}
t_{\mathsf{tran}} - t_{\overline{s-1}}(\tau_s) &\leq \frac{1 + \frac{c_{\mathrm{p}} c_{\mathrm{m}}}{150} \eta (1-\gamma)^2}{\frac{c_{\mathrm{p}} c_{\mathrm{m}}}{150} \eta (1-\gamma)^2 \widehat{\pi}^{(t_{\overline{s-1}}(\tau_s))}(a_2 \,|\, s)} \leq \frac{1 + \frac{c_{\mathrm{p}} c_{\mathrm{m}}}{150} \eta (1-\gamma)^2}{\frac{c_{\mathrm{p}} c_{\mathrm{m}}}{150} \eta (1-\gamma)^2 \exp\big(\theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1)\big)} \\
&\leq \frac{1.5}{\frac{c_{\mathrm{p}} c_{\mathrm{m}}}{150} \eta (1-\gamma)^2 \exp\big(\theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1)\big)},
\end{aligned}$$

provided that $\frac{c_{\mathrm{p}} c_{\mathrm{m}}}{150} \eta (1-\gamma)^2 \leq 0.5$. This, however, contradicts the assumption (189). As a consequence, we conclude that $t_{\overline{s-1}}(\tau_s) + \widetilde{t} > \min\{t_{\mathsf{tran}}, t_{\mathsf{ref}}\}$, thus indicating that

$$\min\{t_{\mathsf{tran}}, t_{\mathsf{ref}}\} - t_{\overline{s-1}}(\tau_s) \leq \widetilde{t} \leq \frac{225}{c_{\mathrm{p}} c_{\mathrm{m}} \eta (1-\gamma)^2 \exp\big(\theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1)\big)}. \tag{191}$$

Showing that $t_{\mathsf{tran}} = \min\{t_{\mathsf{tran}}, t_{\mathsf{ref}}\}$. We now justify that $t_{\mathsf{tran}} < t_{\mathsf{ref}}$, so that the upper bound (191) leads to an upper bound on $t_{\mathsf{tran}} - t_{\overline{s-1}}(\tau_s)$. Suppose instead that

$$t_{\mathsf{tran}} \geq t_{\mathsf{ref}}, \qquad \text{or equivalently,} \qquad t_{\mathsf{ref}} = \min\{t_{\mathsf{tran}}, t_{\mathsf{ref}}\},$$

and we would like to show that this leads to contradiction. By definition of t_{ref}, we have

$$\theta^{(t_{\mathsf{ref}})}(s, a_0) \leq \theta^{(t_{\mathsf{ref}})}(s, a_1) - \log\big(c_{\mathsf{ref}}(1-\gamma)\big).$$

This further yields

$$\begin{aligned}
\max\Big\{\theta^{(t_{\mathsf{ref}})}(s, a_0), \theta^{(t_{\mathsf{ref}})}(s, a_1)\Big\} &\leq \theta^{(t_{\mathsf{ref}})}(s, a_1) - \log\big(c_{\mathsf{ref}}(1-\gamma)\big) \\
&\leq \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1) + 1/2 - \log\big(c_{\mathsf{ref}}(1-\gamma)\big) < 0,
\end{aligned}$$

where the second inequality arises from (182), and the last one makes use of (176) as long as $t_{s-2}(\tau_{s-2})$. However, this together with the constraint $\sum_a \theta^{(t_{\mathsf{ref}})}(s, a) = 0$ implies that

$$\theta^{(t_{\mathsf{ref}})}(s, a_2) = -\theta^{(t_{\mathsf{ref}})}(s, a_0) - \theta^{(t_{\mathsf{ref}})}(s, a_1) > 0 > \max\Big\{\theta^{(t_{\mathsf{ref}})}(s, a_0), \theta^{(t_{\mathsf{ref}})}(s, a_1)\Big\}.$$

which, however, implies that $t_{\mathsf{ref}} > t_{\mathsf{tran}}$ (according to the definition of t_{tran}) and leads to contradiction. As a result, we conclude that

$$t_{\mathsf{tran}} < t_{\mathsf{ref}}, \tag{192}$$

and the bound (191) then indicates that

$$t_{\mathsf{tran}} - t_{\overline{s-1}}(\tau_s) \leq \frac{225}{c_{\mathrm{p}} c_{\mathrm{m}} \eta (1-\gamma)^2 \exp\big(\theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1)\big)}. \tag{193}$$

F.2.1 Proof of the inequality (190)

From the relation (181), one can deduce that

$$\begin{aligned}
\widehat{\pi}^{(t)}(a_2 \,|\, s) - \widehat{\pi}^{(t-1)}(a_2 \,|\, s) &= \exp\Big(2\theta^{(t)}(s,a_2) + \theta^{(t)}(s,a_1)\Big) - \exp\Big(2\theta^{(t-1)}(s,a_2) + \theta^{(t-1)}(s,a_1)\Big) \\
&= \widehat{\pi}^{(t-1)}(a_2 \,|\, s)\Big\{\exp\Big(2\theta^{(t)}(s,a_2) - 2\theta^{(t-1)}(s,a_2) + \theta^{(t)}(s,a_1) - \theta^{(t-1)}(s,a_1)\Big) - 1\Big\} \\
&\geq \widehat{\pi}^{(t-1)}(a_2 \,|\, s)\Big\{2\theta^{(t)}(s,a_2) - 2\theta^{(t-1)}(s,a_2) + \theta^{(t)}(s,a_1) - \theta^{(t-1)}(s,a_1)\Big\} \\
&= \widehat{\pi}^{(t-1)}(a_2 \,|\, s) \cdot \eta\Big(2\frac{\partial V^{(t-1)}(\mu)}{\partial \theta(s,a_2)} + \frac{\partial V^{(t-1)}(\mu)}{\partial \theta(s,a_1)}\Big)
\end{aligned} \tag{194}$$

for any t with $t_{\overline{s-1}}(\tau_s) \leq t \leq \min\{t_{\mathsf{tran}}, t_{\mathsf{ref}}\}$, where the inequality above follows from the elementary fact $e^x - 1 \geq x$ for any $x \in \mathbb{R}$. Therefore, the difference between $\widehat{\pi}^{(t)}(a_2 \,|\, s)$ and $\widehat{\pi}^{(t-1)}(a_2 \,|\, s)$ depends on both $\frac{\partial V^{(t-1)}(\mu)}{\partial \theta(s,a_2)}$ and $\frac{\partial V^{(t-1)}(\mu)}{\partial \theta(s,a_1)}$, motivating us to lower bound these two derivatives separately.

Step 1: bounding $\frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a_2)}$. First, we make the observation that for any $3 \leq s < H$ and any $t \geq t_{\overline{s-1}}(\tau_s)$,

$$Q^{(t)}(s,a_2) - Q^{(t)}(s,a_0) = \gamma p\big(V^{(t)}(\overline{s-2}) - \gamma\tau_{s-2}\big) \geq \gamma p\big(\gamma^{2s-3} - 1/4 - \gamma\tau_{s-2}\big) \geq \frac{p}{8} = \frac{c_{\mathrm{p}}(1-\gamma)}{8} \tag{195}$$

holds as long as $\gamma(\gamma^{2s-3} - 1/4 - \gamma\tau_s) \geq 1/8$. Here, the first identity comes from (47) in Lemma 8, and the first inequality holds for any $t \geq t_{\overline{s-2}}(\gamma^{2s-3} - 1/4)$ — a consequence of the monotonicity property in Lemma 9. As a result, for any t obeying $t_{\overline{s-1}}(\tau_s) \leq t \leq \min\{t_{\mathsf{tran}}, t_{\mathsf{ref}}\}$ we have

$$\begin{aligned}
Q^{(t)}(s,a_2) - V^{(t)}(s) &= \pi^{(t)}(a_0 \,|\, s)\Big(Q^{(t)}(s,a_2) - Q^{(t)}(s,a_0)\Big) + \pi^{(t)}(a_1 \,|\, s)\Big(Q^{(t)}(s,a_2) - Q^{(t)}(s,a_1)\Big) \\
&\geq \frac{c_{\mathrm{p}}(1-\gamma)}{24} - \pi^{(t)}(a_1 \,|\, s) \geq \frac{c_{\mathrm{p}}(1-\gamma)}{24} - c_{\mathsf{ref}}(1-\gamma) \geq \frac{c_{\mathrm{p}}(1-\gamma)}{48},
\end{aligned} \tag{196}$$

where the first inequality combines (195) with the facts that $\pi^{(t)}(a_0 \,|\, s) \geq 1/3$ (see (180)) and $0 \leq Q^{(t)}(s,a_2), Q^{(t)}(s,a_1) \leq 1$ (see Lemma 1), and the last line holds by observing (see (173))

$$\pi^{(t)}(a_1 \,|\, s) \leq c_{\mathsf{ref}}(1-\gamma)\pi^{(t_{\overline{s-1}}(\tau_s))}(a_0 \,|\, s) \leq c_{\mathsf{ref}}(1-\gamma) \qquad \text{for all } t \in [t_{\overline{s-1}}(\tau_s), t_{\mathsf{ref}})$$

and using the assumption $c_{\mathsf{ref}} \leq c_{\mathrm{p}}/2$. Consequently, for any $t \geq t_{\overline{s-1}}(\tau_s)$, the gradient w.r.t. $\theta(s,a_2)$ satisfies

$$\begin{aligned}
\frac{\partial V^{(t)}(s)}{\partial \theta(s,a_2)} &= \frac{1}{1-\gamma} d_\mu^{(t)}(s)\pi^{(t)}(a_2 \,|\, s)\Big(Q^{(t)}(s,a_2) - V^{(t)}(s)\Big) \\
&\geq \frac{c_{\mathrm{p}}c_{\mathrm{m}}\gamma}{48}(1-\gamma)^2\pi^{(t)}(a_2 \,|\, s) \geq \frac{c_{\mathrm{p}}c_{\mathrm{m}}\gamma}{144}(1-\gamma)^2\widehat{\pi}^{(t)}(a_2 \,|\, s) > 0,
\end{aligned} \tag{197}$$

where the first inequality above also makes use of the lower bound in Lemma 2.

In fact, the above lower bound holds true regardless of t, as long as $t \geq t_{\overline{s-1}}(\tau_s)$ where we have shown that $\frac{\partial V^{(t-1)}(\mu)}{\partial \theta(s,a_2)}$ is bounded from below by 0. One can thus conclude that the iterate $\theta^{(t)}(s,a_2)$ increases with t.

Step 2: bounding $\frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a_1)}$. Regarding the gradient w.r.t. $\theta(s,a_1)$, we have

$$\begin{aligned}
\frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a_1)} &= \frac{1}{1-\gamma} d_\mu^{(t)}(s)\pi^{(t)}(a_1 \,|\, s)\Big(Q^{(t)}(s,a_1) - V^{(t)}(s)\Big) \\
&= \frac{1}{1-\gamma} d_\mu^{(t)}(s)\pi^{(t)}(a_1 \,|\, s)\Big\{\pi^{(t)}(a_0 \,|\, s)\Big(Q^{(t)}(s,a_1) - Q^{(t)}(s,a_0)\Big) + \pi^{(t)}(a_2 \,|\, s)\Big(Q^{(t)}(s,a_1) - Q^{(t)}(s,a_2)\Big)\Big\} \\
&\geq \frac{1}{1-\gamma} d_\mu^{(t)}(s)\pi^{(t)}(a_1 \,|\, s)\Big(\pi^{(t)}(a_0 \,|\, s) + \pi^{(t)}(a_2 \,|\, s)\Big)(\gamma\tau_s - \gamma^{\frac{1}{2}}\tau_s),
\end{aligned}$$

where the last line follows since (see Lemma 8 and the fact that $t \geq t_{\overline{s-1}}(\tau_s)$)

$$\max\left\{Q^{(t)}(s,a_0), Q^{\pi}(s,a_2)\right\} \leq \gamma^{\frac{1}{2}}\tau_s, \qquad Q^{(t)}(s,a_1) = \gamma V^{(t)}(\overline{s-1}) \geq \gamma\tau_s.$$

In addition, recognizing that $\pi^{(t)}(a_0 \mid s) + \pi^{(t)}(a_2 \mid s) \leq 1$ and $d_{\mu}^{(t)}(s) \leq 14c_{\mathrm{m}}(1-\gamma)^2$ (see Lemma 3), we can continue the above bound to obtain

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a_1)} \geq \; -14c_{\mathrm{m}}(1-\gamma)\pi^{(t)}(a_1 \mid s)\tau_s\gamma^{\frac{1}{2}}\frac{1-\gamma}{1+\sqrt{\gamma}} \geq -7c_{\mathrm{m}}(1-\gamma)^2\widehat{\pi}^{(t)}(a_1 \mid s), \tag{198}$$

where the last inequality is due to $\tau_s \leq 1/2$ and $0 < \gamma < 1$ and the bound (120).

Step 3: connecting $\widehat{\pi}^{(t)}(a_1 \mid s)$ with $\widehat{\pi}^{(t)}(a_2 \mid s)$. The above lower bound (198) on $\frac{\partial V^{(t)}(\mu)}{\partial \theta(s,a_1)}$ is dependent on $\widehat{\pi}^{(t)}(a_1 \mid s)$. However, the desired lower bound (190) is only a function of $\widehat{\pi}^{(t)}(a_2 \mid s)$. This motivates us to investigate the connection between $\widehat{\pi}^{(t)}(a_1 \mid s)$ and $\widehat{\pi}^{(t)}(a_2 \mid s)$.

To this end, let us write

$$\widehat{\pi}^{(t-1)}(a_1 \mid s) = \widehat{\pi}^{(t-1)}(a_2 \mid s)\exp\Big(\theta^{(t-1)}(s,a_1) - \theta^{(t-1)}(s,a_2)\Big). \tag{199}$$

As a result, one only needs to control the quantity $\exp\big(\theta^{(t-1)}(s,a_1) - \theta^{(t-1)}(s,a_2)\big)$. In order to do so, we make use of the induction hypothesis (182) for the $(t-1)$-th iteration to show that

$$\begin{aligned}
\exp\Big(\theta^{(t-1)}(s,a_1) - \theta^{(t-1)}(s,a_2)\Big) &\leq \exp\Big(\theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_1) + 1/2 - \theta^{(t-1)}(s,a_2)\Big) \\
&\overset{\text{(i)}}{\leq} \exp\Big(\theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_1) + 1/2 - \theta^{(t_{\overline{s-1}}(\tau_s))}(s,a_2)\Big) \\
&\overset{\text{(ii)}}{\leq} \exp\Big(\theta^{(t_{s-2}(\tau_{s-2}))}(s,a_1) + 1/2\Big).
\end{aligned}$$

Here, (i) follows from the fact that $\theta^{(t)}(s,a_2)$ increases with t (see (197)); and (ii) comes from the inequality (36) in Lemma 6 as well as (176). Recalling Lemma 5, one has

$$\begin{aligned}
\exp\Big(\theta^{(t-1)}(s,a_1) - \theta^{(t-1)}(s,a_2)\Big) &\leq \exp\Big(\theta^{(t_{s-2}(\tau_{s-2}))}(s,a_1) + 1/2\Big) \\
&\leq \frac{\exp(1/2)}{\sqrt{1+\frac{c_{\mathrm{m}}\gamma}{35}\eta(1-\gamma)^2 t_{s-2}(\tau_{s-2})}} \leq \frac{\gamma c_{\mathrm{p}}}{1050},
\end{aligned} \tag{200}$$

where the last inequality is satisfied provided that $t_{s-2}(\tau_{s-2}) \geq \frac{1050^2 e}{\frac{c_{\mathrm{m}}\gamma^3}{35}\eta(1-\gamma)^2 c_{\mathrm{p}}^2}$. Combining (198) with (199) and (200), we arrive at

$$\frac{\partial V^{(t-1)}(\mu)}{\partial \theta(s,a_1)} \geq \; -\frac{c_{\mathrm{p}}c_{\mathrm{m}}}{150}(1-\gamma)^2\widehat{\pi}^{(t-1)}(a_2 \mid s). \tag{201}$$

Step 4: combining bounds. Putting the above pieces together and invoking the expression (194) yield for $\gamma > 0.96$,

$$\begin{aligned}
\widehat{\pi}^{(t)}(a_2 \mid s) - \widehat{\pi}^{(t-1)}(a_2 \mid s) &\geq \widehat{\pi}^{(t-1)}(a_2 \mid s)\cdot\eta\Big(2\frac{\partial V^{(t-1)}(\mu)}{\partial \theta(s,a_2)} + \frac{\partial V^{(t-1)}(\mu)}{\partial \theta(s,a_1)}\Big) \\
&\geq \Big[\widehat{\pi}^{(t-1)}(a_2 \mid s)\Big]^2\eta\frac{c_{\mathrm{p}}c_{\mathrm{m}}}{150}(1-\gamma)^2,
\end{aligned}$$

which concludes the proof of the advertised bound (190).

F.3 Stage II: the duration where $\theta^{(t)}(s, a_2) \geq \theta^{(t)}(s, a_0)$

We now turn attention to the case where t lies within $[t_{\mathsf{tran}}, t_{\mathsf{ref}})$, which is a non-empty interval according to (192). In this case one has

$$\theta^{(t)}(s, a_2) \geq \theta^{(t)}(s, a_0), \qquad \text{or equivalently,} \qquad \pi^{(t)}(s, a_2) \geq \pi^{(t)}(s, a_0), \tag{202}$$

as a consequence of the definition (177) of t_{tran}. Again, from the definition (173) of t_{ref}, the inequality $c_{\mathsf{ref}}(1-\gamma)\pi^{(t)}(a_0 \,|\, s) > \pi^{(t)}(a_1 \,|\, s)$ holds true for every $t \in [t_{\mathsf{tran}},\ t_{\mathsf{ref}})$, or equivalently,

$$\theta^{(t)}(s, a_0) \geq \theta^{(t)}(s, a_1) - \log\big(c_{\mathsf{ref}}(1-\gamma)\big) \qquad \text{for all } t \in [t_{\mathsf{tran}},\ t_{\mathsf{ref}}). \tag{203}$$

The goal of this subsection is to show that $\theta^{(t)}(s, a_1) - \theta^{(t_{\mathsf{tran}})}(s, a_1) \leq 1/2$ throughout this stage.

Preparation. From the above conditions (202) and (203), we have

$$\pi^{(t)}(s, a_2) \geq \pi^{(t)}(s, a_0) \geq \pi^{(t)}(s, a_1) \qquad \text{and hence} \qquad \pi^{(t)}(s, a_2) \geq 1/3. \tag{204}$$

We now look at the gradient w.r.t. $\theta(s, a_0)$, for which we first observe that

$$\begin{aligned} Q^{(t)}(s, a_0) - V^{(t)}(s) &= \pi^{(t)}(a_2 \,|\, s)\Big(Q^{(t)}(s, a_0) - Q^{(t)}(s, a_2)\Big) + \pi^{(t)}(a_1 \,|\, s)\Big(Q^{(t)}(s, a_0) - Q^{(t)}(s, a_1)\Big) \\ &\overset{\text{(i)}}{\leq} -\frac{c_{\mathrm{p}}(1-\gamma)}{24} + c_{\mathsf{ref}}(1-\gamma) \overset{\text{(ii)}}{\leq} -\frac{c_{\mathrm{p}}(1-\gamma)}{36} < 0. \end{aligned} \tag{205}$$

Here, (i) follows from the inequalities (195) and (204), whereas (ii) holds true as long as $c_{\mathsf{ref}} \leq c_{\mathrm{p}}/72$. Consequently,

$$\frac{\partial V^{(t)}(\mu)}{\partial \theta(s, a_0)} = \frac{1}{1-\gamma} d_{\mu}^{(t)}(s)\pi^{(t)}(a_0 \,|\, s)\Big(Q^{(t)}(s, a_0) - V^{(t)}(s)\Big) < 0,$$

thus indicating that $\theta^{(t)}(s, a_0)$ is decreasing with t.

Key induction hypotheses. Again, we seek to prove by induction that

$$\theta^{(t)}(s, a_1) - \theta^{(t_{\mathsf{tran}})}(s, a_1) \leq 1/2, \qquad t \in [t_{\mathsf{tran}}, t_{\mathsf{ref}}). \tag{206}$$

For the base case where $t = t_{\mathsf{tran}}$, this claim trivially holds true. Now suppose that the induction hypothesis (206) is satisfied for every iteration up to $t-1$, and we would like to establish it for the t-th iteration. Towards this, we find it helpful to introduce another auxiliary induction hypothesis

$$\widehat{\pi}^{(i)}(a_0 \,|\, s) \leq \frac{1}{1 + \frac{c_{\mathrm{p}} c_{\mathrm{m}}}{288}\eta(1-\gamma)^2(i - t_{\mathsf{tran}})} \qquad \text{for all } i \in [t_{\mathsf{tran}}, t). \tag{207}$$

As an immediate remark, this hypothesis trivially holds true when $t = t_{\mathsf{tran}} + 1$. In what follows, we shall first establish (206) for the t-th iteration assuming satisfaction of (207), and then use to demonstrate that (207) holds for $i = t$ as well.

Inductive step 1: showing that $\theta^{(t)}(s, a_1) - \theta^{(t_{\mathsf{tran}})}(s, a_1) \leq 1/2$. Towards this, let us introduce for convenience another time instance

$$\widetilde{t} \coloneqq \arg \max_{i:\, t_{\mathsf{tran}} \leq i < t} \theta^{(i)}(s, a_1), \tag{208}$$

which reflects the time when $\theta^{(i)}(s, a_1)$ reaches its maximum before iteration t. In order to establish the induction hypothesis (206) for the t-th iteration, it is sufficient to demonstrate that

$$\theta^{(\widetilde{t})}(s, a_1) - \theta^{(t_{\mathsf{tran}})}(s, a_1) \leq 1/2. \tag{209}$$

As before, let us employ the PG update rule (9a) to expand $\theta^{(\widetilde{t})}(s,a_1) - \theta^{(t_{\mathsf{tran}})}(s,a_1)$ as follows

$$\theta^{(\widetilde{t})}(s,a_1) - \theta^{(t_{\mathsf{tran}})}(s,a_1) = \sum_{i=t_{\mathsf{tran}}}^{\widetilde{t}-1} \eta \frac{\partial V^{(i)}(\mu)}{\partial \theta(s,a_1)}. \tag{210}$$

For each gradient $\frac{\partial V^{(i)}(\mu)}{\partial \theta(s,a_1)}$, invoking Lemma 3, Lemma 1 and Lemma 10 tells us that

$$\frac{\partial V^{(i)}(\mu)}{\partial \theta(s,a_1)} = \frac{1}{1-\gamma} d_\mu^{(i)}(s)\pi^{(i)}(a_1 \,|\, s)\big(Q^{(i)}(s,a_1) - V^{(i)}(s)\big) \le 14c_{\mathrm{m}}(1-\gamma)\pi^{(i)}(a_1 \,|\, s). \tag{211}$$

In addition, a little algebra together with (204) leads to

$$\begin{aligned}\pi^{(i)}(a_1 \,|\, s) \le \widehat{\pi}^{(i)}(a_1 \,|\, s) &= \exp\Big(\theta^{(i)}(s,a_1) - \theta^{(i)}(s,a_2)\Big) \overset{\text{(i)}}{=} \exp\Big(\frac{3}{2}\theta^{(i)}(s,a_1) + \frac{1}{2}\theta^{(i)}(s,a_0) - \frac{1}{2}\theta^{(i)}(s,a_2)\Big) \\ &= \exp\Big(\frac{3}{2}\theta^{(i)}(s,a_1)\Big)\sqrt{\widehat{\pi}^{(i)}(a_0 \,|\, s)} \overset{\text{(ii)}}{\le} \exp\Big(\frac{3}{2}\theta^{(\widetilde{t})}(s,a_1)\Big)\frac{1}{\sqrt{1+\frac{c_{\mathrm{p}}c_{\mathrm{m}}}{288}\eta(1-\gamma)^2(i-t_{\mathsf{tran}})}}\end{aligned}$$

for any i obeying $t_{\overline{s-1}}(\tau_s) \le i < \widetilde{t}$, where the first inequality comes from (120), (i) makes use of $\sum_a \theta^{(i)}(s,a) = 0$, and (ii) follows from the induction hypothesis (207) along with the definition (208) of $\widetilde{t}$.

Putting the above bounds together with (210) and (211) guarantees that

$$\begin{aligned}\theta^{(\widetilde{t})}(s,a_1) - \theta^{(t_{\mathsf{tran}})}(s,a_1) &\le \sum_{i=t_{\mathsf{tran}}}^{\widetilde{t}-1} 14c_{\mathrm{m}}\eta(1-\gamma)\exp\Big(\frac{3}{2}\theta^{(\widetilde{t})}(s,a_1)\Big)\frac{1}{\sqrt{1+\frac{c_{\mathrm{p}}c_{\mathrm{m}}}{288}\eta(1-\gamma)^2(i-t_{\mathsf{tran}})}} \\ &= \frac{14c_{\mathrm{m}}\eta\exp\big(\frac{3}{2}\theta^{(\widetilde{t})}(s,a_1)\big)}{\sqrt{\frac{c_{\mathrm{p}}c_{\mathrm{m}}}{288}\eta}}\left\{1+\sum_{i=t_{\mathsf{tran}}+1}^{\widetilde{t}-1}\frac{1}{\sqrt{i-t_{\mathsf{tran}}}}\right\} \\ &\le \sqrt{\frac{225792c_{\mathrm{m}}\eta(\widetilde{t}-t_{\mathsf{tran}})}{c_{\mathrm{p}}}}\exp\Big(\frac{3}{2}\theta^{(\widetilde{t})}(s,a_1)\Big).\end{aligned} \tag{212}$$

Given that $\theta^{(\widetilde{t})}(s,a_0) \ge \theta^{(\widetilde{t})}(s,a_1) - \log\big(c_{\mathsf{ref}}(1-\gamma)\big)$ (see (203)) and $\sum_a \theta^{(\widetilde{t})}(s,a) = 0$, one obtains

$$\begin{aligned}\widehat{\pi}^{(\widetilde{t})}(a_0 \,|\, s) = \exp\Big(\theta^{(\widetilde{t})}(s,a_0) - \theta^{(\widetilde{t})}(s,a_2)\Big) &= \exp\Big(2\theta^{(\widetilde{t})}(s,a_0) + \theta^{(\widetilde{t})}(s,a_1)\Big) \\ &\ge \exp\Big(3\theta^{(\widetilde{t})}(s,a_1) - 2\log\big(c_{\mathsf{ref}}(1-\gamma)\big)\Big),\end{aligned}$$

which combined with the inequality (207) thus implies that

$$\exp\Big(\frac{3}{2}\theta^{(\widetilde{t})}(s,a_1)\Big) \le c_{\mathsf{ref}}(1-\gamma)\sqrt{\widehat{\pi}^{(\widetilde{t})}(a_0 \,|\, s)} \le \frac{c_{\mathsf{ref}}(1-\gamma)}{\sqrt{\frac{c_{\mathrm{p}}c_{\mathrm{m}}}{288}\eta(1-\gamma)^2(\widetilde{t}-t_{\mathsf{tran}})}}. \tag{213}$$

As a consequence of the inequalities (212) and (213), we obtain

$$\begin{aligned}\theta^{(\widetilde{t})}(s,a_1) - \theta^{(t_{\mathsf{tran}})}(s,a_1) &\le \sqrt{\frac{225792c_{\mathrm{m}}\eta(\widetilde{t}-t_{\mathsf{tran}})}{c_{\mathrm{p}}}}\frac{c_{\mathsf{ref}}(1-\gamma)}{\sqrt{\frac{c_{\mathrm{p}}c_{\mathrm{m}}}{288}\eta(1-\gamma)^2(\widetilde{t}-t_{\mathsf{tran}})}} \\ &\le \frac{8064c_{\mathsf{ref}}}{c_{\mathrm{p}}} < \frac{1}{2},\end{aligned} \tag{214}$$

where the last line holds as long as $c_{\mathsf{ref}} < c_{\mathrm{p}}/16128$. This in turn establishes our induction hypothesis (209) — and hence (208) for the t-th iteration — assuming satisfaction of the hypothesis (207).

Inductive step 2: establishing the upper bound (207). The next step is thus to justify the induction hypothesis (207) when $i = t$. To do so, we first pay attention to the dynamics of $\theta^{(i)}(s, a_0)$ for any $t_{\mathsf{tran}} \le i \le t$. Recognizing that $\theta^{(i)}(s, a_2) = \max_a \theta^{(i)}(s, a)$ (see (204)) and $\sum_a \theta^{(i)}(s, a) = 0$, we can express

$$\widehat{\pi}^{(i)}(a_0 \,|\, s) = \exp\Big(\theta^{(i)}(s, a_0) - \theta^{(i)}(s, a_2)\Big) = \exp\Big(2\theta^{(i)}(s, a_0) + \theta^{(i)}(s, a_1)\Big).$$

This allows one to obtain

$$\begin{aligned}
\widehat{\pi}^{(i)}(a_0 \,|\, s) - \widehat{\pi}^{(i+1)}(a_0 \,|\, s) &= \exp\Big(2\theta^{(i)}(s, a_0) + \theta^{(i)}(s, a_1)\Big) - \exp\Big(2\theta^{(i+1)}(s, a_0) + \theta^{(i+1)}(s, a_1)\Big) \\
&= \widehat{\pi}^{(i)}(a_0 \,|\, s)\Big\{1 - \exp\Big(2\theta^{(i+1)}(s, a_0) - 2\theta^{(i)}(s, a_0) + \theta^{(i+1)}(s, a_1) - \theta^{(i)}(s, a_1)\Big)\Big\} \\
&= \widehat{\pi}^{(i)}(a_0 \,|\, s)\Big\{1 - \exp\Big(2\eta\frac{\partial V^{(i)}(\mu)}{\partial \theta(s, a_0)} + \eta\frac{\partial V^{(i)}(\mu)}{\partial \theta(s, a_1)}\Big)\Big\}.
\end{aligned} \tag{215}$$

With the above observation in mind, we claim for the moment the following recursive relation

$$\widehat{\pi}^{(i)}(a_0 \,|\, s) - \widehat{\pi}^{(i+1)}(a_0 \,|\, s) \ge \frac{c_{\mathrm{p}}c_{\mathrm{m}}}{288}\eta(1-\gamma)^2\Big[\widehat{\pi}^{(i)}(a_0 \,|\, s)\Big]^2 \tag{216}$$

for any i obeying $t_{\mathsf{tran}} \le i < t$, whose proof is deferred to the end of this section. If this claim were true, then (55b) in Lemma 11 allows us to conclude the desired bound

$$\widehat{\pi}^{(t)}(a_0 \,|\, s) \le \frac{1}{1 + \frac{c_{\mathrm{p}}c_{\mathrm{m}}}{288}\eta(1-\gamma)^2(t - t_{\mathsf{tran}})}. \tag{217}$$

Proof of the inequality (216). Combining (205) and the lower bound on $d_\mu^{(i)}(s)$ in Lemma 2, we have

$$\frac{\partial V^{(i)}(\mu)}{\partial \theta(s, a_0)} \le c_{\mathrm{m}}\gamma(1-\gamma)\pi^{(i)}(a_0 \,|\, s)\Big(Q^{(i)}(s, a_0) - V^{(i)}(s)\Big) \le -\frac{c_{\mathrm{p}}c_{\mathrm{m}}}{108}(1-\gamma)^2\widehat{\pi}^{(i)}(a_0 \,|\, s),$$

where the last inequality also makes use of (120). In addition, invoking the inequalities (211) and (120) gives

$$\begin{aligned}
\frac{\partial V^{(i)}(\mu)}{\partial \theta(s, a_1)} &\le 14c_{\mathrm{m}}(1-\gamma)\pi^{(i)}(a_1 \,|\, s) \le 14c_{\mathrm{m}}(1-\gamma)\widehat{\pi}^{(i)}(a_1 \,|\, s) \\
&= 14c_{\mathrm{m}}(1-\gamma)\widehat{\pi}^{(i)}(a_0 \,|\, s)\exp\Big(\theta^{(i)}(s, a_1) - \theta^{(i)}(s, a_0)\Big).
\end{aligned} \tag{218}$$

Recall that for any $i \in [t_{\mathsf{tran}},\ t_{\mathsf{ref}})$, one has $\theta^{(i)}(s, a_0) \ge \theta^{(i)}(s, a_1) - \log\big(c_{\mathsf{ref}}(1-\gamma)\big)$, or equivalently,

$$\exp\Big(\theta^{(i)}(s, a_1) - \theta^{(i)}(s, a_0)\Big) \le c_{\mathsf{ref}}(1-\gamma).$$

It thus follows that

$$\frac{\partial V^{(i)}(\mu)}{\partial \theta(s, a_1)} \le 14c_{\mathsf{ref}}c_{\mathrm{m}}(1-\gamma)^2\widehat{\pi}^{(i)}(a_0 \,|\, s).$$

As a result, the above bounds taken collectively lead to

$$2\frac{\partial V^{(i)}(\mu)}{\partial \theta(s, a_0)} + \frac{\partial V^{(i)}(\mu)}{\partial \theta(s, a_1)} \le \Big[-\frac{c_{\mathrm{p}}c_{\mathrm{m}}}{56}(1-\gamma)^2 + 14c_{\mathsf{ref}}c_{\mathrm{m}}(1-\gamma)^2\Big]\widehat{\pi}^{(i)}(a_0 \,|\, s) \le -\frac{c_{\mathrm{p}}c_{\mathrm{m}}}{112}(1-\gamma)^2\widehat{\pi}^{(i)}(a_0 \,|\, s),$$

provided that $c_{\mathsf{ref}}/c_{\mathrm{p}} < 1/1568$. In addition, similar to (218), we can easily see that

$$\eta\Big|\frac{\partial V^{(i)}(\mu)}{\partial \theta(s, a_1)}\Big| \le 14\eta c_{\mathrm{m}}(1-\gamma)\pi^{(i)}(a_1 \,|\, s) \le 14\eta c_{\mathrm{m}}(1-\gamma) \le 1/3, \tag{219a}$$

$$\eta\Big|\frac{\partial V^{(i)}(\mu)}{\partial \theta(s, a_0)}\Big| \le 14\eta c_{\mathrm{m}}(1-\gamma)\pi^{(i)}(a_0 \,|\, s) \le 14\eta c_{\mathrm{m}}(1-\gamma) \le 1/3 \tag{219b}$$

as long as $\eta c_{\mathrm{m}}(1-\gamma) \leq 1/42$.

Substituting the preceding bounds into (215), we immediately arrive at

$$\begin{aligned}
\widehat{\pi}^{(i)}(a_0 \,|\, s) - \widehat{\pi}^{(i+1)}(a_0 \,|\, s) &= \widehat{\pi}^{(i)}(a_0 \,|\, s) \left\{1 - \exp\left(2\eta \frac{\partial V^{(i)}(\mu)}{\partial \theta(s, a_0)} + \eta \frac{\partial V^{(i)}(\mu)}{\partial \theta(s, a_1)}\right)\right\} \\
&\geq \widehat{\pi}^{(i)}(a_0 \,|\, s) \frac{\eta}{2}\Big(-2 \frac{\partial V^{(i)}(\mu)}{\partial \theta(s, a_0)} - \frac{\partial V^{(i)}(\mu)}{\partial \theta(s, a_1)}\Big) \\
&\geq \frac{\eta c_{\mathrm{p}} c_{\mathrm{m}}}{224}(1-\gamma)^2 \Big[\widehat{\pi}^{(i)}(a_0 \,|\, s)\Big]^2,
\end{aligned}$$

where the first inequality holds due to the fact $-1 \leq 2\eta \frac{\partial V^{(i)}(\mu)}{\partial \theta(s,a_0)} + \eta \frac{\partial V^{(i)}(\mu)}{\partial \theta(s,a_1)} \leq 0$ as well as the elementary inequality $1 - e^x \geq -x/2$ as long as $-1 \leq x \leq 0$. This establishes the inequality (216).

F.4 Proof of the claims (38a) and (38b)

We are now ready to justify the claims (38a) and (38b). Combining (182) and (206), we reach

$$\begin{aligned}
\theta^{(t)}(s, a_1) - \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1) &= \begin{cases} \theta^{(t)}(s, a_1) - \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1), & \text{if } t_{\overline{s-1}}(\tau_s) \leq t \leq t_{\mathsf{tran}} \\ \Big(\theta^{(t_{\mathsf{tran}})}(s, a_1) - \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1)\Big) + \Big(\theta^{(t)}(s, a_1) - \theta^{(t_{\mathsf{tran}})}(s, a_1)\Big), & \text{if } t_{\mathsf{tran}} \leq t \leq t_{\mathsf{ref}} \end{cases} \\
&\leq \max_{t_{\overline{s-1}}(\tau_s) \leq i \leq t_{\mathsf{tran}}} \Big(\theta^{(i)}(s, a_1) - \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1)\Big) + \max_{t_{\mathsf{tran}} \leq i < t_{\mathsf{ref}}} \Big(\theta^{(i)}(s, a_1) - \theta^{(t_{\mathsf{tran}})}(s, a_1)\Big) \leq 1.
\end{aligned}$$

This taken collectively with (176) leads to

$$\theta^{(t_{\mathsf{ref}})}(s, a_1) \leq \theta^{(t_{\overline{s-1}}(\tau_s))}(s, a_1) + 1 \leq -\frac{1}{2}\log\Big(1 + \frac{c_{\mathrm{m}}\gamma}{35}\eta(1-\gamma)t_{s-2}(\tau_{s-2})\Big) + 1,$$

as claimed in (38b).

In addition, recalling the definition (173) of t_{ref}, we have

$$\theta^{(t_{\mathsf{ref}})}(s, a_0) \leq \theta^{(t_{\mathsf{ref}})}(s, a_1) - \log\big(c_{\mathsf{ref}}(1-\gamma)\big),$$

which clearly satisfies (38a) as long as $c_{\mathsf{ref}} \geq c_{\mathrm{p}}/16128$.

F.5 Proof of the claim (38c)

Finally, we move on to analyze what happens after iteration t_{ref}, for which we focus on tracking the changes of $\widehat{\pi}^{(t)}(a_1 \,|\, s)$. In this part, let us only consider the set of t satisfying

$$\pi^{(t)}(a_1 \,|\, s) \leq \pi^{(t)}(a_2 \,|\, s).$$

Note that at time t_{ref}, the inequalities (38a) and (38b) are both satisfied, which together with the property $\pi^{(t)}(a_1 \,|\, s) \leq \pi^{(t)}(a_2 \,|\, s)$ yield

$$\widehat{\pi}^{(t)}(a_1 \,|\, s) := \exp\Big(\theta^{(t)}(s, a_1) - \max_a \theta^{(t)}(s, a)\Big) = \exp\Big(\theta^{(t)}(s, a_1) - \theta^{(t)}(s, a_2)\Big).$$

Then, if $c_{\mathsf{ref}} < c_{\mathrm{p}}/1000$, we have

$$\begin{aligned}
\widehat{\pi}^{(t+1)}(a_1 \,|\, s) - \widehat{\pi}^{(t)}(a_1 \,|\, s) &= \exp\Big(\theta^{(t+1)}(s, a_1) - \theta^{(t+1)}(s, a_2)\Big) - \exp\Big(\theta^{(t)}(s, a_1) - \theta^{(t)}(s, a_2)\Big) \\
&= \widehat{\pi}^{(t)}(a_1 \,|\, s)\Big\{\exp\Big(\theta^{(t+1)}(s, a_1) - \theta^{(t+1)}(s, a_2) - \theta^{(t)}(s, a_1) + \theta^{(t)}(s, a_2)\Big) - 1\Big\} \\
&= \widehat{\pi}^{(t)}(a_1 \,|\, s) \max\Big\{\exp\Big(\eta \frac{\partial V^{(t)}(\mu)}{\partial \theta(s, a_1)} - \eta \frac{\partial V^{(t)}(\mu)}{\partial \theta(s, a_2)}\Big) - 1,\ 0\Big\} \\
&\leq \widehat{\pi}^{(t)}(a_1 \,|\, s) \cdot 2\eta \max\Big\{\frac{\partial V^{(t)}(\mu)}{\partial \theta(s, a_1)} - \frac{\partial V^{(t)}(\mu)}{\partial \theta(s, a_2)},\ 0\Big\}
\end{aligned}$$

$$\leq 56c_{\mathrm{m}}\eta(1-\gamma)^2\Big[\widehat{\pi}^{(t)}(a_1\,|\,s)\Big]^2. \tag{220}$$

Here, the first inequality holds if $\eta\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)}-\eta\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_2)}\leq 1$ (given the elementary fact $e^x-1\leq 2x$ for any $0\leq x\leq 1$), and the last line is valid since

$$\begin{aligned}
\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)} &= \frac{1}{1-\gamma}d_\mu^{(t)}(s)\pi^{(t)}(a_1\,|\,s)\Big(Q^{(t)}(s,a_1)-V^{(t)}(s)\Big)\overset{\text{(i)}}{\leq}14c_{\mathrm{m}}(1-\gamma)\pi^{(t)}(a_1\,|\,s),\\
\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_2)} &= \frac{1}{1-\gamma}d_\mu^{(t)}(s)\Big\{\pi^{(t)}(a_1\,|\,s)\Big(Q^{(t)}(s,a_2)-Q^{(t)}(s,a_1)\Big)+\pi^{(t)}(a_0\,|\,s)\Big(Q^{(t)}(s,a_2)-Q^{(t)}(s,a_0)\Big)\Big\}\\
&\overset{\text{(ii)}}{\geq}\frac{1}{1-\gamma}d_\mu^{(t)}(s)\pi^{(t)}(a_1\,|\,s)\Big(Q^{(t)}(s,a_2)-Q^{(t)}(s,a_1)\Big)\overset{\text{(iii)}}{\geq}-14c_{\mathrm{m}}(1-\gamma)\pi^{(t)}(a_1\,|\,s),
\end{aligned}$$

where (ii) holds since $Q^{(t)}(s,a_2)\geq Q^{(t)}(s,a_0)$ (cf. (195)), and (i) and (iii) make use of Lemma 1 and Lemma 3. In addition, these bounds also imply that $\eta\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_1)}-\eta\frac{\partial V^{(t)}(\mu)}{\partial\theta(s,a_2)}\leq 1$ hold as long as $28\eta c_{\mathrm{m}}(1-\gamma)\leq 1$, thus validating the argument for the first inequality in (220).

Armed with the above recursive relation (220), we can invoke Lemma 11 to show that

$$t_s(\tau_s)-t_{\mathsf{ref}}\geq\frac{\frac{1}{\widehat{\pi}^{(t_{\mathsf{ref}})}(a_1\,|\,s)}-\frac{1}{\widehat{\pi}^{(t_s(\tau_s))}(a_1\,|\,s)}}{56c_{\mathrm{m}}\eta(1-\gamma)^2}\geq\frac{\frac{1}{\widehat{\pi}^{(t_{\mathsf{ref}})}(a_1\,|\,s)}-\frac{2}{1-\gamma}}{56c_{\mathrm{m}}\eta(1-\gamma)^2}, \tag{222}$$

where the last inequality holds since (in view of (120) and (48)).

$$\widehat{\pi}^{(t)}(a_1\,|\,s)\geq\pi^{(t)}(a_1\,|\,s)\geq(1-\gamma)/2\qquad\text{for any } t\geq t_s(\tau_s).$$

In order to control $t_s(\tau_s)-t_{\mathsf{ref}}$ via (222), it remains to upper bound $\widehat{\pi}^{(t_{\mathsf{ref}})}(a_1\,|\,s)$. Towards this end, it is seen that

$$\begin{aligned}
\widehat{\pi}^{(t_{\mathsf{ref}})}(a_1\,|\,s) &= \exp\Big(\theta^{(t_{\mathsf{ref}})}(s,a_1)-\theta^{(t_{\mathsf{ref}})}(s,a_2)\Big)=\exp\Big(2\theta^{(t_{\mathsf{ref}})}(s,a_1)+\theta^{(t_{\mathsf{ref}})}(s,a_0)\Big)\\
&\leq\exp\Big(3\theta^{(t_{\mathsf{ref}})}(s,a_1)-\log\Big(\frac{c_{\mathrm{p}}(1-\gamma)}{16128}\Big)\Big)\\
&\leq\frac{16128e^3}{c_{\mathrm{p}}(1-\gamma)\Big(1+\frac{c_{\mathrm{m}}\gamma}{35}\eta(1-\gamma)^2t_{s-2}(\tau_{s-2})\Big)^{1.5}}\leq\frac{16128e^3}{c_{\mathrm{p}}(1-\gamma)\Big(\frac{c_{\mathrm{m}}\gamma}{35}\eta(1-\gamma)^2t_{s-2}(\tau_{s-2})\Big)^{1.5}},
\end{aligned} \tag{223}$$

where the first line uses $\sum_a\theta^{(t_{\mathsf{ref}})}(s,a)=0$, the second line relies on the inequality (38a), and the last one applies the inequality (38b). Substitution into the relation (222) yields

$$t_s(\tau_s)-t_{\mathsf{ref}}\geq 10^{-10}c_{\mathrm{p}}c_{\mathrm{m}}^{0.5}\eta^{0.5}(1-\gamma)^2\Big(t_{s-2}(\tau_{s-2})\Big)^{1.5},$$

thus establishing the advertised bound.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2019). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv preprint arXiv:1908.00261*.

Agazzi, A. and Lu, J. (2021). Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime. *International Conference on Learning Representations (ICLR)*.

Asadi, K. and Littman, M. L. (2017). An alternative softmax operator for reinforcement learning. In *International Conference on Machine Learning*, pages 243–252. PMLR.

Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Bhandari, J. (2020). *Optimization Foundations of Reinforcement Learning*. PhD thesis, Columbia University.

Bhandari, J. and Russo, D. (2019). Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*.

Bhandari, J. and Russo, D. (2020). A note on the linear convergence of policy gradient methods. *arXiv preprint arXiv:2007.11120*.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2019). Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2020). Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*.

Ding, D., Zhang, K., Basar, T., and Jovanovic, M. (2020). Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33.

Du, S., Jin, C., Jordan, M., Póczos, B., Singh, A., and Lee, J. (2017). Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1068–1078.

Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476.

Jansch-Porto, J. P., Hu, B., and Dullerud, G. (2020). Convergence guarantees of policy optimization methods for Markovian jump linear systems. *arXiv preprint arXiv:2002.04090*.

Kakade, S. M. (2002). A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538.

Khamaru, K., Pananjady, A., Ruan, F., Wainwright, M. J., and Jordan, M. I. (2020). Is temporal difference learning optimal? an instance-dependent analysis. *arXiv preprint arXiv:2003.07337*.

Khodadadian, S., Doan, T. T., Maguluri, S. T., and Romberg, J. (2021). Finite sample analysis of two-time-scale natural actor-critic algorithm. *arXiv preprint arXiv:2101.10506*.

Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014.

Lan, G. (2021). Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135*.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. (2016). Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257. PMLR.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. (2019). Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*.

Liu, Y., Zhang, K., Basar, T., and Yin, W. (2020). An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33.

Mei, J., Xiao, C., Dai, B., Li, L., Szepesvári, C., and Schuurmans, D. (2020a). Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, 33.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020b). On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Mohammadi, H., Zare, A., Soltanolkotabi, M., and Jovanović, M. R. (2019). Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem. *arXiv preprint arXiv:1912.11899*.

Pananjady, A. and Wainwright, M. J. (2020). Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585.

Peters, J. and Schaal, S. (2008). Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shani, L., Efroni, Y., and Mannor, S. (2019). Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. *arXiv preprint arXiv:1909.02769*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Sutton, R. S. (1984). Temporal credit assignment in reinforcement learning. *PhD thesis, University of Massachusetts*.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063.

Tu, S. and Recht, B. (2019). The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory*, pages 3036–3083.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. (2019). Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Wu, Y., Zhang, W., Xu, P., and Gu, Q. (2020). A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350*.

Xie, Q., Yang, Z., Wang, Z., and Minca, A. (2020). Provable fictitious play for general mean-field games. *arXiv preprint arXiv:2010.04211*.

Xu, T., Liang, Y., and Lan, G. (2020a). A primal approach to constrained policy optimization: Global optimality and finite-time analysis. *arXiv preprint arXiv:2011.05869*.

Xu, T., Wang, Z., and Liang, Y. (2020b). Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*.

Yang, W., Li, X., Xie, G., and Zhang, Z. (2020). Finding the near optimal policy via adaptive reduced regularization in mdps. *arXiv preprint arXiv:2011.00213*.

Zhang, J., Kim, J., O'Donoghue, B., and Boyd, S. (2020a). Sample efficient reinforcement learning with REINFORCE. *arXiv preprint arXiv:2010.11364*.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. (2020b). Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33.

Zhang, K., Hu, B., and Basar, T. (2019). Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. *arXiv preprint arXiv:1910.09496*.

Zhang, K., Koppel, A., Zhu, H., and Basar, T. (2020c). Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612.

Zhang, X., Chen, Y., Zhu, X., and Sun, W. (2021). Robust policy gradient against strong data corruption. *arXiv preprint arXiv:2102.05800*.

Zhao, Y., Tian, Y., Lee, J., and Du, S. (2021). Provably efficient policy gradient methods for two-player zero-sum Markov games. *arXiv preprint arXiv:2102.08903*.