On the Global Convergence of Risk-Averse Policy Gradient Methods with Expected Conditional Risk Measures

Xian Yu [1] **Lei Ying** [2]

Abstract

Risk-sensitive reinforcement learning (RL) has become a popular tool to control the risk of uncertain outcomes and ensure reliable performance in various sequential decision-making problems. While policy gradient methods have been developed for risk-sensitive RL, it remains unclear if these methods enjoy the same global convergence guarantees as in the risk-neutral case (Bhandari & Russo, 2019; Mei et al., 2020; Agarwal et al., 2021; Cen et al., 2022). In this paper, we consider a class of dynamic time-consistent risk measures, called Expected Conditional Risk Measures (ECRMs), and derive policy gradient updates for ECRM-based objective functions. Under both constrained direct parameterization and unconstrained softmax parameterization, we provide global convergence and iteration complexities of the corresponding risk-averse policy gradient algorithms. We further test risk-averse variants of REINFORCE (Williams, 1992) and actor-critic algorithms (Konda & Tsitsiklis, 1999) to demonstrate the efficacy of our method and the importance of risk control.

1. Introduction

As reinforcement learning (RL) becomes a popular technique for solving Markov Decision Processes (MDPs) (Puterman, 2014), a stream of research has been devoted to managing *risk*. In risk-neutral RL, one seeks a policy that minimizes the expected total discounted cost. However, minimizing the expected cost does not necessarily avoid the rare occurrences of undesirably high cost, and in a situation where it is important to maintain reliable performance, we aim to evaluate and control the *risk*.

In particular, coherent risk measures (Artzner et al., 1999) have been used in many risk-sensitive RL research as they satisfy several natural and desirable properties. Among them, conditional value-at-risk (CVaR) (Rockafellar et al., 2000; Rockafellar & Uryasev, 2002; Ruszczyński & Shapiro, 2006; Shapiro et al., 2009) quantifies the amount of tail risk. When the risk is calculated in a nested way via dynamic risk measures, a desirable property is called *time consistency* (Ruszczyński, 2010), which ensures consistent risk preferences over time. Informally, it says that if a certain cost is considered less risky at stage k, then it should also be considered less risky at an earlier stage $l < k$. In this paper, we consider a class of dynamic risk measures, called expected conditional risk measures (ECRMs) (Homem-de Mello & Pagnoncelli, 2016), that are both coherent and time-consistent.

Broadly speaking, there are two classes of RL algorithms, value-based and policy-gradient-based methods. Policy gradient methods have captured a lot of attention as they are applicable to any differentiable policy parameterization and have been recently proved to have global convergence guarantees (Bhandari & Russo, 2019; Mei et al., 2020; Agarwal et al., 2021; Cen et al., 2022). While Tamar et al. (2015a) have developed policy gradient updates for both static coherent risk measures and time-consistent Markov coherent risk measures (MCR), they do not provide any discussions related to their global convergence. Recently, Huang et al. (2021) show that the MCR objectives (unlike the risk-neutral case) are not gradient dominated, and thus the stationary points that policy gradient methods find are not, in general, guaranteed to be globally optimal. To the best of our knowledge, it still remains an open question to develop policy gradient methods for RL with dynamic time-consistent risk measures that possess the same global convergence properties as in the risk-neutral case.

This step aims at answering this open question. We apply ECRMs on infinite-horizon MDPs and propose policy gradient updates for ECRMs-based objectives. Under both constrained direct parameterization and unconstrained softmax parameterization, we provide global convergence guarantees

[1]Department of Integrated Systems Engineering, The Ohio State University, Columbus, OH, USA [2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA. Correspondence to: Xian Yu <yu.3610@osu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

and iteration complexities for the corresponding risk-averse policy gradient methods, analogous to the risk-neutral case (Bhandari & Russo, 2019; Mei et al., 2020; Agarwal et al., 2021; Cen et al., 2022). Using the proposed policy gradient updates, any policy gradient algorithms can be tailored to solve risk-averse ECRM-based RL problems. Specifically, we apply a risk-averse variant of the REINFORCE algorithm (Williams, 1992) on a stochastic Cliffwalk environment (Sutton & Barto, 2018) and a risk-averse variant of the actor-critic algorithm (Konda & Tsitsiklis, 1999) on a Cartpole environment (Barto et al., 1983). Our numerical results show that the risk-averse algorithms enhance policy safety by choosing safer actions and reducing the cost variance, compared to the risk-neutral counterparts.

Table 1. Iteration complexity comparison between the risk-neutral results in Agarwal et al. (2021) and our risk-averse setting, where $\mathcal{S}$, $\mathcal{A}$ are the state and action space, $\mathcal{H}$ is the space of an auxiliary variable η, $\gamma \in (0,1)$ is the discount factor, ϵ is the optimality gap, $D_\infty = ||\frac{d_{s_1}^{\pi^*}}{\mu}||_\infty$ is used in Agarwal et al. (2021), and $D_1 = \max\{||\frac{\rho}{\mu}||_\infty, \frac{1}{(1-\gamma)\pi_1^{LB}}||\frac{d_{\rho\pi^*}^{\pi^*}}{\mu^P}||_\infty\}$, $D_2 = ||\frac{\rho}{\mu}||_\infty + \frac{1}{(1-\gamma)\pi_1^{LB}}||\frac{d_{\rho\pi^*}^{\pi^*}}{\mu^P}||_\infty$ are defined in Theorems 3.7 and 3.12 in our paper, respectively.

Iteration Complexity	Direct Parameter	Softmax Parameter
Agarwal et al. (2021) (Risk-neutral)	$O(\frac{D_\infty^2 \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^6\epsilon^2})$	$O(\frac{D_\infty^2 \lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2}{(1-\gamma)^6\epsilon^2})$
Our work (Risk-averse)	$O(\frac{D_1^2 \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\lvert\mathcal{H}\rvert^2}{(1-\gamma)^3\epsilon^2})$	$O(\frac{D_2^2 \lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2\lvert\mathcal{H}\rvert^4}{(1-\gamma)^3\epsilon^2})$

Related Work Risk-sensitive MDPs have been studied in several different settings, where the objectives are to maximize the worst-case outcome (Heger, 1994; Coraluppi & Marcus, 2000), to reduce variance (Howard & Matheson, 1972; Markowitz & Todd, 2000; Borkar, 2002; Tamar et al., 2012; La & Ghavamzadeh, 2013), to optimize a static risk measure (Chow & Ghavamzadeh, 2014; Tamar et al., 2015b; Yu et al., 2017) or to optimize a dynamic risk measure (Ruszczyński, 2010; Chow & Pavone, 2013; 2014; Köse & Ruszczyński, 2021; Yu & Shen, 2022).

Recently, Tamar et al. (2015a) derive policy gradient algorithms for both static coherent risk measures and dynamic MCR using the dual representation of coherent risk measures. Later, Huang et al. (2021) show that the dynamic MCR objective function is not gradient dominated and thus the corresponding policy gradient method does not have the same global convergence guarantees as it has for the risk-neutral case (Bhandari & Russo, 2019; Mei et al., 2020; Agarwal et al., 2021; Cen et al., 2022).

The major contributions of this paper are three-fold. First, we take the first step to answer an open question by providing global optimality guarantees for risk-averse policy gradient algorithms using a class of dynamic time-consistent risk measures – ECRMs, first introduced by Homem-de Mello & Pagnoncelli (2016). We would like to note that, although Yu & Shen (2022) have shown the ECRM-based risk-averse Bellman operator is a contraction mapping, it does not necessarily imply the global convergence of policy gradient algorithms for ECRM-based RL. Second, we derive iteration complexity bounds for the corresponding risk-averse policy gradient methods under both constrained direct parameterization and unconstrained softmax parameterization, which closely match the risk-neutral results in Agarwal et al. (2021) (see Table 1). Third, our method can be extended to any policy gradient algorithms, including actor-critic algorithms, for solving problems with continuous state and action space.

2. Preliminaries

We consider an infinite horizon discounted MDP: $M = (\mathcal{S}, \mathcal{A}, C, P, \gamma, \rho)$, where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, $C(s,a) \in [0,1]$ is a bounded, deterministic cost given state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, $P(\cdot|s,a)$ is the transition probability distribution, $\gamma \in (0,1)$ is the discount factor, and ρ is the initial state distribution over $\mathcal{S}$.

A stationary Markov policy $\pi^\theta : \mathcal{S} \to \Delta(\mathcal{A})$ parameterized by θ specifies a probability distribution over the action space given each state $s \in \mathcal{S}$, where $\Delta(\cdot)$ denotes the probability simplex, i.e., $0 \le \pi^\theta(a|s) \le 1$, $\sum_{a\in\mathcal{A}} \pi^\theta(a|s) = 1$, $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$. A policy induces a distribution over trajectories $\{(s_t, a_t, C(s_t,a_t))\}_{t=1}^\infty$, where s_1 is drawn from the initial state distribution ρ, and for all time steps t, $a_t \sim \pi^\theta(\cdot|s_t)$, $s_{t+1} \sim P(\cdot|s_t,a_t)$. The value function $V^{\pi^\theta} : \mathcal{S} \to \mathbb{R}$ is defined as the discounted sum of future costs starting at state s and executing π, i.e., $V^{\pi^\theta}(s) = \mathbb{E}[\sum_{t=1}^\infty \gamma^{t-1} C(s_t,a_t)|\pi^\theta, s_1 = s]$. We overload the notation and define $V^{\pi^\theta}(\rho)$ as the expected value under initial state distribution ρ, i.e., $V^{\pi^\theta}(\rho) = \mathbb{E}_{s_1\sim\rho}[V^{\pi_\theta}(s_1)]$. The action-value (or Q-value) function $Q^{\pi^\theta} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is defined as $Q^{\pi^\theta}(s,a) = \mathbb{E}[\sum_{t=1}^\infty \gamma^{t-1} C(s_t,a_t)|\pi^\theta, s_1 = s, a_1 = a]$.

In a risk-neutral RL framework, the goal of the agent is to find a policy π^θ that minimizes the expected total cost from the initial state, i.e., the agent seeks to solve $\min_{\theta\in\Theta} V^{\pi^\theta}(\rho)$ where $\{\pi^\theta|\theta \in \Theta\}$ is some class of parametric stochastic policies. The famous theorem of Bellman & Dreyfus (1959) shows that there exists a policy π^* that simultaneously minimizes $V^{\pi^\theta}(s_1)$ for all states $s_1 \in \mathcal{S}$. It is worth noting that $V^{\pi^\theta}(s)$ is non-convex in θ, so the standard tools from convex optimization literature are not applicable. We refer interested readers to Agarwal et al. (2021) for a non-convex example in Figure 1.

2.1. Policy Gradient Methods

Policy gradient algorithms have received lots of attention in the RL community due to their simple structure. The basic idea is to adjust the parameter θ of the policy in the gradient descent direction. Before introducing the policy gradient methods, we first define the discounted state visitation distribution $d_{s_1}^{\pi^\theta}$ of a policy π^θ as $d_{s_1}^{\pi^\theta}(s) = (1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}\text{Pr}^{\pi^\theta}(s_t = s|s_1)$, where $\text{Pr}^{\pi^\theta}(s_t = s|s_1)$ is the probability that $s_t = s$ after executing π starting from state s_1. Correspondingly, we define the discounted state visitation distribution under initial distribution ρ as $d_{\rho}^{\pi^\theta}(s) = \mathbb{E}_{s_1\sim\rho}[d_{s_1}^{\pi^\theta}(s)]$.

The fundamental result underlying policy gradient algorithms is the *policy gradient theorem* (Williams, 1992; Sutton et al., 1999), i.e., $\nabla_\theta V^{\pi^\theta}(s_1)$ takes the following form

$$\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_1}^{\pi^\theta}}\mathbb{E}_{a\sim\pi^\theta(\cdot|s)}[\nabla_\theta \log \pi^\theta(a|s) Q^{\pi^\theta}(s,a)],$$

where the policy gradient is surprisingly simple and does not depend on the gradient of the state distribution.

Recently, Bhandari & Russo (2019); Mei et al. (2020); Agarwal et al. (2021); Cen et al. (2022) demonstrate the global optimality and convergence rate of policy gradient methods in a risk-neutral setting. This paper aims to extend the results to risk-averse objective functions with dynamic time-consistent risk measures. Next, we first define coherent one-step conditional risk measures.

2.2. Coherent One-Step Conditional Risk Measures

Consider a probability space $(\Xi, \mathcal{F}, P)$, and let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ be sub-sigma-algebras of $\mathcal{F}$ such that each $\mathcal{F}_t$ corresponds to the information available up to (and including) stage t, with $Z_t,\ t = 1, 2\ldots$ being an adapted sequence of random variables. In this paper, we interpret the variables Z_t as immediate costs. We assume that $\mathcal{F}_1 = \{\emptyset, \Xi\}$, and thus Z_1 is in fact deterministic. Let $\mathcal{Z}_t$ denote a space of $\mathcal{F}_t$-measurable functions from Ξ to $\mathbb{R}$.

Definition 2.1. (Artzner et al., 1999) A conditional risk measure $\rho : \mathcal{Z}_{k+1} \to \mathcal{Z}_k$ is coherent if it satisfies the following four properties: (i) [**Monotonicity**] If $Z_1, Z_2 \in \mathcal{Z}_{k+1}$ and $Z_1 \geq Z_2$, then $\rho(Z_1) \geq \rho(Z_2)$; (ii) [**Convexity**] $\rho(\gamma Z_1 + (1-\gamma)Z_2) \leq \gamma\rho(Z_1) + (1-\gamma)\rho(Z_2)$ for all $Z_1,\ Z_2 \in \mathcal{Z}_{k+1}$ and all $\gamma \in [0, 1]$; (iii) [**Translation invariance**] If $W \in \mathcal{Z}_k$ and $Z \in \mathcal{Z}_{k+1}$, then $\rho(Z + W) = \rho(Z) + W$ and (iv) [**Positive Homogeneity**] If $\gamma \geq 0$ and $Z \in \mathcal{Z}_{k+1}$, then $\rho(\gamma Z) = \gamma\rho(Z)$.

For ease of presentation, we rewrite $C(s_t, a_t)$ as c_t for all $t \geq 1$ and denote vector $(a_1, \ldots, a_t)$ as $a_{[1,t]}$ in the rest of this paper. For our problem, we consider a special class of coherent one-step conditional risk measures $\rho_t^{s_{[1,t-1]}}$ mapping from $\mathcal{Z}_t$ to $\mathcal{Z}_{t-1}$, which is a convex combination of conditional expectation and Conditional Value-at-Risk (CVaR):

$$\rho_t^{s_{[1,t-1]}}(c_t) = (1-\lambda)\mathbb{E}[c_t|s_{[1,t-1]}] + \lambda\text{CVaR}_\alpha[c_t|s_{[1,t-1]}], \tag{1}$$

where $\lambda \in [0, 1]$ is a weight parameter to balance the expected cost and tail risk, and $\alpha \in (0, 1)$ represents the confidence level. Notice that this risk measure is more general than CVaR and expectation because it has CVaR or expectation as a special case when $\lambda = 1$ or $\lambda = 0$.

Following the results by Rockafellar & Uryasev (2002), the upper α-tail CVaR can be expressed as the optimization problem below:

$$\text{CVaR}_\alpha[c_t|s_{[1,t-1]}] := \min_{\eta_t\in\mathbb{R}}\left\{\eta_t + \frac{1}{\alpha}\mathbb{E}[[c_t - \eta_t]_+|s_{[1,t-1]}]\right\}, \tag{2}$$

where $[a]_+ := \max\{a, 0\}$, and η_t is an auxiliary variable. The minimum of the right-hand side of the above definition is attained at $\eta_t^* = \text{VaR}_\alpha[c_t|s_{[1,t-1]}] := \inf\{v : \mathbb{P}(c_t \leq v) \geq 1-\alpha\}$, and thus CVaR is the mean of the upper α-tail distribution of c_t, i.e., $\mathbb{E}[c_t|c_t > \eta_t^*]$. Please see Figure 1 for an illustration of the CVaR measure. Selecting a small α value makes CVaR sensitive to rare but very high costs. Because $c_t \in [0, 1]$, we can restrict the η-variable to be within $[0, 1]$, i.e., $\eta_t \in \mathcal{H} = [0, 1]$ for all $t \geq 1$.

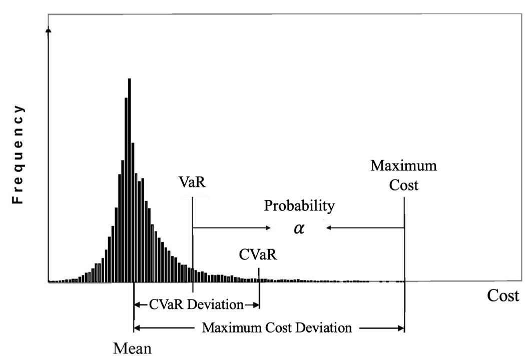


Figure 1. Illustration of CVaR.

2.3. Expected Conditional Risk Measures

We consider a class of multi-period risk function $\mathbb{F}$ mapping from $\mathcal{Z}_{1,\infty} := \mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_t \times \cdots$ to $\mathbb{R}$ as follows:

$$\begin{aligned}\mathbb{F}(c_{[1,\infty]}|s_1) = & c_1 + \gamma\rho_2^{s_1}(c_2) \\ & + \lim_{T\to\infty}\sum_{t=3}^{T}\gamma^{t-1}\mathbb{E}_{s_{[1,t-1]}}\left[\rho_t^{s_{[1,t-1]}}(c_t)\right],\end{aligned} \tag{3}$$

where $\rho_t^{s_{[1,t-1]}}$ is the coherent one-step conditional risk measure mapping from $\mathcal{Z}_t$ to $\mathcal{Z}_{t-1}$ defined in Eq. (1) to represent the risk given the information available up to (including) stage $t-1$, and the expectation is taken with respect to the random history $s_{[1,t-1]}$. This class of multi-period risk measures is called expected conditional risk measures (ECRMs), first introduced in (Homem-de Mello & Pagnoncelli, 2016).

Using the specific risk measure defined in (1) and (2) and applying tower property of expectations on (3), we have

$$\begin{aligned}
&\min_{a_{[1,\infty]}} \mathbb{F}(c_{[1,\infty]}|s_1) \\
&= \min_{a_1,\eta_2} \Big\{ C(s_1,a_1) + \gamma\lambda\eta_2 + \gamma\mathbb{E}^{s_1}_{s_2}\Big[\min_{a_2,\eta_3} \Big\{ \frac{\lambda}{\alpha}[C(s_2,a_2)-\eta_2]_+ \\
&+ (1-\lambda)C(s_2,a_2) + \gamma\lambda\eta_3 + \gamma\mathbb{E}^{s_2}_{s_3}\Big[\min_{a_3,\eta_4} \Big\{ \frac{\lambda}{\alpha}[C(s_3,a_3)-\eta_3]_+ \\
&+ (1-\lambda)C(s_3,a_3) + \gamma\lambda\eta_4 + \cdots \Big\}\Big]\Big\}\Big]\Big\},
\end{aligned} \tag{4}$$

where $\mathbb{E}^{s_{t-1}}_{s_t} = \mathbb{E}_{s_t}[\cdot|s_{t-1}]$ is the conditional expectation and we apply the Markov property to recast $\mathbb{E}^{s_{[1,t-1]}}_{s_t}$ as $\mathbb{E}^{s_{t-1}}_{s_t}$. The auxiliary variable η_t from Eq. (2) is decided before taking conditional expectation $\mathbb{E}^{s_{t-1}}_{s_t}$ and thus it can be regarded as a $(t-1)$-stage action, similar to a_{t-1}. Here, η_t denotes the tail information of state s_t's immediate cost (i.e., $\eta^*_t = \mathrm{VaR}_\alpha[C(s_t,a_t)|s_{[1,t-1]}]$), which helps us take risk into account when making decisions. We refer interested readers to Yu & Shen (2022) for discussions on the time-consistency of ECRMs and contraction property of the corresponding risk-averse Bellman equation.

Based on formulation (4), we observe that the key differences between (4) and risk-neutral RL are (i) the augmentation of the action space $a_t \in \mathcal{A}$ to be $(a_t,\eta_{t+1}) \in \mathcal{A}\times\mathcal{H}$ for all time steps $t \geq 1$ to help learn the tail information of cost distribution, and (ii) the manipulation on immediate costs, i.e., replacing $C(s_1,a_1)$ with $\bar{C}_1(s_1,a_1,\eta_2) = C(s_1,a_1) + \gamma\lambda\eta_2$ and replacing $C(s_t,a_t)$ with $\bar{C}_t(s_t,\eta_t,a_t,\eta_{t+1}) = \frac{\lambda}{\alpha}[C(s_t,a_t)-\eta_t]_+ + (1-\lambda)C(s_t,a_t) + \gamma\lambda\eta_{t+1}$ for $t \geq 2$. Note that for time steps $t \geq 2$, the calculations of immediate costs $\bar{C}_t(s_t,\eta_t,a_t,\eta_{t+1})$ involve both the action η_t from the previous time step $t-1$ and η_{t+1} from the current time step t. As a result, we augment the state space $s_t \in \mathcal{S}$ to be $(s_t,\eta_t) \in \mathcal{S}\times\mathcal{H}$ for all $t \geq 2$, where η_t is the previous action taken in time step $t-1$. To discretize the space $\mathcal{H} = [0,1]$, we assume that η_t has $H+1$ possible values in total and the h-th element of the $\mathcal{H}$-space is $\frac{h}{H}$ for all $h = 0,1,\ldots,H$. This leads to the following proposition.

Proposition 2.2. *Define the augmented action space as $\bar{\mathcal{A}} = \mathcal{A}\times\mathcal{H}$, augmented state space as $\bar{\mathcal{S}} = \mathcal{S}\times\mathcal{H}$, and the state-action transition matrix under policy π as $\bar{P}_{(s_t,\eta_t,a_t,\eta_{t+1})\to(s'_t,\eta'_t,a'_t,\eta'_{t+1})} = P(s'_t|s_t,a_t)\pi(a'_t,\eta'_{t+1}|s'_t,\eta'_t)$, if $\eta'_t = \eta_{t+1}$; otherwise, $\bar{P}_{(s_t,\eta_t,a_t,\eta_{t+1})\to(s'_t,\eta'_t,a'_t,\eta'_{t+1})} = 0$. Then the risk-averse RL with ECRM-based objective function (4) is equivalent to a risk-neutral RL with $\bar{M} = (\bar{\mathcal{S}},\bar{\mathcal{A}},\bar{C},\bar{P},\gamma,\rho)$.*

The proof of Proposition 2.2 is straightforward and omitted here. Note that although the risk-averse RL with ECRM can be reformulated as a risk-neutral RL, the modified immediate cost $\bar{C}_1(s_1,a_1,\eta_2)$ in the first time step has a different form than $\bar{C}_t(s_t,\eta_t,a_t,\eta_{t+1})$ in other time steps $t \geq 2$. Due to this, the conventional Bellman equation used in risk-neutral RL is not applicable here, thereby preventing us from directly employing the results of the risk-neutral policy gradient algorithms.

3. Global Optimality and Convergence of Risk-Averse Policy Gradient Methods

According to formulation (4), we should distinguish the value functions and policies for ECRMs-based objectives between the first time step and others because of the differences in the immediate costs. Furthermore, starting from time step 2, problem (4) reduces to a risk-neutral RL with the same form of manipulated costs $\bar{C}_t(s_t,\eta_t,a_t,\eta_{t+1})$ for time steps $t \geq 2$, and according to (Puterman, 2014), there exists a deterministic stationary Markov optimal policy. As a result, we consider a class of policies $\pi^\theta = (\pi_1^{\theta_1},\pi_2^{\theta_2}) \in \Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}$ where $\pi_1^{\theta_1}(a_1,\eta_2|s_1)$ is the policy for the first time step parameterized by θ_1 and $\pi_2^{\theta_2}(a_t,\eta_{t+1}|s_t,\eta_t)$, $\forall t \geq 2$ is the stationary policy for the following time steps parameterized by θ_2. We omit the dependence of π on θ in the following for ease of presentation. The goal is to solve the optimization problem below

$$\min_{\pi\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}} J^\pi(\rho) \tag{5}$$

where we denote π^* as the optimal policy and $J^*(\rho)$ as the optimal objective value. The value and action-value functions for the first time step are defined as

$$\begin{aligned}
J^\pi(\rho) =& \mathbb{E}_{s_1\sim\rho}[J^\pi(s_1)] \\
=& \mathbb{E}_{s_1\sim\rho}\mathbb{E}_{(a_1,\eta_2)\sim\pi_1(\cdot,\cdot|s_1)}[Q^{\pi_2}(s_1,a_1,\eta_2)]
\end{aligned} \tag{6}$$

and

$$\begin{aligned}
&Q^{\pi_2}(s_1,a_1,\eta_2) = C(s_1,a_1) + \gamma\lambda\eta_2 \\
&+ \gamma\mathbb{E}^{s_1,a_1}_{s_2}\mathbb{E}_{(a_2,\eta_3)\sim\pi_2(\cdot,\cdot|s_2,\eta_2)}\Big[\Big\{\frac{\lambda}{\alpha}[C(s_2,a_2)-\eta_2]_+ \\
&+ (1-\lambda)C(s_2,a_2) + \gamma\lambda\eta_3 \\
&+ \gamma\mathbb{E}^{s_2,a_2}_{s_3}\mathbb{E}_{(a_3,\eta_4)\sim\pi_2(\cdot,\cdot|s_3,\eta_3)}\Big[\Big\{\frac{\lambda}{\alpha}[C(s_3,a_3)-\eta_3]_+ \\
&+ (1-\lambda)C(s_3,a_3) + \gamma\lambda\eta_4 + \cdots\Big\}\Big]\Big\}\Big],
\end{aligned}$$

respectively. Correspondingly, we define value functions for time steps $t \geq 2$ with state (s_t,η_t) as

$$\hat{J}^{\pi_2}(s_t,\eta_t) = \mathbb{E}_{(a_t,\eta_{t+1})\sim\pi_2(\cdot,\cdot|s_t,\eta_t)}[\hat{Q}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})]$$

where

$$\hat{Q}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1}) = \frac{\lambda}{\alpha}[C(s_t,a_t)-\eta_t]_+$$

$$+ (1-\lambda)C(s_t, a_t) + \gamma\lambda\eta_{t+1}$$
$$+ \gamma\mathbb{E}_{s_{t+1}}^{s_t,a_t}\mathbb{E}_{(a_{t+1},\eta_{t+2})\sim\pi_2}\Big[\Big\{\frac{\lambda}{\alpha}[C(s_{t+1}, a_{t+1}) - \eta_{t+1}]_+$$
$$+ (1-\lambda)C(s_{t+1}, a_{t+1}) + \gamma\lambda\eta_{t+2} + \cdots\Big\}\Big]$$
$$= \frac{\lambda}{\alpha}[C(s_t, a_t) - \eta_t]_+ + (1-\lambda)C(s_t, a_t) + \gamma\lambda\eta_{t+1}$$
$$+ \gamma\mathbb{E}_{s_{t+1}}^{s_t,a_t}[\hat{J}^{\pi_2}(s_{t+1}, \eta_{t+1})]. \tag{7}$$

As a result, we have the following equation:

$$Q^{\pi_2}(s_1, a_1, \eta_2) = C(s_1, a_1) + \gamma\lambda\eta_2 + \gamma\mathbb{E}_{s_2}^{s_1,a_1}[\hat{J}^{\pi_2}(s_2, \eta_2)]. \tag{8}$$

We also define advantage functions $A^\pi : \mathcal{S}\times\mathcal{A}\times\mathcal{H}\to\mathbb{R}$ for the first time step and $\hat{A}^{\pi_2} : \mathcal{S}\times\mathcal{H}\times\mathcal{A}\times\mathcal{H}\to\mathbb{R}$ for time steps $t \geq 2$ as

$$A^\pi(s_1, a_1, \eta_2) = J^\pi(s_1) - Q^{\pi_2}(s_1, a_1, \eta_2)$$
$$\hat{A}^{\pi_2}(s_t, \eta_t, a_t, \eta_{t+1}) = \hat{J}^{\pi_2}(s_t, \eta_t) - \hat{Q}^{\pi_2}(s_t, \eta_t, a_t, \eta_{t+1}),$$

respectively. Note that because of the differences between the first time step and others, we cannot directly apply the risk-neutral policy gradient updates and their convergence results. Instead, we need to derive risk-averse policy gradients, i.e., $\nabla_\theta J^\pi(\rho) = (\nabla_{\theta_1}J^\pi(\rho), \nabla_{\theta_2}J^\pi(\rho))$, and use this joint gradient vector to provide convergence guarantees.

Next, we first derive a performance difference lemma in Lemma 3.1 and the policy gradients for ECRM-based objective functions in Theorem 3.2, which are applicable to any parameterizations. The proofs are presented in Appendix A.

Lemma 3.1. *Let $Pr^\pi(\tau|s_1 = s)$ denote the probability of observing a trajectory τ when starting in state s and following policy π. For all policies π, π' and states s_1,*

$$J^\pi(s_1) - J^{\pi'}(s_1) =$$
$$\mathbb{E}_{\tau\sim Pr^{\pi'}(\tau|s_1)}\Big[A^\pi(s_1, a_1, \eta_2) + \sum_{t=2}^{\infty}\gamma^{t-1}\hat{A}^{\pi_2}(s_t, \eta_t, a_t, \eta_{t+1})\Big].$$

Let $\Pr^\pi(s_{t+1} = s, \eta_{t+1} = \eta|s_t, \eta_t) = \sum_{a_t}\pi_2(a_t, \eta|s_t, \eta_t)P(s|s_t, a_t)$ denote the probability that $s_{t+1} = s, \eta_{t+1} = \eta$ when starting in state s_t, η_t and following policy π and let $d^\pi_{s_2,\eta_2}(s,\eta) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\Pr^\pi(s_{t+2} = s, \eta_{t+2} = \eta|s_2, \eta_2)$ denote the discounted state visitation distribution starting from state s_2, η_2. We present the policy gradients of ECRMs-based objective function (5) in the next theorem.

Theorem 3.2. *The policy gradients of (5) take the following forms*

$$\nabla_{\theta_1}J^\pi(\rho) = \mathbb{E}_{s_1\sim\rho}\mathbb{E}_{(a_1,\eta_2)\sim\pi_1(\cdot|s_1)}[$$
$$\nabla_{\theta_1}\log\pi_1(a_1, \eta_2|s_1)Q^{\pi_2}(s_1, a_1, \eta_2)]$$
$$\nabla_{\theta_2}J^\pi(\rho) = \frac{\gamma}{1-\gamma}\mathbb{E}_{(s_t,\eta_t)\sim d^\pi_{\rho^\pi}}\mathbb{E}_{(a_t,\eta_{t+1})\sim\pi_2(\cdot|s_t,\eta_t)}[$$
$$\nabla_{\theta_2}\log\pi_2(a_t, \eta_{t+1}|s_t, \eta_t)\hat{Q}^{\pi_2}(s_t, \eta_t, a_t, \eta_{t+1})]$$

where $\rho^\pi(s,\eta) = \sum_{s_1}\rho(s_1)Pr^{\pi_1}(s_2 = s, \eta_2 = \eta|s_1)$.

The differences between risk-averse ECRM-based and risk-neutral policy gradients are twofold. First, we break the policy parameters into two parts, θ_1 and θ_2, and derive the gradient for each one separately. Second, as reflected in the state visitation distribution $d^\pi_{\rho^\pi}$ in $\nabla_{\theta_2}J^\pi(\rho)$, the initial state becomes (s_2, η_2) with initial distribution $\rho^\pi(s,\eta)$, different from the risk-neutral gradients where the initial state is s_1 with distribution ρ and the state visitation distribution is d^π_ρ.

Next, we consider two types of parameterizations: (i) constrained direct parameterization in Section 3.1 and (ii) unconstrained softmax parameterization in Section 3.2. Both parameterizations are complete in the sense that any stochastic policy can be represented in the class, and for each of them, we provide global convergence of the risk-averse policy gradient methods with iteration complexities.

3.1. Constrained Direct Parameterization

For direct parameterization, the policies are $\pi_1(a_1, \eta_2|s_1) = \theta_1(s_1, a_1, \eta_2)$ and $\pi_2(a_t, \eta_{t+1}|s_t, \eta_t) = \theta_2(s_t, \eta_t, a_t, \eta_{t+1})$, $\forall t \geq 2$, where $\theta_1 \in \Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|}$ and $\theta_2 \in \Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}||\mathcal{H}|}$. In this section, we may write $\nabla_\pi J^\pi(\rho)$ instead of $\nabla_\theta J^{\pi^\theta}(\rho)$, and the gradients are

$$\frac{\partial J^\pi(\rho)}{\partial\pi_1(a,\eta|s)} = \rho(s)Q^{\pi_2}(s, a, \eta) \tag{9}$$
$$\frac{\partial J^\pi(\rho)}{\partial\pi_2(a,\eta'|s,\eta)} = \frac{\gamma}{1-\gamma}d^\pi_{\rho^\pi}(s,\eta)\hat{Q}^{\pi_2}(s, \eta, a, \eta') \tag{10}$$

using Theorem 3.2. Next, we show that the objective function $J^\pi(s_1)$ is smooth. From standard optimization results (see Appendix E), for a smooth function, a small gradient descent update will guarantee to improve the objective value. The omitted proofs of this section are provided in Appendix B.

Lemma 3.3. *For all starting states s_1, $J^\pi(s_1)$ is $\frac{2\gamma|\mathcal{A}||\mathcal{H}|}{(1-\gamma)^3}||\bar{c}||_\infty$-smooth in π, i.e.,*

$$||\nabla_\pi J^\pi(s_1) - \nabla_\pi J^{\pi'}(s_1)||_2 \leq \frac{2\gamma|\mathcal{A}||\mathcal{H}|}{(1-\gamma)^3}||\bar{c}||_\infty||\pi - \pi'||_2,$$

where $||\bar{c}||_\infty = \frac{\lambda}{\alpha} + (1-\lambda) + \gamma\lambda$.

However, smoothness alone can only guarantee the convergence of the gradient descent method to a stationary point (i.e., $\nabla_\pi J^\pi(s_1) = 0$). For non-convex objective functions, in order to ensure convergence to global minima, we need to establish that the gradient of the objective at any parameter dominates the sub-optimality of the parameter, such as Polyak-like gradient domination conditions (Polyak, 1963). We give a formal definition of gradient domination below.

Definition 3.4. (Bhandari & Russo, 2019) We say f is (ν, μ)-gradient dominated over Θ if there exists constants $\nu > 0$ and $\mu \geq 0$ such that for all $\theta \in \Theta$,

$$\min_{\theta' \in \Theta} f(\theta') \geq f(\theta) + \min_{\theta' \in \Theta} \left[\nu \langle \nabla f(\theta), \theta' - \theta \rangle + \frac{\mu}{2} ||\theta - \theta'||_2^2 \right].$$

The function is said to be gradient dominated with degree one if $\mu = 0$ and with degree two if $\mu > 0$.

Any stationary point of a gradient-dominated function is globally optimal. To see this, we note that for any stationary point θ, we have $\langle \nabla f(\theta), \theta' - \theta \rangle \geq 0$ for all $\theta' \in \Theta$. Then the minimizer of the right-hand side in Definition 3.4 is θ, implying $\min_{\theta' \in \Theta} f(\theta') \geq f(\theta)$.

In the next theorem, we show that the value function $J^\pi(\rho)$ is gradient dominated with degree one, which will be used to quantify the convergence rate of projected gradient descent methods in Theorem 3.7 later. Following Agarwal et al. (2021), even though we are interested in the value $J^\pi(\rho)$, we will consider the gradient with respect to another state distribution $\mu \in \Delta(\mathcal{S})$, which allows for greater flexibility in our analysis.

Theorem 3.5. *Let $\pi_1^{LB} = \min_{s,a,\eta} \pi_1(a, \eta|s)$. For the direct policy parameterization, for all state distributions $\rho, \ \mu \in \Delta(\mathcal{S})$, we have*

$$J^\pi(\rho) - J^*(\rho) \leq D_1 \max_{\bar{\pi} \in \Delta(\mathcal{A} \times \mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}} (\pi - \bar{\pi})^\intercal \nabla_\pi J^\pi(\mu)$$

where $D_1 = \max\{||\frac{\rho}{\mu}||_\infty, \frac{1}{(1-\gamma)\pi_1^{LB}} ||\frac{d_{\rho \pi^*}^{\pi^*}}{\mu^P}||_\infty\}$ and $\mu^P(s, \eta) = \sum_{s_1, a_1} \mu(s_1) P(s|s_1, a_1), \forall s \in \mathcal{S}, \ \eta \in \mathcal{H}$.

Note that the significance of Theorem 3.5 is that although the gradient is with respect to $J^\pi(\mu)$, the final guarantee applies to all distributions ρ.

Remark 3.6. Compared to Lemma 4 in Agarwal et al. (2021):

$$V^\pi(\rho) - V^*(\rho) \leq \frac{1}{1-\gamma} ||\frac{d_\rho^{\pi^*}}{\mu}||_\infty \max_{\bar{\pi}} (\pi - \bar{\pi})^\intercal \nabla_\pi V^\pi(\mu),$$

in risk-neutral policy gradient methods, some conditions on the state distribution of π, or equivalently μ, are necessary for stationarity to imply optimality as illustrated in Section 4.3 of Agarwal et al. (2021). For example, if the starting state distribution is not strictly positive, i.e., $\mu(s_1) = 0$ for some state s_1, then the coefficient $||\frac{d_\rho^{\pi^*}}{\mu}||_\infty = +\infty$.

Similarly, in risk-averse policy gradient methods, according to our Theorem 3.5, we not only need a strictly positive distribution μ for the starting state s_1, but also need a strictly positive distribution μ^π for second-step states s_2, η_2. To achieve this, we need to ensure that (i) each possible value of η_2 is achieved with a positive probability (i.e., $\pi_1^{LB} > 0$), and (ii) each possible value of s_2 is reachable from the initial distribution μ (i.e., $\mu^P > 0$).

With the policy gradient results in Theorem 3.2, we consider a *projected gradient descent* method, where we directly update the policy parameter in the gradient descent direction and then project it back onto the simplex if the constraints are violated after a gradient update. The projected gradient descent algorithm updates

$$\pi^{(t+1)} = P_{\Delta(\mathcal{A} \times \mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}} (\pi^{(t)} - \beta \nabla_\pi J^{(t)}(\mu))$$

where $P_{\Delta(\mathcal{A} \times \mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}}$ is the projection onto $\Delta(\mathcal{A} \times \mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}$ in the Euclidean norm, and β is the step size. Using Theorem 3.5, we now give an iteration complexity bound for projected gradient descent methods.

Theorem 3.7. *Let $\pi_1^{LB} = \inf_{t \geq 1} \min_{s,a,\eta} \pi_1^{(t)}(a, \eta|s)$, $D_1 = \max\{||\frac{\rho}{\mu}||_\infty, \frac{1}{(1-\gamma)\pi_1^{LB}} ||\frac{d_{\rho \pi^*}^{\pi^*}}{\mu^P}||_\infty\}$. The projected gradient descent algorithm on $J^\pi(\mu)$ with stepsize $\beta = \frac{(1-\gamma)^3}{2\gamma|\mathcal{A}||\mathcal{H}|||\bar{c}||_\infty}$ satisfies for all distributions $\rho \in \Delta(\mathcal{S})$,*

$$\min_{t \leq T} J^{(t)}(\rho) - J^*(\rho) \leq \epsilon$$

whenever

$$T \geq D_1^2 \frac{128\gamma|\mathcal{S}||\mathcal{A}||\mathcal{H}|^2}{(1-\gamma)^3 \epsilon^2} \bar{C} ||\bar{c}||_\infty$$

with $\bar{C} = (\frac{\lambda}{\alpha} + \lambda + \frac{1}{1-\gamma} ||\bar{c}||_\infty), \ ||\bar{c}||_\infty = \frac{\lambda}{\alpha} + (1-\lambda) + \gamma\lambda$.

A proof is provided in Appendix B, where we invoke a standard iteration complexity result of projected gradient descent on smooth functions to show that the gradient magnitude with respect to all feasible directions is small. Then, we use Theorem 3.5 to complete the proof. Note that the guarantee we provide is for the best policy found over T rounds, which is standard in the non-convex optimization literature. As can be seen from Theorems 3.5 and 3.7, when $\pi_1^{LB} \to 0$, the iteration bound $T \to +\infty$. To circumvent this issue, we consider a softmax parameterization with log barrier regularizer in the next section, which ensures that $\pi_1^{LB} > 0$.

3.2. Unconstrained Softmax Parameterization

In this section, we aim to solve the optimization problem (5) with the following softmax parameterization: for all s_1, a_1, η_2 and $s_t, \eta_t, a_t, \eta_{t+1}, \ t \geq 2$, we have

$$\pi_1^\theta(a_1, \eta_2|s_1) = \frac{\exp(\theta_1(s_1, a_1, \eta_2))}{\sum_{a_1', \eta_2'} \exp(\theta_1(s_1, a_1', \eta_2'))},$$

$$\pi_2^\theta(a_t, \eta_{t+1}|s_t, \eta_t) = \frac{\exp(\theta_2(s_t, \eta_t, a_t, \eta_{t+1}))}{\sum_{a_t', \eta_{t+1}'} \exp(\theta_2(s_t, \eta_t, a_t', \eta_{t+1}'))}.$$

Note that the softmax parameterization is preferable to the direct parameterization, since the parameters θ are unconstrained (i.e., $\pi_1^\theta, \pi_2^\theta$ belong to the probability simplex automatically) and standard unconstrained optimization algorithms can be employed. The omitted proofs of this section are provided in Appendix C.

Lemma 3.8. *Using the softmax parameterization, the gradients take the following forms:*

$$\frac{\partial J^{\pi^\theta}(\mu)}{\partial \theta_1(s_1, a_1, \eta_2)} = \mu(s_1)\pi_1^\theta(a_1, \eta_2|s_1)(-A^{\pi^\theta}(s_1, a_1, \eta_2))$$

$$\frac{\partial J^{\pi^\theta}(\mu)}{\partial \theta_2(s_t, \eta_t, a_t, \eta_{t+1})} = \frac{\gamma}{1-\gamma} d_{\mu^\pi}^\pi(s_t, \eta_t)\pi_2^\theta(a_t, \eta_{t+1}|s_t, \eta_t) (-\hat{A}^{\pi_2^\theta}(s_t, \eta_t, a_t, \eta_{t+1})).$$

Because the action probabilities depend exponentially on the parameters θ, policies can quickly become near deterministic and lack exploration, leading to slow convergence. In this section, we add an entropy-based regularization term to the objective function to keep the probabilities from getting too small. Recall that the relative-entropy for distributions p and q is defined as $\mathrm{KL}(p, q) := \mathbb{E}_{x\sim p}[-\log q(x)/p(x)]$. Denote the uniform distribution over set $\mathcal{X}$ as $\mathrm{Unif}_\mathcal{X}$, and consider the following log barrier regularized objective:

$$\begin{aligned} L_\kappa(\theta) :=& J^{\pi^\theta}(\mu) + \kappa\mathbb{E}_{s_1\sim \mathrm{Unif}_\mathcal{S}}[\mathrm{KL}(\mathrm{Unif}_{\mathcal{A}\times\mathcal{H}}, \pi_1^\theta(\cdot,\cdot|s_1))] \\ &+ \kappa\mathbb{E}_{s_t,\eta_t\sim \mathrm{Unif}_{\mathcal{S}\times\mathcal{H}}}[\mathrm{KL}(\mathrm{Unif}_{\mathcal{A}\times\mathcal{H}}, \pi_2^\theta(\cdot,\cdot|s_t,\eta_t))] \\ =& J^{\pi^\theta}(\mu) - \frac{\kappa}{|\mathcal{S}||\mathcal{A}||\mathcal{H}|}\sum_{s_1,a_1,\eta_2}\log \pi_1^\theta(a_1,\eta_2|s_1) \\ &- \frac{\kappa}{|\mathcal{S}||\mathcal{H}||\mathcal{A}||\mathcal{H}|}\sum_{s_t,\eta_t,a_t,\eta_{t+1}}\log \pi_2^\theta(a_t,\eta_{t+1}|s_t,\eta_t) \\ &- 2\kappa\log|\mathcal{A}||\mathcal{H}|, \end{aligned}$$

where $\kappa > 0$ is a regularization parameter and the last (constant) term is not relevant to optimization. We first show that $L_\kappa(\theta)$ is smooth in the next lemma.

Lemma 3.9. *The log barrier regularized objective function $L_\kappa(\theta)$ is σ_κ-smooth with $\sigma_\kappa = 6(\frac{\lambda}{\alpha} + \lambda) + \frac{8}{(1-\gamma)^3}||\bar{c}||_\infty + \frac{2\kappa}{|\mathcal{S}|} + \frac{2\kappa}{|\mathcal{S}||\mathcal{H}|}$.*

Our next theorem shows that the approximate first-order stationary points of $L_\kappa(\theta)$ are approximately globally optimal with respect to $J^\pi(\rho)$, as long as the regularization parameter κ is small enough.

Theorem 3.10. *Let $\pi_1^{LB} = \min_{s,a,\eta}\pi_1^\theta(a,\eta|s)$. Suppose*

$$||\nabla_{\theta_1}L_\kappa(\theta)||_2 \le \epsilon_1 \le \frac{\kappa}{2|\mathcal{S}||\mathcal{A}||\mathcal{H}|},$$
$$||\nabla_{\theta_2}L_\kappa(\theta)||_2 \le \epsilon_2 \le \frac{\kappa}{2|\mathcal{S}||\mathcal{H}||\mathcal{A}||\mathcal{H}|},$$

then we have that for all starting state distributions ρ:

$$J^{\pi^\theta}(\rho) - J^*(\rho) \le 2\kappa||\frac{\rho}{\mu}||_\infty + \frac{2\kappa}{(1-\gamma)\pi_1^{LB}}||\frac{d_{\rho^{\pi^*}}^{\pi^*}}{\mu^P}||_\infty.$$

Now consider policy gradient descent updates for $L_\kappa(\theta)$ as follows:

$$\theta^{(t+1)} := \theta^{(t)} - \beta\nabla_\theta L_\kappa(\theta^{(t)}). \tag{11}$$

Lemma 3.11. *Using the policy gradient updates (11) for $L_\kappa(\theta)$ with $\beta = \frac{1}{\sigma_\kappa}$, we have $\pi_1^{LB} := \inf_{t\ge1}\min_{s,a,\eta}\pi_1^{\theta^{(t)}}(a,\eta|s) > 0$ and $\pi_2^{LB} := \inf_{t\ge1}\min_{s,a,\eta,\eta'}\pi_2^{\theta^{(t)}}(a,\eta'|s,\eta) > 0$.*

Using Theorem 3.10, Lemma 3.11 and standard results on the convergence of gradient descent, we obtain the following iteration complexity for softmax parameterization with log barrier regularization.

Theorem 3.12. *Let $\sigma_\kappa = 6(\frac{\lambda}{\alpha} + \lambda) + \frac{8}{(1-\gamma)^3}||\bar{c}||_\infty + \frac{2\kappa}{|\mathcal{S}|} + \frac{2\kappa}{|\mathcal{S}||\mathcal{H}|}$ and $D_2 = ||\frac{\rho}{\mu}||_\infty + \frac{1}{(1-\gamma)\pi_1^{LB}}||\frac{d_{\rho^{\pi^*}}^{\pi^*}}{\mu^P}||_\infty$. Starting from any initial finite-value $\theta^{(0)}$, consider the updates (11) with $\kappa = \frac{\epsilon}{2D_2}$ and $\beta = \frac{1}{\sigma_\kappa}$. Then for all starting state distributions ρ, we have*

$$\min_{t\le T} J^{(t)}(\rho) - J^*(\rho) \le \epsilon$$

whenever

$$T \ge \frac{64(3(\frac{\lambda}{\alpha}+\lambda) + 4||\bar{c}||_\infty + 2)\bar{B}|\mathcal{S}|^2|\mathcal{A}|^2|\mathcal{H}|^4}{(1-\gamma)^3\epsilon^2}D_2^2$$

with $\bar{B} = \bar{C} - \log\pi_1^{LB} - \log\pi_2^{LB}$.

4. Numerical Results

In this section, we propose a risk-averse REINFORCE algorithm for handling discrete state and action space in Section 4.1 and a risk-averse actor-critic algorithm for handling continuous state and action space in Section 4.2, respectively.

4.1. Algorithms for Discrete State and Action Space

We first propose a risk-averse REINFORCE algorithm (Williams, 1992) in Algorithm 1 with softmax parameterization, which works for discrete state and action space.

We implement Algorithm 1 on a stochastic version of the 4×4 Cliffwalk environment (Sutton & Barto, 2018) (see Figure 2), where the agent needs to travel from a start ($[3,0]$) to a goal state ($[3,3]$) while incurring as little cost as possible. Each action incurs 1 cost, but the shortest path lies next to a cliff ($[3,1]$ and $[3,2]$), where entering the cliff corresponds to a cost of 5 and the agent will return to the start. Because the classical Cliffwalk environment is deterministic, we consider a stochastic variant following (Huang et al., 2021), where the row of cells above the cliff ($[2,1]$ and $[2,2]$) are slippery, and entering these slippery cells induces a transition into the cliff with probability $p = 0.1$. The discount

Algorithm 1 Risk-averse REINFORCE with softmax parameterization

Initialize $\theta_1(s_1, a_1, \eta_2)$, $\forall s_1 \in \mathcal{S}$, $a_1 \in \mathcal{A}$, $\eta_2 \in \mathcal{H}$ and set softmax policy $\pi_1^\theta(a_1, \eta_2|s_1) = \frac{\exp(\theta_1(s_1,a_1,\eta_2))}{\sum_{a_1',\eta_2'} \exp(\theta_1(s_1,a_1',\eta_2'))}$.

Initialize $\theta_2(s_t, \eta_t, a_t, \eta_{t+1})$, $\forall s_t \in \mathcal{S}$, $a_t \in \mathcal{A}$, η_t, $\eta_{t+1} \in \mathcal{H}$ and set softmax policy $\pi_2^\theta(a_t, \eta_{t+1}|s_t, \eta_t) = \frac{\exp(\theta_2(s_t,\eta_t,a_t,\eta_{t+1}))}{\sum_{a_t',\eta_{t+1}'} \exp(\theta_2(s_t,\eta_t,a_t',\eta_{t+1}'))}$.

while not converged **do**

  Generate one trajectory on policy $\pi^\theta = (\pi_1^\theta, \pi_2^\theta)$: $s_1, a_1, \eta_2, c_1, s_2, \ldots, s_{T-1}, a_{T-1}, \eta_T, c_{T-1}, s_T$.

  Modify immediate costs as $\bar{c}_1 = c_1 + \gamma\lambda\eta_2$, $\bar{c}_t = \frac{\lambda}{\alpha}[c_t - \eta_t]_+ + (1-\lambda)c_t + \gamma\lambda\eta_{t+1}$, $\forall t \geq 2$.

  Update $\theta_1 := \theta_1 - \beta\nabla_{\theta_1} \log \pi_1^\theta(a_1, \eta_2|s_1) \sum_{\tau=1}^{T-1} \gamma^{\tau-1}\bar{c}_\tau$.

  Update $\theta_2 := \theta_2 - \beta \sum_{t=2}^{T-1} \nabla_{\theta_2} \log \pi_2^\theta(a_t, \eta_{t+1}|s_t, \eta_t) \sum_{\tau=t}^{T-1} \gamma^{\tau-t}\bar{c}_\tau$.

end while

factor γ is set to 0.98, and the confidence level for CVaR is set to $\alpha = 0.05$. Because each immediate cost can only take values 1 or 5, we set the η-space as $\mathcal{H} = \{1, 5\}$. For this stochastic environment, the shortest path is the blue path in Figure 2, which takes a cost of 5 if not entering the cliff; however, a safer path is the orange path, which induces a deterministic cost of 7.

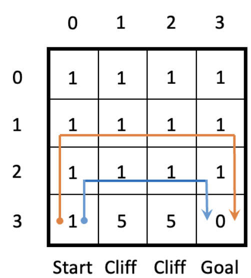
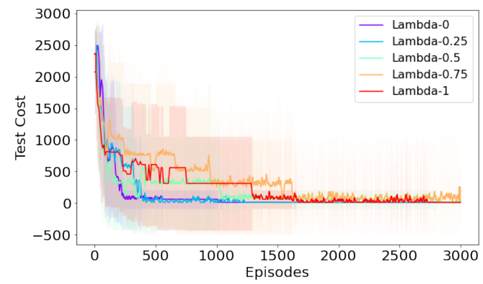


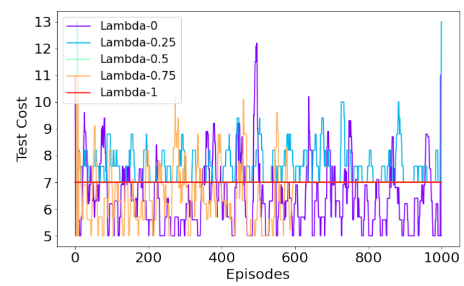
Figure 2. A 4×4 stochastic Cliffwalk environment.

We train the model over 10000 episodes where each episode starts at a random state and continues until the goal state is reached or the maximum time step (i.e., 500) is reached. All the hyperparameters are kept the same across different trials, except for the risk parameters λ which is swept across 0, 0.25, 0.5, 0.75, 1. For each of these values, we perform the task over 10 independent simulation runs. The average test costs (when we start at state $[3, 0]$ and choose the action having the largest probability) in the first 3000 episodes over the 10 simulation runs are recorded in Figure 3(a), where the colored shades represent the standard deviation. To present a detailed description of these cost profiles, we also look at a specific run and plot the test cost in the last 1000 episodes in Figure 3(b).

From Figure 3(a), $\lambda = 1$ produces the largest variance of test cost in the first 1500 episodes, after which the policy



(a) Average test cost over 10 runs with varying λ.



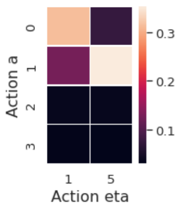
(b) Test cost in the last 1000 episodes in one run with varying λ.

Figure 3. Test cost over 10 independent runs with varying λ.

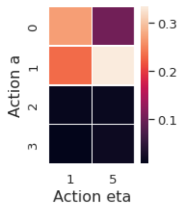
gradually converges to the safer path with a steady cost of 7. When we look at Figure 3(b), $\lambda = 0$ converges to the shortest path with the highest variance, and $\lambda = 0.25$ chooses a combination of the shortest and safer paths (i.e., move right at state $[2, 0]$, move up at state $[2, 1]$, and then follow the safer path). On the other hand, $\lambda = 0.5, 0.75, 1$ all converge to the same safer path with a steady cost of 7 in the last 400 episodes, so they overlap in the figure.

Next, we present the average action probabilities ($\pi_2^\theta(\cdot, \cdot|s_t, \eta_t)$ in the last episode) at state $s_t = [2, 0]$, $\eta_t = 0$ over the 10 independent runs with varying λ in Figure 4, where actions $a = 0, 1, 2, 3$ represent moving up, right, down and left respectively, and actions $\eta = 1, 5$ represent the VaR of the next state's cost distribution. A lighter color denotes a higher probability. From Figure 4, when $\lambda = 0, 0.25$, the learned optimal action at state $s = [2, 0]$ is to move right ($a = 1$) with $\eta = 5$, as entering state $[2, 1]$ will induce a cost of 5 with probability 0.1. As λ increases, the optimal action shifts to move up ($a = 0$) with $\eta = 1$. We present the average action probabilities at each state in the learned optimal path in Figures 7–11 in Appendix F.

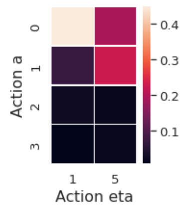
Lastly, we display the impact of modifying the regularizer parameter κ from 0 to 0.5 in Figure 5. A higher κ helps speed up the convergence to a steady path while $\kappa = 0$ generates the highest variance of test cost after 2000 episodes.



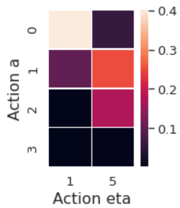
(a) $\lambda = 0$



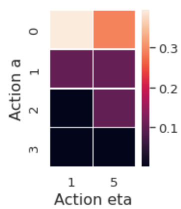
(b) $\lambda = 0.25$



(c) $\lambda = 0.5$



(d) $\lambda = 0.75$



(e) $\lambda = 1$

Figure 4. Average action probabilities at state $s_t = [2, 0],\ \eta_t = 0$.

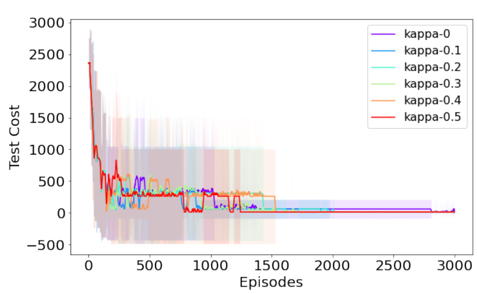


Figure 5. Average test cost over 10 runs with varying κ.

4.2. Algorithms for Continuous State and Action Space

Our method can be also extended to solve problems with continuous state and action space. To design a risk-averse actor-critic algorithm, we replace the tabular policy π_1 and π_2 in Algorithm 1 with neural networks parameterized by θ_1 and θ_2, respectively. We also construct neural networks for value critics $v_1(s_1)$ with parameter w_1 and value critic $v_2(s_t, \eta_t)$ with parameter w_2. The main steps for the risk-averse actor-critic are outlined in Algorithm 2. We apply Algorithm 2 on the CartPole environment (Barto et al., 1983), where the reward is $+1$ for every step taken and the goal is to keep the pole upright for as long as possible. The average test rewards over 5 simulation runs are displayed in Figure 6, where we vary the risk parameter λ from 0 to 1. From Figure 6, $\lambda = 1$ outperforms all other configurations by producing the highest reward over time.

Algorithm 2 Risk-Averse Actor-Critic

Initialize policy $\pi_1(a_1, \eta_2 | s_1)$ with parameter θ_1 and policy $\pi_2(a_t, \eta_{t+1} | s_t, \eta_t)$ with parameter θ_2.
Initialize value critic $v_1(s_1)$ with parameter w_1 and value critic $v_2(s_t, \eta_t)$ with parameter w_2.
while not converged **do**
  Generate one trajectory on policy $\pi^\theta = (\pi_1^{\theta_1}, \pi_2^{\theta_2})$: $s_1, a_1, \eta_2, c_1, s_2, \ldots, s_{T-1}, a_{T-1}, \eta_T, c_{T-1}, s_T$.
  Modify immediate costs as $\bar{c}_1 = c_1 + \gamma\lambda\eta_2$, $\bar{c}_t = \frac{\lambda}{\alpha}[c_t - \eta_t]_+ + (1-\lambda)c_t + \gamma\lambda\eta_{t+1}$, $\forall t \geq 2$.
  Compute discounted costs: $V_t = \sum_{\tau=t}^{T-1} \gamma^{\tau-t} \bar{c}_\tau$ for all $t = 1, \ldots, T-1$.
  Update w_1 to minimize $||v_1^{w_1}(s_1) - V_1||^2$ and update w_2 to minimize $\sum_{t=2}^{T-1} ||v_2^{w_2}(s_t, \eta_t) - V_t||^2$.
  Update $\theta_1 := \theta_1 - \beta \nabla_{\theta_1} \log \pi_1^{\theta_1}(a_1, \eta_2 | s_1) V_1$.
  Update $\theta_2 := \theta_2 - \beta \sum_{t=2}^{T-1} \nabla_{\theta_2} \log \pi_2^{\theta_2}(a_t, \eta_{t+1} | s_t, \eta_t) V_t$.
end while

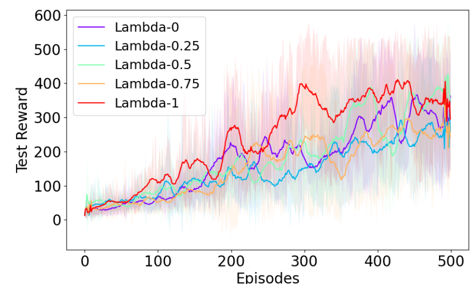


Figure 6. Average test reward over 5 runs with varying λ.

5. Conclusions

In this paper, we applied a class of dynamic time-consistent coherent risk measures (i.e., ECRMs) on infinite-horizon MDPs and provided global convergence guarantees for risk-averse policy gradient methods under constrained direct parameterization and unconstrained softmax parameterization. Our iteration complexity results closely matched the risk-neutral counterparts in (Agarwal et al., 2021).

For future research, it is worth investigating iteration complexities for policy gradient algorithms with restricted policy classes (e.g., log-linear policy and neural policy) and natural policy gradient. It would also be interesting to incorporate distributional RL (Bellemare et al., 2017) into this risk-sensitive setting and derive global convergence guarantees.

Acknowledgements

The authors would like to thank three anonymous reviewers and program chairs for providing helpful feedback on this manuscript. The work of Lei Ying is supported in part by NSF under grants 2112471, 2207548, and 2228974.

References

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

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Barto, A. G., Sutton, R. S., and Anderson, C. W. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13(5):834–846, 1983.

Beck, A. *First-order methods in optimization*. SIAM, 2017.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458. PMLR, 2017.

Bellman, R. and Dreyfus, S. Functional approximations and dynamic programming. *Mathematical Tables and Other Aids to Computation*, pp. 247–251, 1959.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Borkar, V. S. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27(2):294–311, 2002.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4): 2563–2578, 2022.

Chow, Y. and Ghavamzadeh, M. Algorithms for CVaR optimization in MDPs. *Advances in Neural Information Processing Systems*, 27, 2014.

Chow, Y.-L. and Pavone, M. Stochastic optimal control with dynamic, time-consistent risk constraints. In *2013 American Control Conference*, pp. 390–395. IEEE, 2013.

Chow, Y.-L. and Pavone, M. A framework for time-consistent, risk-averse model predictive control: Theory and algorithms. In *2014 American Control Conference*, pp. 4204–4211. IEEE, 2014.

Coraluppi, S. P. and Marcus, S. I. Mixed risk-neutral/minimax control of discrete-time, finite-state markov decision processes. *IEEE Transactions on Automatic Control*, 45(3):528–532, 2000.

Ghadimi, S. and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.

Heger, M. Consideration of risk in reinforcement learning. In *Machine Learning Proceedings 1994*, pp. 105–111. Elsevier, 1994.

Homem-de Mello, T. and Pagnoncelli, B. K. Risk aversion in multistage stochastic programming: A modeling and algorithmic perspective. *European Journal of Operational Research*, 249(1):188–199, 2016.

Howard, R. A. and Matheson, J. E. Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369, 1972.

Huang, A., Leqi, L., Lipton, Z. C., and Azizzadenesheli, K. On the convergence and optimality of policy gradient for markov coherent risk. *arXiv preprint arXiv:2103.02827*, 2021.

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Köse, Ü. and Ruszczyński, A. Risk-averse learning by temporal difference methods with markov risk measures. *The Journal of Machine Learning Research*, 22(1):1800–1833, 2021.

La, P. and Ghavamzadeh, M. Actor-critic algorithms for risk-sensitive MDPs. *Advances in neural information processing systems*, 26, 2013.

Markowitz, H. M. and Todd, G. P. *Mean-variance analysis in portfolio choice and capital markets*, volume 66. John Wiley & Sons, 2000.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020.

Polyak, B. T. Gradient methods for minimizing functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):643–653, 1963.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Rockafellar, R. T. and Uryasev, S. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

Rockafellar, R. T., Uryasev, S., et al. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–42, 2000.

Ruszczyński, A. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, 125(2):235–261, 2010.

Ruszczyński, A. and Shapiro, A. Optimization of convex risk functions. *Mathematics of Operations Research*, 31 (3):433–452, 2006.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12, 1999.

Tamar, A., Di Castro, D., and Mannor, S. Policy gradients with variance related risk criteria. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1651–1658, 2012.

Tamar, A., Chow, Y., Ghavamzadeh, M., and Mannor, S. Policy gradient for coherent risk measures. *Advances in Neural Information Processing Systems*, 28, 2015a.

Tamar, A., Glassner, Y., and Mannor, S. Optimizing the CVaR via sampling. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015b.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

Yu, P., Haskell, W. B., and Xu, H. Dynamic programming for risk-aware sequential optimization. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 4934–4939. IEEE, 2017.

Yu, X. and Shen, S. Risk-averse reinforcement learning via dynamic time-consistent risk measures. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 2307–2312, 2022. doi: 10.1109/CDC51059.2022.9992450.

Appendix

The appendix is organized as follows.

- Appendix A: proofs of Lemma 3.1 and Theorem 3.2.
- Appendix B: proofs in Section 3.1.
- Appendix C: proofs in Section 3.2.
- Appendix D: smoothness proofs.
- Appendix E: standard optimization results.
- Appendix F: additional computational results.

A. Proofs of Lemma 3.1 and Theorem 3.2

Proof of Lemma 3.1. [Performance difference lemma] Using a telescoping argument, we have

$$\begin{aligned}
J^{\pi}(s_1) - J^{\pi'}(s_1) =& J^{\pi}(s_1) - \mathbb{E}_{\tau\sim Pr^{\pi'}}[\sum_{t=1}^{\infty}\gamma^{t-1}\bar{c}_t] \\
=& J^{\pi}(s_1) - \mathbb{E}_{\tau\sim Pr^{\pi'}}[\sum_{t=1}^{\infty}\gamma^{t-1}(\bar{c}_t + J^{\pi}(s_t) - J^{\pi}(s_t)] \\
\overset{(a)}{=}& - \mathbb{E}_{\tau\sim Pr^{\pi'}}[\sum_{t=1}^{\infty}\gamma^{t-1}(\bar{c}_t + \gamma J^{\pi}(s_{t+1}) - J^{\pi}(s_t)] \\
\overset{(b)}{=}& - \mathbb{E}_{\tau\sim Pr^{\pi'}}[\sum_{t=1}^{\infty}\gamma^{t-1}(\bar{c}_t + \gamma \mathbb{E}_{s_{t+1}}^{s_t,a_t}[J^{\pi}(s_{t+1})] - J^{\pi}(s_t))],
\end{aligned}$$

where (a) rearranges terms in the summation and cancels the $J^{\pi}(s_1)$ term with $J^{\pi}(s_1)$ outside the summation, and (b) uses the tower property of conditional expectations. For the term $t = 1$ inside the summation, we have

$$\begin{aligned}
& - \mathbb{E}_{(a_1,\eta_2)\sim\pi_1'(s_1)}[C(s_1,a_1) + \gamma\lambda\eta_2 + \gamma\mathbb{E}_{s_2}^{s_1,a_1}[J^{\pi}(s_2)] - J^{\pi}(s_1)] \\
=& - \mathbb{E}_{(a_1,\eta_2)\sim\pi_1'(s_1)}[Q^{\pi_2}(s_1,a_1,\eta_2) + \gamma\mathbb{E}_{s_2}^{s_1,a_1}[J^{\pi}(s_2) - \hat{J}^{\pi_2}(s_2,\eta_2)] - J^{\pi}(s_1)] \\
=& \mathbb{E}_{(a_1,\eta_2)\sim\pi_1'(s_1)}[A^{\pi}(s_1,a_1,\eta_2) + \gamma\mathbb{E}_{s_2}^{s_1,a_1}[\hat{J}^{\pi_2}(s_2,\eta_2) - J^{\pi}(s_2)]]
\end{aligned} \tag{12}$$

where the first equality is because of Eq. (8). For terms $t \geq 2$, we have

$$\begin{aligned}
& - \mathbb{E}_{\tau\sim Pr^{\pi'}}\Big[\gamma^{t-1}\Big(\frac{\lambda}{\alpha}[C(s_t,a_t) - \eta_t]_+ + (1-\lambda)C(s_t,a_t) + \gamma\lambda\eta_{t+1} + \gamma\mathbb{E}_{s_{t+1}}^{s_t,a_t}[J^{\pi}(s_{t+1})] - J^{\pi}(s_t)\Big)\Big] \\
=& - \mathbb{E}_{\tau\sim Pr^{\pi'}}\Big[\gamma^{t-1}\Big(\hat{Q}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1}) - \hat{J}^{\pi_2}(s_t,\eta_t) + \gamma\mathbb{E}_{s_{t+1}}^{s_t,a_t}[J^{\pi}(s_{t+1}) - \hat{J}^{\pi_2}(s_{t+1},\eta_{t+1})] + (\hat{J}^{\pi_2}(s_t,\eta_t) - J^{\pi}(s_t))\Big)\Big] \\
=& \mathbb{E}_{\tau\sim Pr^{\pi'}}\Big[\gamma^{t-1}\Big(\hat{A}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1}) + \gamma\mathbb{E}_{s_{t+1}}^{s_t,a_t}[\hat{J}^{\pi_2}(s_{t+1},\eta_{t+1}) - J^{\pi}(s_{t+1})] - (\hat{J}^{\pi_2}(s_t,\eta_t) - J^{\pi}(s_t))\Big)\Big].
\end{aligned} \tag{13}$$

Observing that $\mathbb{E}_{s_t,a_t,\eta_{t+1}\sim Pr^{\pi'}}\mathbb{E}_{s_{t+1}}^{s_t,a_t}[\hat{J}^{\pi_2}(s_{t+1},\eta_{t+1}) - J^{\pi}(s_{t+1})] = \mathbb{E}_{s_{t+1},\eta_{t+1}\sim Pr^{\pi'}}[\hat{J}^{\pi_2}(s_{t+1},\eta_{t+1}) - J^{\pi}(s_{t+1})]$ for all $t \geq 1$, the second term in Eq. (12) cancels out the third term in Eq. (13) with $t = 2$. Moreover, the second term in Eq. (13) with time step t cancels out the third term in Eq. (13) with time step $t + 1$ for all $t \geq 2$. As a result, we have

$$J^{\pi}(s_1) - J^{\pi'}(s_1) = \mathbb{E}_{\tau\sim \Pr^{\pi'}(\tau|s_1)}\Big[A^{\pi}(s_1,a_1,\eta_2) + \sum_{t=2}^{\infty}\gamma^{t-1}\hat{A}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})\Big].$$

This completes the proof. $\square$

Proof of Theorem 3.2. [Risk-averse policy gradients] According to Eq. (6), we have

$$\begin{aligned}\nabla_{\theta_1} J^{\pi}(s_1) &= \nabla_{\theta_1}\Big(\sum_{a_1,\eta_2} \pi_1(a_1,\eta_2|s_1) Q^{\pi_2}(s_1,a_1,\eta_2)\Big)\\ &\overset{(a)}{=} \sum_{a_1,\eta_2} \nabla_{\theta_1}\pi_1(a_1,\eta_2|s_1) Q^{\pi_2}(s_1,a_1,\eta_2)\\ &= \mathbb{E}_{(a_1,\eta_2)\sim\pi_1}[\nabla_{\theta_1}\log\pi_1(a_1,\eta_2|s_1) Q^{\pi_2}(s_1,a_1,\eta_2)]\end{aligned}$$

where (a) is true because $\nabla_{\theta_1} Q^{\pi_2}(s_1,a_1,\eta_2)=0$. As a result,

$$\nabla_{\theta_1} J^{\pi}(\rho) = \nabla_{\theta_1}\mathbb{E}_{s_1\sim\rho}[J^{\pi}(s_1)] = \mathbb{E}_{s_1\sim\rho}[\nabla_{\theta_1} J^{\pi}(s_1)].$$

Based on the definition of $Q^{\pi_2}(s_1,a_1,\eta_2)$, we have

$$\begin{aligned}\nabla_{\theta_2} J^{\pi}(s_1) =&\nabla_{\theta_2}\Big(\sum_{a_1,\eta_2}\pi_1(a_1,\eta_2|s_1)Q^{\pi_2}(s_1,a_1,\eta_2)\Big)\\ =&\sum_{a_1,\eta_2}(\pi_1(a_1,\eta_2|s_1)\nabla_{\theta_2}(c_1+\gamma\lambda\eta_2+\gamma\mathbb{E}_{s_2}^{s_1}[\hat{J}^{\pi_2}(s_2,\eta_2)]))\\ =&\gamma\sum_{a_1,\eta_2}\pi_1(a_1,\eta_2|s_1)\sum_{s_2}P(s_2|s_1,a_1)\nabla_{\theta_2}\hat{J}^{\pi_2}(s_2,\eta_2)\\ =&\gamma\sum_{s_2,\eta_2}\mathrm{Pr}^{\pi_1}(s_2,\eta_2|s_1)\nabla_{\theta_2}\hat{J}^{\pi_2}(s_2,\eta_2)\end{aligned}$$

Now for $\nabla_{\theta_2}\hat{J}^{\pi_2}(s_2,\eta_2)$, we have

$$\begin{aligned}&\nabla_{\theta_2}\hat{J}^{\pi_2}(s_2,\eta_2)\\ =&\nabla_{\theta_2}\Big(\sum_{a_2,\eta_3}\pi_2(a_2,\eta_3|s_2,\eta_2)\hat{Q}^{\pi_2}(s_2,\eta_2,a_2,\eta_3)\Big)\\ =&\sum_{a_2,\eta_3}\Big(\nabla_{\theta_2}\pi_2(a_2,\eta_3|s_2,\eta_2)\hat{Q}^{\pi_2}(s_2,\eta_2,a_2,\eta_3)+\pi_2(a_2,\eta_3|s_2,\eta_2)\nabla_{\theta_2}\hat{Q}^{\pi_2}(s_2,\eta_2,a_2,\eta_3)\Big)\\ =&\sum_{a_2,\eta_3}(\nabla_{\theta_2}\pi_2(a_2,\eta_3|s_2,\eta_2)\hat{Q}^{\pi_2}(s_2,\eta_2,a_2,\eta_3))+\gamma\sum_{a_2,\eta_3}(\pi_2(a_2,\eta_3|s_2,\eta_2)\sum_{s_3}P(s_3|s_2,a_2)\nabla_{\theta_2}\hat{J}^{\pi_2}(s_3,\eta_3)),\end{aligned}$$

where the last equation follows from the definition of $\hat{Q}^{\pi_2}(s_2,\eta_2,a_2,\eta_3))$ in Eq. (7). Using a similar argument in risk-neutral policy gradient theorems (Williams, 1992; Sutton et al., 1999) and denoting $\phi(s_t,\eta_t,a_t,\eta_{t+1}) := \nabla_{\theta_2}\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)\hat{Q}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})$, we obtain

$$\begin{aligned}&\nabla_{\theta_2}\hat{J}^{\pi_2}(s_2,\eta_2)\\ =&\sum_{a_2,\eta_3}\phi(s_2,\eta_2,a_2,\eta_3)+\gamma\sum_{s_3,\eta_3}\mathrm{Pr}^{\pi_2}(s_3,\eta_3|s_2,\eta_2)\sum_{a_3,\eta_4}\phi(s_3,\eta_3,a_3,\eta_4)\\ &+\gamma^2\sum_{s_4,\eta_4}\mathrm{Pr}^{\pi_2}(s_4,\eta_4|s_2,\eta_2)\sum_{a_4,\eta_5}\phi(s_4,\eta_4,a_4,\eta_5)+\cdots\\ =&\frac{1}{1-\gamma}\mathbb{E}_{(s_t,\eta_t)\sim d^{\pi}_{s_2,\eta_2}}\mathbb{E}_{(a_t,\eta_{t+1})\sim\pi_2(\cdot|s_t,\eta_t)}[\nabla_{\theta_2}\log\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)\hat{Q}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})]\end{aligned}$$

As a result,

$$\begin{aligned}&\nabla_{\theta_2}J^{\pi}(\rho)\\ =&\gamma\sum_{s_2,\eta_2}\sum_{s_1}\rho(s_1)\mathrm{Pr}^{\pi_1}(s_2,\eta_2|s_1)\nabla_{\theta_2}\hat{J}^{\pi_2}(s_2,\eta_2)\\ =&\gamma\sum_{s_2,\eta_2}\rho^{\pi}(s_2,\eta_2)\nabla_{\theta_2}\hat{J}^{\pi_2}(s_2,\eta_2)\\ =&\frac{\gamma}{1-\gamma}\mathbb{E}_{(s_t,\eta_t)\sim d^{\pi}_{\rho^{\pi}}}\mathbb{E}_{(a_t,\eta_{t+1})\sim\pi_2(\cdot|s_t,\eta_t)}[\nabla_{\theta_2}\log\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)\hat{Q}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})].\end{aligned}$$

This completes the proof. □

B. Proofs in Section 3.1

Proof of Lemma 3.3. [Smoothness for direct parameterization] Consider a unit vector u and let $\pi^\alpha := \pi^{\theta+\alpha u}$, $\tilde{J}(\alpha) := J^{\pi^\alpha}(s_1)$. For direct parameterization, we have $\pi^\alpha = \pi^{\theta+\alpha u} = \theta + \alpha u$. Differentiating π^α with respect to α gives

$$\sum_{a_1,\eta_2}\left|\frac{d\pi_1^\alpha(a_1,\eta_2|s_1)}{d\alpha}\right| \le \sum_{a_1,\eta_2}|u_1(s_1,a_1,\eta_2)| \le \sqrt{\sum_{a_1,\eta_2}|u_1(s_1,a_1,\eta_2)|^2}\sqrt{\sum_{a_1,\eta_2}1^2} \le ||u||_2\sqrt{|\mathcal{A}||\mathcal{H}|} = \sqrt{|\mathcal{A}||\mathcal{H}|}$$

$$\sum_{a_t,\eta_{t+1}}\left|\frac{d\pi_2^\alpha(a_t,\eta_{t+1}|s_t,\eta_t)}{d\alpha}\right| \le \sum_{a_t,\eta_{t+1}}|u_2(s_t,\eta_t,a_t,\eta_{t+1})| \le \sqrt{\sum_{a_t,\eta_{t+1}}|u_2(s_t,\eta_t,a_t,\eta_{t+1})|^2}\sqrt{\sum_{a_t,\eta_{t+1}}1^2} \le \sqrt{|\mathcal{A}||\mathcal{H}|}$$

$$\sum_{a_1,\eta_2}\left|\frac{d^2\pi_1^\alpha(a_1,\eta_2|s_1)}{(d\alpha)^2}\right| = 0,\ \sum_{a_t,\eta_{t+1}}\left|\frac{d^2\pi_2^\alpha(a_t,\eta_{t+1}|s_t,\eta_t)}{(d\alpha)^2}\right| = 0.$$

Using Lemma D.1 with $C_1 = \sqrt{|\mathcal{A}||\mathcal{H}|}$ and $C_2 = 0$, we get

$$\begin{aligned}\max_{||u||_2=1}\left|\frac{d^2\tilde{J}(\alpha)}{(d\alpha)^2}|_{\alpha=0}\right| &\le C_2\left(\frac{\lambda}{\alpha}+\lambda+\frac{||\bar{c}||_\infty}{(1-\gamma)^2}\right)+\frac{2\gamma C_1^2}{(1-\gamma)^3}||\bar{c}||_\infty \\ &\le \frac{2\gamma|\mathcal{A}||\mathcal{H}|}{(1-\gamma)^3}||\bar{c}||_\infty\end{aligned}$$

Thus $J^\pi(s_1)$ is $\frac{2\gamma|\mathcal{A}||\mathcal{H}|}{(1-\gamma)^3}||\bar{c}||_\infty$-smooth. This completes the proof. □

Proof of Theorem 3.5. [Gradient domination] According to Lemma 3.1, we have

$$J^\pi(\rho) - J^*(\rho) = \mathbb{E}_{\substack{s_1\sim\rho \\ \tau\sim\Pr^{\pi^*}(\tau|s_1)}}\Big[A^\pi(s_1,a_1,\eta_2) + \sum_{t=2}^\infty \gamma^{t-1}\hat{A}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})\Big].$$

Then for the first term, we have

$$\begin{aligned}&\mathbb{E}_{s_1\sim\rho}\mathbb{E}_{(a_1,\eta_2)\sim\pi_1^*(\cdot|s_1)}[A^\pi(s_1,a_1,\eta_2)] \\ =&\sum_{s_1,a_1,\eta_2}\rho(s_1)\pi_1^*(a_1,\eta_2|s_1)A^\pi(s_1,a_1,\eta_2) \\ \le&\sum_{s_1}\rho(s_1)\max_{a_1,\eta_2}A^\pi(s_1,a_1,\eta_2) \\ =&\sum_{s_1}\frac{\rho(s_1)}{\mu(s_1)}\mu(s_1)\max_{a_1,\eta_2}A^\pi(s_1,a_1,\eta_2) \\ \overset{(a)}{\le}&||\frac{\rho}{\mu}||_\infty\sum_{s_1}\mu(s_1)\max_{a_1,\eta_2}A^\pi(s_1,a_1,\eta_2) \\ \overset{(b)}{=}&||\frac{\rho}{\mu}||_\infty\max_{\bar{\pi}_1\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|}}\sum_{s_1,a_1,\eta_2}\mu(s_1)\bar{\pi}_1(a_1,\eta_2|s_1)A^\pi(s_1,a_1,\eta_2) \\ \overset{(c)}{=}&||\frac{\rho}{\mu}||_\infty\max_{\bar{\pi}_1\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|}}\sum_{s_1,a_1,\eta_2}\mu(s_1)(\bar{\pi}_1(a_1,\eta_2|s_1)-\pi_1(a_1,\eta_2|s_1))A^\pi(s_1,a_1,\eta_2) \\ \overset{(d)}{=}&||\frac{\rho}{\mu}||_\infty\max_{\bar{\pi}_1\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|}}\sum_{s_1,a_1,\eta_2}\mu(s_1)(\pi_1(a_1,\eta_2|s_1)-\bar{\pi}_1(a_1,\eta_2|s_1))Q^{\pi_2}(s_1,a_1,\eta_2) \\ \overset{(e)}{=}&||\frac{\rho}{\mu}||_\infty\max_{\bar{\pi}_1\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|}}(\pi_1-\bar{\pi}_1)^\mathsf{T}\nabla_{\pi_1}J^\pi(\mu)\end{aligned}$$

where (a) is true because $\max_{a_1,\eta_2}A^\pi(s_1,a_1,\eta_2) = J^\pi(s_1) - \min_{a_1,\eta_2}Q^{\pi_2}(s_1,a_1,\eta_2) \ge 0$; (b) is true because $\max_{\bar{\pi}_1}$ is attained at an action (a_1,η_2) that maximizes $A^\pi(s_1,\cdot,\cdot)$ for each state s_1; (c) follows since

$\sum_{a_1,\eta_2} \pi_1(a_1,\eta_2|s_1)A^\pi(s_1,a_1,\eta_2) = J^\pi(s_1) - \sum_{a_1,\eta_2} \pi_1(a_1,\eta_2|s_1)Q^{\pi_2}(s_1,a_1,\eta_2) = 0$; (d) follows since $\sum_{a_1,\eta_2}(\bar\pi_1(a_1,\eta_2|s_1) - \pi_1(a_1,\eta_2|s_1))J^\pi(s_1) = 0$; and (e) follows from Theorem 3.2 and Eq. (9).

For the second term, we have

$$
\begin{aligned}
&\mathbb{E}_{s_1\sim\rho}\mathbb{E}_{\tau\sim Pr^{\pi^*}}[\sum_{t=2}^\infty \gamma^{t-1}(\hat A^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})]\\
=&\gamma\mathbb{E}_{\substack{s_1\sim\rho\\(a_1,\eta_2)\sim\pi_1^*(s_1)\\ s_2\sim P(\cdot|s_1,a_1)\\ \tau\sim Pr^{\pi^*}}}[\sum_{t=0}^\infty \gamma^t \hat A^{\pi_2}(s_{t+2},\eta_{t+2},a_{t+2},\eta_{t+3})]\\
=&\frac{\gamma}{1-\gamma}\mathbb{E}_{\substack{(s_2,\eta_2)\sim\rho^{\pi^*}\\(s_t,\eta_t)\sim d^{\pi^*}_{(s_2,\eta_2)}}}\mathbb{E}_{(a_t,\eta_{t+1})\sim\pi^*}[\hat A^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})]\\
\le&\frac{\gamma}{1-\gamma}\sum_{s_t,\eta_t} d^{\pi^*}_{\rho^{\pi^*}}(s_t,\eta_t)\max_{a_t,\eta_{t+1}}\hat A^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})\\
=&\frac{\gamma}{1-\gamma}\sum_{s_t,\eta_t}\frac{d^{\pi^*}_{\rho^{\pi^*}}(s_t,\eta_t)}{d^\pi_{\mu^\pi}(s_t,\eta_t)}d^\pi_{\mu^\pi}(s_t,\eta_t)\max_{a_t,\eta_{t+1}}\hat A^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})\\
\overset{(a)}{\le}&\frac{\gamma}{1-\gamma}||\frac{d^{\pi^*}_{\rho^{\pi^*}}}{d^\pi_{\mu^\pi}}||_\infty\sum_{s_t,\eta_t}d^\pi_{\mu^\pi}(s_t,\eta_t)\max_{a_t,\eta_{t+1}}\hat A^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})\\
\overset{(b)}{=}&\frac{\gamma}{1-\gamma}||\frac{d^{\pi^*}_{\rho^{\pi^*}}}{d^\pi_{\mu^\pi}}||_\infty\max_{\bar\pi_2\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}||\mathcal{H}|}}\sum_{\substack{s_t,\eta_t\\a_t,\eta_{t+1}}}d^\pi_{\mu^\pi}(s_t,\eta_t)\bar\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)\hat A^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})\\
\overset{(c)}{=}&\frac{\gamma}{1-\gamma}||\frac{d^{\pi^*}_{\rho^{\pi^*}}}{d^\pi_{\mu^\pi}}||_\infty\max_{\bar\pi_2\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}||\mathcal{H}|}}\sum_{\substack{s_t,\eta_t\\a_t,\eta_{t+1}}}d^\pi_{\mu^\pi}(s_t,\eta_t)(\bar\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)-\pi_2(a_t,\eta_{t+1}|s_t,\eta_t))\hat A^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})\\
\overset{(d)}{=}&||\frac{d^{\pi^*}_{\rho^{\pi^*}}}{d^\pi_{\mu^\pi}}||_\infty\max_{\bar\pi_2\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}||\mathcal{H}|}}\sum_{\substack{s_t,\eta_t\\a_t,\eta_{t+1}}}\frac{\gamma}{1-\gamma}d^\pi_{\mu^\pi}(s_t,\eta_t)(\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)-\bar\pi_2(a_t,\eta_{t+1}|s_t,\eta_t))\hat Q^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})\\
\overset{(e)}{=}&||\frac{d^{\pi^*}_{\rho^{\pi^*}}}{d^\pi_{\mu^\pi}}||_\infty\max_{\bar\pi_2\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}||\mathcal{H}|}}(\pi_2-\bar\pi)^\intercal\nabla_{\pi_2}J^\pi(\mu)
\end{aligned}
$$

where (a) is true because $\max_{a_t,\eta_{t+1}}\hat A^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1}) = \hat J^{\pi_2}(s_t,\eta_t) - \min_{a_t,\eta_{t+1}}\hat Q^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1}) \ge 0$; (b) follows since $\max_{\bar\pi_2}$ is attained at an action (a_t,η_{t+1}) that maximizes $\hat A^{\pi_2}(s_t,\eta_t,\cdot,\cdot)$ for each state s_t,η_t; (c) follows since $\sum_{a_t,\eta_{t+1}}\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)\hat A^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1}) = \hat J^{\pi_2}(s_t,\eta_t) - \sum_{a_t,\eta_{t+1}}\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)\hat Q^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})$ $= 0$; (d) follows since $\sum_{a_t,\eta_{t+1}}(\bar\pi_2(a_t,\eta_{t+1}|s_t,\eta_t) - \pi_2(a_t,\eta_{t+1}|s_t,\eta_t))\hat J^{\pi_2}(s_t,\eta_t) = 0$; and (e) follows from Theorem 3.2 and Eq. (10).

Combining these two terms, we obtain

$$
\begin{aligned}
&J^\pi(\rho) - J^{\pi^*}(\rho)\\
\le&||\frac{\rho}{\mu}||_\infty\max_{\bar\pi_1\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|}}(\pi_1-\bar\pi_1)^\intercal\nabla_{\pi_1}J^\pi(\mu) + ||\frac{d^{\pi^*}_{\rho^{\pi^*}}}{d^\pi_{\mu^\pi}}||_\infty\max_{\bar\pi_2\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}||\mathcal{H}|}}(\pi_2-\bar\pi_2)^\intercal\nabla_{\pi_2}J^\pi(\mu)\\
\le&\max\{||\frac{\rho}{\mu}||_\infty, ||\frac{d^{\pi^*}_{\rho^{\pi^*}}}{d^\pi_{\mu^\pi}}||_\infty\}\max_{\bar\pi\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}}(\pi-\bar\pi)^\intercal\nabla_\pi J^\pi(\mu)
\end{aligned}
$$

where the last inequality is because $\max_{\bar\pi_1\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|}}(\pi_1-\bar\pi_1)^\intercal\nabla_{\pi_1}J^\pi(\mu) \ge 0$, $\max_{\bar\pi_2\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}||\mathcal{H}|}}(\pi_2-\bar\pi_2)^\intercal\nabla_{\pi_2}J^\pi(\mu) \ge 0$. Furthermore, we have

$$
\begin{aligned}
&d^\pi_{\mu^\pi}(s,\eta)\\
=&(1-\gamma)\sum_{s_2,\eta_2}\mu^\pi(s_2,\eta_2)\sum_{t=0}^\infty\gamma^t\mathrm{Pr}^\pi(s_{t+2}=s,\eta_{t+2}=\eta|s_2,\eta_2)
\end{aligned}
$$

$$\begin{aligned}
&\geq (1-\gamma)\mu^{\pi}(s,\eta) \\
&= (1-\gamma)\sum_{s_1}\mu(s_1)\sum_{a_1}\pi_1(a_1,\eta|s_1)P(s|s_1,a_1) \\
&\geq (1-\gamma)\pi_1^{LB}\sum_{s_1,a_1}\mu(s_1)P(s|s_1,a_1) = (1-\gamma)\pi_1^{LB}\mu^P(s,\eta)
\end{aligned}$$

then

$$\begin{aligned}
&J^{\pi}(\rho) - J^{\pi^*}(\rho) \\
\leq &\max\{||\frac{\rho}{\mu}||_\infty, \frac{1}{(1-\gamma)\pi_1^{LB}}||\frac{d^{\pi^*}_{\rho^{\pi^*}}}{\mu^P}||_\infty\}\max_{\bar{\pi}}(\pi-\bar{\pi})^\mathsf{T}\nabla_\pi J^{\pi}(\mu)
\end{aligned}$$

This concludes the proof. □

Next we define the gradient mapping and first-order optimality for constrained optimization in Definitions B.1 and B.2, respectively.

Definition B.1. Define the gradient mapping $G^\beta(\pi)$ as

$$G^\beta(\pi) = \frac{1}{\beta}(\pi - P_{\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}}(\pi - \beta\nabla_\pi J^{\pi}(\mu)))$$

where P_C is the projection onto C.

Then the update rule for the projected gradient descent is $\pi^+ = \pi - \beta G^\beta(\pi)$.

Definition B.2. A policy $\pi \in \Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}$ is ϵ-stationary with respect to the initial state distribution μ if

$$\min_{\pi+\delta\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}, ||\delta||_2\leq 1}\delta^\mathsf{T}\nabla_\pi J^{\pi}(\mu) \geq -\epsilon,$$

where $\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}$ is the set of all feasible policies.

Definition B.2 says that if $\epsilon = 0$, then any feasible direction of movement is positively correlated with the gradient. Since our goal is to minimize the objective function, this means that π is first-order stationary.

Proposition B.3 (Proposition B.1 in (Agarwal et al., 2021)). *Suppose that $J^\pi(\mu)$ is σ-smooth in π. Let $\pi^+ = \pi - \beta G^\beta(\pi)$. If $||G^\beta(\pi)||_2 \leq \epsilon$, then*

$$\min_{\pi^++\delta\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}, ||\delta||_2\leq 1}\delta^\mathsf{T}\nabla_\pi J^{\pi^+}(\mu) \geq -\epsilon(\beta\sigma+1).$$

Proof of Proposition B.3. By Lemma E.3,

$$-\nabla_\pi J^{\pi^+}(\mu) \in N_{\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}}(\pi^+) + \epsilon(\beta\sigma+1)B_2$$

where B_2 is the unit ℓ_2 ball, and N_C is the normal cone of the set C. Since $-\nabla_\pi J^{\pi^+}(\mu)$ is $\epsilon(\beta\sigma+1)$ distance from the normal cone $N_{\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}}(\pi^+)$ and δ is in the tangent cone of $\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}$ at π^+, we have $-\delta^\mathsf{T}\nabla_\pi J^{\pi^+}(\mu) \leq \epsilon(\beta\sigma+1)$. Thus

$$\min_{\pi^++\delta\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}, ||\delta||_2\leq 1}\delta^\mathsf{T}\nabla_\pi J^{\pi^+}(\mu) \geq -\epsilon(\beta\sigma+1).$$

This completes the proof. □

Proof of Theorem 3.7. [Iteration complexity for projected gradient descent] From Lemma 3.3, we know that $J^\pi(s_1)$ is σ-smooth for all state s_1 and $J^\pi(\mu)$ is also σ-smooth with $\sigma = \frac{2\gamma|\mathcal{A}||\mathcal{H}|}{(1-\gamma)^3}||\bar{c}||_\infty$. Then using Theorem E.2, we have that for stepsize $\beta = \frac{1}{\sigma}$,

$$\min_{t=0,1,\ldots,T-1}||G^\beta(\pi^{(t)})||_2 \leq \frac{\sqrt{2\sigma(J^{\pi^{(0)}}(\mu) - J^*(\mu))}}{\sqrt{T}}$$

From Proposition B.3, we have

$$\max_{t=0,1,\dots,T-1}\min_{\pi^{(t+1)}+\delta\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|},||\delta||_2\le 1}\delta^{\mathsf{T}}\nabla_\pi J^{\pi^{(t+1)}}(\mu)\ge-(\beta\sigma+1)\frac{\sqrt{2\sigma(J^{\pi^{(0)}}(\mu)-J^*(\mu))}}{\sqrt{T}}$$

Observe that

$$\begin{aligned}\max_{\bar\pi\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}}(\pi-\bar\pi)^{\mathsf{T}}\nabla_\pi J^\pi(\mu)=&-2\sqrt{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}\min_{\bar\pi\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}}\frac{1}{2\sqrt{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}}(\bar\pi-\pi)^{\mathsf{T}}\nabla_\pi J^\pi(\mu)\\ \le&-2\sqrt{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}\min_{\pi+\delta\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|},||\delta||_2\le1}\delta^{\mathsf{T}}\nabla_\pi J^\pi(\mu)\end{aligned}$$

where the last step follows as $||\bar\pi-\pi||_2\le||\bar\pi||_2+||\pi||_2=\sqrt{\sum_s\sum_{a,\eta}\bar\pi_1(a,\eta|s)^2+\sum_{s,\eta}\sum_{a,\eta'}\bar\pi_2(a,\eta'|s,\eta)^2}+\sqrt{\sum_s\sum_{a,\eta}\pi_1(a,\eta|s)^2+\sum_{s,\eta}\sum_{a,\eta'}\pi_2(a,\eta'|s,\eta)^2}\le\sqrt{\sum_s\sum_{a,\eta}\bar\pi_1(a,\eta|s)+\sum_{s,\eta}\sum_{a,\eta'}\bar\pi_2(a,\eta'|s,\eta)}+\sqrt{\sum_s\sum_{a,\eta}\pi_1(a,\eta|s)+\sum_{s,\eta}\sum_{a,\eta'}\pi_2(a,\eta'|s,\eta)}=2\sqrt{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}$ and $\pi+\frac{1}{2\sqrt{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}}(\bar\pi-\pi)\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}$ following the convexity of the probability simplex.

Then using Theorem 3.5 and $\beta\sigma=1$, we have

$$\begin{aligned}&\min_{t=1,\dots,T}J^{\pi^{(t)}}(\rho)-J^*(\rho)\\ \le&\min_{t=1,\dots,T}D_1\max_{\bar\pi\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}}(\pi^{(t)}-\bar\pi)^{\mathsf{T}}\nabla_\pi J^{\pi^{(t)}}(\mu)\\ \le&-2D_1\sqrt{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}\max_{t=1,\dots,T}\min_{\pi^{(t)}+\delta\in\Delta(\mathcal{A}\times\mathcal{H})^{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|},||\delta||_2\le1}\delta^{\mathsf{T}}\nabla_\pi J^{\pi^{(t)}}(\mu)\\ \le&4D_1\sqrt{|\mathcal{S}|+|\mathcal{S}||\mathcal{H}|}\frac{\sqrt{2\sigma(J^{\pi^{(0)}}(\mu)-J^*(\mu))}}{\sqrt{T}}\\ \overset{(a)}{\le}&4D_1\sqrt{2|\mathcal{S}||\mathcal{H}|}\frac{\sqrt{2\sigma(\frac{\lambda}{\alpha}+\lambda+\frac{1}{1-\gamma}||\bar c||_\infty)}}{\sqrt{T}}\end{aligned}$$

where (a) is true because $J^{\pi^{(0)}}(\mu)-J^*(\mu)\le J^{\pi^{(0)}}(\mu)\le\frac{\lambda}{\alpha}+\lambda+\frac{1}{1-\gamma}||\bar c||_\infty$ from Eq. (20) and $|\mathcal{H}|\ge1$. If we set T such that

$$4D_1\sqrt{2|\mathcal{S}||\mathcal{H}|}\frac{\sqrt{2\sigma(\frac{\lambda}{\alpha}+\lambda+\frac{1}{1-\gamma}||\bar c||_\infty)}}{\sqrt{T}}\le\epsilon$$

or, equivalently,

$$T\ge D_1^2\frac{64|\mathcal{S}||\mathcal{H}|}{\epsilon^2}\sigma(\frac{\lambda}{\alpha}+\lambda+\frac{1}{1-\gamma}||\bar c||_\infty)$$

then $\min_{t=1,\dots,T}J^{\pi^{(t)}}(\rho)-J^*(\rho)\le\epsilon$. Using $\sigma=\frac{2\gamma|\mathcal{A}||\mathcal{H}|}{(1-\gamma)^3}||\bar c||_\infty$ from Lemma 3.3 and $||\bar c||_\infty\le\frac{\lambda}{\alpha}+(1-\lambda)+\gamma\lambda$ leads to the desired result. □

C. Proofs in Section 3.2

Proof of Lemma 3.8. [Gradients for softmax parameterization] According to Theorem 3.2, we have

$$\begin{aligned}\nabla_{\theta_1}J^\pi(\mu)=&\mathbb{E}_{s_1\sim\mu}\mathbb{E}_{(a_1,\eta_2)\sim\pi_1(\cdot|s_1)}[\nabla_{\theta_1}\log\pi_1(a_1,\eta_2|s_1)Q^{\pi_2}(s_1,a_1,\eta_2)]\\ \nabla_{\theta_2}J^\pi(\mu)=&\frac{\gamma}{1-\gamma}\mathbb{E}_{(s_t,\eta_t)\sim d_{\mu}^{\pi}{}^{\pi}}\mathbb{E}_{(a_t,\eta_{t+1})\sim\pi_2(\cdot|s_t,\eta_t)}[\nabla_{\theta_2}\log\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)\hat{Q}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})].\end{aligned}$$

Because of the softmax parameterization, we have $\sum_{a_1,\eta_2}\pi_1(a_1,\eta_2|s_1)=\sum_{a_t,\eta_{t+1}}\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)=1$ satisfied automatically for all $s_1,s_t\in\mathcal{S},\ \eta_t\in\mathcal{H},\ t\geq 2$. As a result, $\sum_{a_1,\eta_2}\nabla_{\theta_1}\pi_1(a_1,\eta_2|s_1)=\sum_{a_t,\eta_{t+1}}\nabla_{\theta_2}\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)=0$ and we have

$$\begin{aligned}
\nabla_{\theta_1}J^\pi(\mu)=&\mathbb{E}_{s_1\sim\mu}\mathbb{E}_{(a_1,\eta_2)\sim\pi_1(\cdot|s_1)}[\nabla_{\theta_1}\log\pi_1(a_1,\eta_2|s_1)(-A^\pi(s_1,a_1,\eta_2))]\\
\nabla_{\theta_2}J^\pi(\mu)=&\frac{\gamma}{1-\gamma}\mathbb{E}_{(s_t,\eta_t)\sim d^\pi_{\mu^\pi}}\mathbb{E}_{(a_t,\eta_{t+1})\sim\pi_2(\cdot|s_t,\eta_t)}[\nabla_{\theta_2}\log\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)(-\hat{A}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1}))].
\end{aligned}$$

Because $\log\pi_1(a_1,\eta_2|s_1)=\theta_1(s_1,a_1,\eta_2)-\log\sum_{a_1',\eta_2'}\exp\left(\theta_1(s_1,a_1',\eta_2')\right)$ and $\log\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)=\theta_2(s_t,\eta_t,a_t,\eta_{t+1})-\log\sum_{a_t',\eta_{t+1}'}\exp\left(\theta_2(s_t,\eta_t,a_t',\eta_{t+1}')\right)$, we have

$$\begin{aligned}
\frac{\partial\log\pi_1(a_1,\eta_2|s_1)}{\partial\theta_1(s,a,\eta)}&=\mathbf{1}(s_1=s,a_1=a,\eta_2=\eta)-\frac{\exp(\theta_1(s,a,\eta))}{\sum_{a_1',\eta_2'}\exp(\theta_1(s,a_1',\eta_2'))}\mathbf{1}(s_1=s)\\
&=\mathbf{1}(s_1=s)\Big(\mathbf{1}(a_1=a,\eta_2=\eta)-\pi_1(a,\eta|s)\Big)\\
\frac{\partial\log\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)}{\partial\theta_2(s,\eta,a,\eta')}&=\mathbf{1}(s_t=s,\eta_t=\eta,a_t=a,\eta_{t+1}=\eta')-\frac{\exp(\theta_2(s,\eta,a,\eta'))}{\sum_{a_t',\eta_{t+1}'}\exp(\theta_2(s,\eta,a_t',\eta_{t+1}'))}\mathbf{1}(s_t=s,\eta_t=\eta)\\
&=\mathbf{1}(s_t=s,\eta_t=\eta)\Big(\mathbf{1}(a_t=a,\eta_{t+1}=\eta')-\pi_2(a,\eta'|s,\eta)\Big)
\end{aligned}$$

and thus,

$$\begin{aligned}
\frac{\partial J^{\pi^\theta}(\mu)}{\partial\theta_1(s,a,\eta)}=&\sum_{s_1}\mu(s_1)\sum_{a_1,\eta_2}\pi_1(a_1,\eta_2|s_1)\mathbf{1}(s_1=s)\Big(\mathbf{1}(a_1=a,\eta_2=\eta)-\pi_1(a,\eta|s)\Big)(-A^\pi(s_1,a_1,\eta_2))\\
=&\sum_{s_1}\mu(s_1)\sum_{a_1,\eta_2}\pi_1(a_1,\eta_2|s_1)\mathbf{1}((s_1,a_1,\eta_2)=(s,a,\eta))(-A^\pi(s_1,a_1,\eta_2))\\
&-\pi_1(a,\eta|s)\sum_{s_1}\mu(s_1)\sum_{a_1,\eta_2}\pi_1(a_1,\eta_2|s_1)\mathbf{1}(s_1=s)(-A^\pi(s_1,a_1,\eta_2))\\
\overset{(a)}{=}&\mu(s)\pi_1(a,\eta|s)(-A^\pi(s,a,\eta))
\end{aligned}$$

and

$$\begin{aligned}
&\frac{\partial J^{\pi^\theta}(\mu)}{\partial\theta_2(s,\eta,a,\eta')}\\
=&\frac{\gamma}{1-\gamma}\sum_{s_t,\eta_t}d^\pi_{\mu^\pi}(s_t,\eta_t)\sum_{a_t,\eta_{t+1}}\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)\mathbf{1}(s_t=s,\eta_t=\eta)\Big(\mathbf{1}(a_t=a,\eta_{t+1}=\eta')-\pi_2(a,\eta'|s,\eta)\Big)(-\hat{A}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1}))\\
=&\frac{\gamma}{1-\gamma}\sum_{s_t,\eta_t}d^\pi_{\mu^\pi}(s_t,\eta_t)\sum_{a_t,\eta_{t+1}}\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)\mathbf{1}((s_t,\eta_t,a_t,\eta_{t+1})=(s,\eta,a,\eta'))(-\hat{A}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1}))\\
&-\pi_2(a,\eta'|s,\eta)\frac{\gamma}{1-\gamma}\sum_{s_t,\eta_t}d^\pi_{\mu^\pi}(s_t,\eta_t)\sum_{a_t,\eta_{t+1}}\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)\mathbf{1}((s_t,\eta_t)=(s,\eta))(-\hat{A}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1}))\\
\overset{(b)}{=}&\frac{\gamma}{1-\gamma}d^\pi_{\mu^\pi}(s,\eta)\pi_2(a,\eta'|s,\eta)(-\hat{A}^{\pi_2}(s,\eta,a,\eta'))
\end{aligned}$$

where (a) is true because $\sum_{a_1,\eta_2}\pi_1(a_1,\eta_2|s)A^\pi(s,a_1,\eta_2)=0$ and (b) is true because $\sum_{a_t,\eta_{t+1}}\pi_2(a_t,\eta_{t+1}|s,\eta)\hat{A}^{\pi_2}(s,\eta,a_t,\eta_{t+1})=0$. This completes the proof. □

Proof of Lemma 3.9. [Smoothness for softmax parameterization] For $J^{\pi^\theta}(\mu)$, let $\theta_s\in\mathbb{R}^{|\mathcal{A}||\mathcal{H}|}$ denote the parameters associated with a given state s and $\theta_{s,\eta}\in\mathbb{R}^{|\mathcal{A}||\mathcal{H}|}$ denote the parameters associated with a given state s,η. We have

$$\begin{aligned}
&\nabla_{\theta_s}\pi_1^\theta(a,\eta|s)=\pi_1^\theta(a,\eta|s)(e_{a,\eta}-\pi_1^\theta(\cdot,\cdot|s))\\
&\nabla_{\theta_{s_t,\eta_t}}\pi_2^\theta(a_t,\eta_{t+1}|s_t,\eta_t)=\pi_2^\theta(a_t,\eta_{t+1}|s_t,\eta_t)(e_{a_t,\eta_{t+1}}-\pi_2^\theta(\cdot,\cdot|s_t,\eta_t))
\end{aligned}$$

where $e_{a,\eta}$ is a basis vector that only equals 1 at (a, η)-location and 0 elsewhere. Differentiating them with respect to θ again, we get

$$\nabla^2_{\theta_s}\pi_1^\theta(a,\eta|s) = \pi_1^\theta(a,\eta|s)\left(e_{a,\eta}e_{a,\eta}^\mathsf{T} - e_{a,\eta}\pi_1^\theta(\cdot,\cdot|s)^\mathsf{T} - \pi_1^\theta(\cdot,\cdot|s)e_{a,\eta}^\mathsf{T} + 2\pi_1^\theta(\cdot,\cdot|s)\pi_1^\theta(\cdot,\cdot|s)^\mathsf{T} - \text{diag}(\pi_1^\theta(\cdot,\cdot|s))\right)$$

$$\begin{aligned}\nabla^2_{\theta_{s_t,\eta_t}}\pi_2^\theta(a_t,\eta_{t+1}|s_t,\eta_t) = \pi_2^\theta(a_t,\eta_{t+1}|s_t,\eta_t)\Big(&e_{a_t,\eta_{t+1}}e_{a_t,\eta_{t+1}}^\mathsf{T} - e_{a_t,\eta_{t+1}}\pi_2^\theta(\cdot,\cdot|s_t,\eta_t)^\mathsf{T} - \pi_2^\theta(\cdot,\cdot|s_t,\eta_t)e_{a_t,\eta_{t+1}}^\mathsf{T} \\ &+ 2\pi_2^\theta(\cdot,\cdot|s_t,\eta_t)\pi_2^\theta(\cdot,\cdot|s_t,\eta_t)^\mathsf{T} - \text{diag}(\pi_2^\theta(\cdot,\cdot|s_t,\eta_t))\Big)\end{aligned}$$

Define $\pi_1^\alpha := \pi_1^{\theta+\alpha u}$ where $u \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}||\mathcal{H}|}$ is a unit vector and denote $u_s \in \mathbb{R}^{|\mathcal{A}||\mathcal{H}|}$ as the parameters associated with a given state s. Differentiating π_1^α with respect to α, we obtain

$$\begin{aligned}\sum_{a_1,\eta_2}\left|\frac{d\pi_1^\alpha(a_1,\eta_2|s_1)}{d\alpha}|_{\alpha=0}\right| &\le \sum_{a_1,\eta_2}\left|u^\mathsf{T}\nabla_{\theta+\alpha u}\pi_1^\alpha(a_1,\eta_2|s_1)|_{\alpha=0}\right| \\ &\le \sum_{a_1,\eta_2}\pi_1^\alpha(a_1,\eta_2|s_1)|u_{s_1}^\mathsf{T}e_{a_1,\eta_2} - u_{s_1}^\mathsf{T}\pi_1^\alpha(\cdot,\cdot|s_1)| \\ &\le \max_{a_1,\eta_2}\left(|u_{s_1}^\mathsf{T}e_{a_1,\eta_2}| + |u_{s_1}^\mathsf{T}\pi_1^\alpha(\cdot,\cdot|s_1)|\right) \le 2\end{aligned}$$

Now differentiating once again with respect to α, we get

$$\begin{aligned}\sum_{a_1,\eta_2}\left|\frac{d^2\pi_1^\alpha(a_1,\eta_2|s_1)}{(d\alpha)^2}|_{\alpha=0}\right| &\le \sum_{a_1,\eta_2}\left|u^\mathsf{T}\nabla^2_{\theta+\alpha u}\pi_1^\alpha(a_1,\eta_2|s_1)|_{\alpha=0}u\right| \\ &\le \max_{a_1,\eta_2}\Big(|u_s^\mathsf{T}e_{a_1,\eta_2}e_{a_1,\eta_2}^\mathsf{T}u_s| + |u_s^\mathsf{T}e_{a_1,\eta_2}\pi_1^\alpha(\cdot,\cdot|s_1)^\mathsf{T}u_s| + |u_s^\mathsf{T}\pi_1^\alpha(\cdot,\cdot|s_1)e_{a_1,\eta_2}^\mathsf{T}u_s| \\ &\quad + 2|u_s^\mathsf{T}\pi_1^\alpha(\cdot,\cdot|s)\pi_1^\alpha(\cdot,\cdot|s)^\mathsf{T}u_s| + |u_s^\mathsf{T}\text{diag}(\pi_1^\alpha(\cdot,\cdot|s_1)u_s|\Big) \le 6\end{aligned}$$

The same arguments apply to π_2^α, where we get $\sum_{a_t,\eta_{t+1}}\left|\frac{d\pi_2^\alpha(a_t,\eta_{t+1}|s_t,\eta_t)}{d\alpha}|_{\alpha=0}\right| \le 2$ and $\sum_{a_t,\eta_{t+1}}\left|\frac{d^2\pi_2^\alpha(a_t,\eta_{t+1}|s_t,\eta_t)}{(d\alpha)^2}|_{\alpha=0}\right| \le 6$.

Let $\tilde{J}(\alpha) := J^{\pi^\alpha}(s_1)$. Using this with Lemma D.1 for $C_1 = 2,\ C_2 = 6$, we get

$$\begin{aligned}\max_{||u||_2=1}\left|\frac{d^2\tilde{J}(\alpha)}{(d\alpha)^2}|_{\alpha=0}\right| &\le C_2\left(\frac{\lambda}{\alpha} + \lambda + \frac{||\bar{c}||_\infty}{(1-\gamma)^2}\right) + \frac{2\gamma C_1^2}{(1-\gamma)^3}||\bar{c}||_\infty \\ &\le 6\left(\frac{\lambda}{\alpha} + \lambda + \frac{||\bar{c}||_\infty}{(1-\gamma)^2}\right) + \frac{8\gamma}{(1-\gamma)^3}||\bar{c}||_\infty \\ &\le 6\left(\frac{\lambda}{\alpha} + \lambda\right) + \frac{8}{(1-\gamma)^3}||\bar{c}||_\infty\end{aligned}$$

where the last inequality uses the fact that $\gamma \le 1$. Therefore $J^{\pi^\theta}(\mu)$ is $6(\frac{\lambda}{\alpha} + \lambda) + \frac{8}{(1-\gamma)^3}||\bar{c}||_\infty$-smooth.

Next let us bound the smoothness of the regularizer $-\frac{\kappa}{|\mathcal{S}|}R_1(\theta)$, where

$$R_1(\theta) := \frac{1}{|\mathcal{A}||\mathcal{H}|}\sum_{s_1,a_1,\eta_2}\log\pi_1^\theta(a_1,\eta_2|s_1).$$

We have

$$\frac{\partial R_1(\theta)}{\partial\theta_{s_1,a_1,\eta_2}} = \frac{1}{|\mathcal{A}||\mathcal{H}|} - \pi_1^\theta(a_1,\eta_2|s_1)$$

Equivalently,

$$\nabla_{\theta_{s_1}}R_1(\theta) = \frac{1}{|\mathcal{A}||\mathcal{H}|}\mathbf{1} - \pi_1^\theta(\cdot,\cdot|s_1)$$

As a result,

$$\nabla^2_{\theta_{s_1}} R_1(\theta) = -\text{diag}(\pi_1^\theta(\cdot,\cdot|s_1)) + \pi_1^\theta(\cdot,\cdot|s_1)\pi_1^\theta(\cdot,\cdot|s_1)^\top$$

and for any vector u_{s_1},

$$|u_{s_1}^\top \nabla^2_{\theta_{s_1}} R_1(\theta) u_{s_1}| = |u_{s_1}^\top \text{diag}(\pi_1^\theta(\cdot,\cdot|s_1))u_{s_1} - (u_{s_1}\pi_1^\theta(\cdot,\cdot|s_1))^2| \le 2||u_{s_1}||_\infty^2$$

Because $\nabla_{\theta_s}\nabla_{\theta_{s'}} R_1(\theta) = 0$ for $s \neq s'$, we have

$$|u^\top \nabla^2_\theta R_1(\theta) u| = |\sum_{s_1} u_{s_1}^\top \nabla^2_{\theta_{s_1}} R_1(\theta) u_{s_1}| \le 2\sum_{s_1} ||u_{s_1}||_\infty^2 \le 2||u||_2^2$$

Therefore $R_1(\theta)$ is 2-smooth and $-\frac{\kappa}{|\mathcal{S}|}R_1(\theta)$ is $\frac{2\kappa}{|\mathcal{S}|}$-smooth. The same arguments apply to $R_2(\theta) = \frac{1}{|\mathcal{A}||\mathcal{H}|}\sum_{s_t,\eta_t,a_t,\eta_{t+1}} \log \pi_2^\theta(a_t,\eta_{t+1}|s_t,\eta_t)$, where we get $-\frac{\kappa}{|\mathcal{S}||\mathcal{H}|}R_2(\theta)$ is $\frac{2\kappa}{|\mathcal{S}||\mathcal{H}|}$-smooth.

As a result, $L_\kappa(\theta)$ is σ_κ-smooth with $\sigma_\kappa = 6(\frac{\lambda}{\alpha}+\lambda) + \frac{8}{(1-\gamma)^3}||\bar{c}||_\infty + \frac{2\kappa}{|\mathcal{S}|} + \frac{2\kappa}{|\mathcal{S}||\mathcal{H}|}$. This completes the proof. $\square$

Proof of Theorem 3.10. [Suboptimality for softmax parameterization] We only need to show that $\max_{a_1,\eta_2} A^\pi(s_1,a_1,\eta_2) \le \frac{2\kappa}{\mu(s_1)|\mathcal{S}|}$, $\forall s_1 \in \mathcal{S}$ and $\max_{a_t,\eta_{t+1}} \hat{A}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1}) \le \frac{2\kappa}{\gamma\pi_1^{LB}|\mathcal{S}||\mathcal{H}|\mu^P(s_t,\eta_t)}$, $\forall s_t \in \mathcal{S}$, $\eta_t \in \mathcal{H}$. To see why this is sufficient, observe that by the performance difference lemma (Lemma 3.1),

$$\begin{aligned}
&J^\pi(\rho) - J^{\pi^*}(\rho)\\
=&\mathbb{E}_{s_1\sim\rho}\mathbb{E}_{\tau\sim\Pr^{\pi^*}(\tau|s_1)}\Big[A^\pi(s_1,a_1,\eta_2) + \sum_{t=2}^\infty \gamma^{t-1}\hat{A}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})\Big]\\
=&\sum_{s_1,a_1,\eta_2}\rho(s_1)\pi_1^*(a_1,\eta_2|s_1)A^\pi(s_1,a_1,\eta_2) + \frac{\gamma}{1-\gamma}\sum_{s_t,\eta_t} d_{\rho^{\pi^*}}^{\pi^*}(s_t,\eta_t)\sum_{a_t,\eta_{t+1}}\pi_2^*(a_t,\eta_{t+1}|s_t,\eta_t)\hat{A}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})\\
\le&\sum_{s_1}\rho(s_1)\max_{a_1,\eta_2}A^\pi(s_1,a_1,\eta_2) + \frac{\gamma}{1-\gamma}\sum_{s_t,\eta_t} d_{\rho^{\pi^*}}^{\pi^*}(s_t,\eta_t)\max_{a_t,\eta_{t+1}}\hat{A}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})\\
\le&\sum_{s_1}\rho(s_1)\frac{2\kappa}{\mu(s_1)|\mathcal{S}|} + \frac{\gamma}{1-\gamma}\sum_{s_t,\eta_t} d_{\rho^{\pi^*}}^{\pi^*}(s_t,\eta_t)\frac{2\kappa}{\gamma\pi_1^{LB}|\mathcal{S}||\mathcal{H}|\mu^P(s_t,\eta_t)}\\
\le&2\kappa||\frac{\rho}{\mu}||_\infty + \frac{2\kappa}{(1-\gamma)\pi_1^{LB}}||\frac{d_{\rho^{\pi^*}}^{\pi^*}}{\mu^P}||_\infty
\end{aligned}$$

Now to prove $\max_{a_1,\eta_2} A^\pi(s_1,a_1,\eta_2) \le \frac{2\kappa}{\mu(s_1)|\mathcal{S}|}$, it suffices to bound $A^\pi(s_1,a_1,\eta_2)$ for any state-action pair s_1,a_1,η_2 where $A^\pi(s_1,a_1,\eta_2) \ge 0$ otherwise the claim holds trivially. Consider an (s_1,a_1,η_2) pair such that $A^\pi(s_1,a_1,\eta_2) \ge 0$. Using Lemma 3.8,

$$\frac{\partial L_\kappa(\theta)}{\partial\theta_1(s_1,a_1,\eta_2)} = \mu(s_1)\pi_1^\theta(a_1,\eta_2|s_1)(-A^{\pi^\theta}(s_1,a_1,\eta_2)) - \frac{\kappa}{|\mathcal{S}|}(\frac{1}{|\mathcal{A}||\mathcal{H}|} - \pi_1^\theta(a_1,\eta_2|s_1)) \tag{14}$$

The gradient norm assumption $||\nabla_{\theta_1}L_\kappa(\theta)||_2 \le \epsilon_1 \le \frac{\kappa}{2|\mathcal{S}||\mathcal{A}||\mathcal{H}|}$ implies that

$$\begin{aligned}
-\frac{\kappa}{2|\mathcal{S}||\mathcal{A}||\mathcal{H}|} \le -\epsilon_1 \overset{(a)}{\le} \frac{\partial L_\kappa(\theta)}{\partial\theta_1(s_1,a_1,\eta_2)} &= \mu(s)\pi_1^\theta(a_1,\eta_2|s_1)(-A^{\pi^\theta}(s_1,a_1,\eta_2)) - \frac{\kappa}{|\mathcal{S}|}(\frac{1}{|\mathcal{A}||\mathcal{H}|} - \pi_1^\theta(a_1,\eta_2|s_1)) \\
&\overset{(b)}{\le} -\frac{\kappa}{|\mathcal{S}|}(\frac{1}{|\mathcal{A}||\mathcal{H}|} - \pi_1^\theta(a_1,\eta_2|s_1))
\end{aligned} \tag{15}$$

where (a) is due to $|\frac{\partial L_\kappa(\theta)}{\partial\theta_1(s_1,a_1,\eta_2)}| \le ||\nabla_{\theta_1}L_\kappa(\theta)||_2 \le \epsilon_1$ and (b) uses the fact that $-A^\pi(s_1,a_1,\eta_2) \le 0$. Rearranging the terms, we get

$$\pi_1^\theta(a_1,\eta_2|s_1) \ge \frac{1}{|\mathcal{A}||\mathcal{H}|} - \frac{1}{2|\mathcal{A}||\mathcal{H}|} = \frac{1}{2|\mathcal{A}||\mathcal{H}|}$$

From (14), we have

$$
\begin{aligned}
A^{\pi^\theta}(s_1,a_1,\eta_2) =& \frac{1}{\mu(s_1)}\Big(\frac{1}{\pi_1^\theta(a_1,\eta_2|s_1)}\frac{-\partial L_\kappa(\theta)}{\partial\theta_1(s,a,\eta)} + \frac{\kappa}{|\mathcal{S}|}\big(1-\frac{1}{\pi_1^\theta(a_1,\eta_2|s_1)|\mathcal{A}||\mathcal{H}|}\big)\Big)\\
\leq& \frac{1}{\mu(s_1)}\Big(\frac{1}{\pi_1^\theta(a_1,\eta_2|s_1)}\left|\frac{-\partial L_\kappa(\theta)}{\partial\theta_1(s,a,\eta)}\right| + \frac{\kappa}{|\mathcal{S}|}\Big)\\
\leq& \frac{1}{\mu(s_1)}\big(2|\mathcal{A}||\mathcal{H}|\epsilon_1 + \frac{\kappa}{|\mathcal{S}|}\big)\\
\leq& \frac{2\kappa}{\mu(s_1)|\mathcal{S}|}
\end{aligned}
$$

To prove $\max_{a_t,\eta_{t+1}} \hat{A}^\pi(s_t,\eta_t,a_t,\eta_{t+1}) \leq \frac{2\kappa}{\gamma\pi_1^{LB}|\mathcal{S}||\mathcal{H}|\mu^P(s_t,\eta_t)}$, $\forall s_t,\eta_t$, it suffices to bound $\hat{A}^\pi(s_t,\eta_t,a_t,\eta_{t+1})$ for any state-action pair $s_t,\eta_t,a_t,\eta_{t+1}$ where $\hat{A}^\pi(s_t,\eta_t,a_t,\eta_{t+1}) \geq 0$. Consider an $(s_t,\eta_t,a_t,\eta_{t+1})$ pair such that $\hat{A}^\pi(s_t,\eta_t,a_t,\eta_{t+1}) \geq 0$. Using Lemma 3.8, we have

$$
\frac{\partial L_\kappa(\theta)}{\partial\theta_2(s_t,\eta_t,a_t,\eta_{t+1})} = \frac{\gamma}{1-\gamma}d^\pi_{\mu^\pi}(s_t,\eta_t)\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)(-\hat{A}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1})) - \frac{\kappa}{|\mathcal{S}||\mathcal{H}|}\Big(\frac{1}{|\mathcal{A}||\mathcal{H}|} - \pi_2(a_t,\eta_{t+1}|s_t,\eta_t)\Big) \tag{16}
$$

The gradient norm assumption $||\nabla_{\theta_2}L_\kappa(\theta)||_2 \leq \epsilon_2 \leq \frac{\kappa}{2|\mathcal{S}||\mathcal{H}||\mathcal{A}||\mathcal{H}|}$ implies that

$$
-\frac{\kappa}{2|\mathcal{S}||\mathcal{H}||\mathcal{A}||\mathcal{H}|} \leq -\epsilon_2 \leq \frac{\partial L_\kappa(\theta)}{\partial\theta_2(s_t,\eta_t,a_t,\eta_{t+1})} \leq -\frac{\kappa}{|\mathcal{S}||\mathcal{H}|}\Big(\frac{1}{|\mathcal{A}||\mathcal{H}|} - \pi_2(a_t,\eta_{t+1}|s_t,\eta_t)\Big)
$$

where the last inequality uses the fact that $-\hat{A}^\pi(s_t,\eta_t,a_t,\eta_{t+1}) \leq 0$. Rearranging the terms, we get

$$
\pi_2(a_t,\eta_{t+1}|s_t,\eta_t) \geq \frac{1}{|\mathcal{A}||\mathcal{H}|} - \frac{1}{2|\mathcal{A}||\mathcal{H}|} = \frac{1}{2|\mathcal{A}||\mathcal{H}|}
$$

From (16), we have

$$
\begin{aligned}
\hat{A}^{\pi_2}(s_t,\eta_t,a_t,\eta_{t+1}) =& \frac{1-\gamma}{\gamma}\frac{1}{d^\pi_{\mu^\pi}(s_t,\eta_t)}\Big(\frac{1}{\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)}\frac{-\partial L_\kappa(\theta)}{\partial\theta_2(s_t,\eta_t,a_t,\eta_{t+1})} + \frac{\kappa}{|\mathcal{S}||\mathcal{H}|}\big(1-\frac{1}{\pi_2(a_t,\eta_{t+1}|s_t,\eta_t)|\mathcal{A}||\mathcal{H}|}\big)\Big)\\
\leq& \frac{1-\gamma}{\gamma}\frac{1}{d^\pi_{\mu^\pi}(s_t,\eta_t)}\big(2|\mathcal{A}||\mathcal{H}|\epsilon_2 + \frac{\kappa}{|\mathcal{S}||\mathcal{H}|}\big)\\
\leq& \frac{1-\gamma}{\gamma}\frac{2\kappa}{d^\pi_{\mu^\pi}(s_t,\eta_t)|\mathcal{S}||\mathcal{H}|}\\
\leq& \frac{2\kappa}{\gamma\pi_1^{LB}|\mathcal{S}||\mathcal{H}|\mu^P(s_t,\eta_t)}
\end{aligned}
$$

where the last inequality uses the fact that $d^\pi_{\mu^\pi}(s_t,\eta_t) \geq (1-\gamma)\mu^\pi(s_t,\eta_t) \geq (1-\gamma)\pi_1^{LB}\mu^P(s_t,\eta_t)$ because $\pi_1(a,\eta_t|s) \geq \pi_1^{LB}$. This completes the proof. □

Proof of Lemma 3.11. [Positive action probabilities] Given any finite values of initialized parameters $\theta^{(0)} = (\theta_1^{(0)},\theta_2^{(0)})$, the action probabilities $\pi_1^{\theta^{(0)}}$ and $\pi_2^{\theta^{(0)}}$ will be bounded away from 0, i.e., $\pi_1^{\theta^{(0)}},\pi_2^{\theta^{(0)}} > 0$. As a result, the initial regularized objective function can be upper bounded, i.e., $L_\kappa(\theta^{(0)}) < +\infty$. Indeed, $L_\kappa(\theta^{(0)}) \leq \bar{C} - \frac{\kappa}{|\mathcal{S}||\mathcal{A}||\mathcal{H}|}\sum_{s_1,a_1,\eta_2}\log\pi_1^{\theta^{(0)}}(a_1,\eta_2|s_1) - \frac{\kappa}{|\mathcal{S}||\mathcal{H}||\mathcal{A}||\mathcal{H}|}\sum_{s_t,\eta_t,a_t,\eta_{t+1}}\log\pi_2^{\theta^{(0)}}(a_t,\eta_{t+1}|s_t,\eta_t) - 2\kappa\log|\mathcal{A}||\mathcal{H}| < +\infty$. Because $L_\kappa(\theta)$ is σ_κ-smooth based on Lemma 3.9, according to Theorem E.2, $L_\kappa(\theta^{(t)})$ is non-increasing following gradient updates (11). Thus, $L_\kappa(\theta^{(t)}) \leq L_\kappa(\theta^{(0)}) < +\infty$, $\forall t \geq 1$. Now assume, for the sake of contradiction, that there exists $t \in \mathbb{Z}_+ \cup \{+\infty\}$, s,a,η such that $\pi_1^{\theta^{(t)}}(a,\eta|s) = 0$. Then we have $-\log\pi_1^{\theta^{(t)}}(a,\eta|s) = +\infty$ and $L_\kappa(\theta^{(t)}) = +\infty$, which contradicts with $L_\kappa(\theta^{(t)}) < +\infty$. Similar arguments can be also applied to $\pi_2^{\theta^{(t)}}(a,\eta'|s,\eta)$ where we conclude that $\pi_2^{LB} := \inf_{t\geq 1}\min_{s,a,\eta,\eta'}\pi_2^{\theta^{(t)}}(a,\eta'|s,\eta) > 0$. This completes the proof. □

Proof of Theorem 3.12. [Iteration complexity for softmax parameterization with log barrier regularizer] Using Theorem 3.10, the desired optimality gap ϵ will follow if we set

$$\kappa = \frac{\epsilon}{2||\frac{\rho}{\mu}||_\infty + \frac{2}{(1-\gamma)\pi_1^{LB}}||\frac{d_\rho^{\pi^*}}{\mu^P}||_\infty} \tag{17}$$

and if $||\nabla_\theta L_\kappa(\theta)||_2 \le \frac{\kappa}{2|\mathcal{S}||\mathcal{H}||\mathcal{A}||\mathcal{H}|}$ (then we have $||\nabla_{\theta_1} L_\kappa(\theta)||_2 \le ||\nabla_\theta L_\kappa(\theta)||_2 \le \frac{\kappa}{2|\mathcal{S}||\mathcal{H}||\mathcal{A}||\mathcal{H}|} \le \frac{\kappa}{2|\mathcal{S}||\mathcal{A}||\mathcal{H}|}$ and $||\nabla_{\theta_2} L_\kappa(\theta)||_2 \le ||\nabla_\theta L_\kappa(\theta)||_2 \le \frac{\kappa}{2|\mathcal{S}||\mathcal{H}||\mathcal{A}||\mathcal{H}|}$). To proceed, we need to bound the iteration number to make the gradient sufficiently small. Using Theorem E.2, after T iterations of gradient descent with stepsize $1/\sigma_\kappa$, we have

$$\begin{aligned}
\min_{t<T} ||\nabla_\theta L_\kappa(\theta^{(t)})||_2 &\le \sqrt{\frac{2\sigma_\kappa(L_\kappa(\theta^{(0)}) - L_\kappa(\theta^*))}{T}} \\
&\overset{(a)}{\le} \sqrt{\frac{2\sigma_\kappa(\bar{C} - \kappa \log \pi_1^{LB} - \kappa \log \pi_2^{LB})}{T}} \\
&\overset{(b)}{\le} \sqrt{\frac{2\sigma_\kappa(\bar{C} - \log \pi_1^{LB} - \log \pi_2^{LB})}{T}}
\end{aligned}$$

where (a) is true because $L_\kappa(\theta^{(0)}) - L_\kappa(\theta^*) = J^{(0)}(\mu) - J^*(\mu) - \frac{\kappa}{|\mathcal{S}||\mathcal{A}||\mathcal{H}|}\sum_{s_1,a_1,\eta_2} \log \frac{\pi_1^{\theta^{(0)}}(a_1,\eta_2|s_1)}{\pi_1^{\theta^*}(a_1,\eta_2|s_1)} - \frac{\kappa}{|\mathcal{S}||\mathcal{H}||\mathcal{A}||\mathcal{H}|}\sum_{s_t,\eta_t,a_t,\eta_{t+1}} \log \frac{\pi_2^{\theta^{(0)}}(a_t,\eta_{t+1}|s_t,\eta_t)}{\pi_2^{\theta^*}(a_t,\eta_{t+1}|s_t,\eta_t)} \le \bar{C} - \kappa \log \pi_1^{LB} - \kappa \log \pi_2^{LB}$; (b) uses the fact that $\kappa \le 1$. Denoting $\bar{B} = \bar{C} - \log \pi_1^{LB} - \log \pi_2^{LB}$, we seek to ensure

$$\sqrt{\frac{2\sigma_\kappa \bar{B}}{T}} \le \frac{\kappa}{2|\mathcal{S}||\mathcal{H}||\mathcal{A}||\mathcal{H}|} \tag{18}$$

Choosing $T \ge \frac{8\sigma_\kappa \bar{B}|\mathcal{S}|^2|\mathcal{H}|^4|\mathcal{A}|^2}{\kappa^2}$ satisfies the above inequality. By Lemma 3.9, we can plug in $\sigma_\kappa = 6(\frac{\lambda}{\alpha}+\lambda) + \frac{8}{(1-\gamma)^3}||\bar{c}||_\infty + \frac{2\kappa}{|\mathcal{S}|} + \frac{2\kappa}{|\mathcal{S}||\mathcal{H}|}$, which gives us

$$\begin{aligned}
\frac{8\sigma_\kappa \bar{B}|\mathcal{S}|^2|\mathcal{H}|^4|\mathcal{A}|^2}{\kappa^2} &= \frac{48(\frac{\lambda}{\alpha}+\lambda)\bar{B}|\mathcal{S}|^2|\mathcal{H}|^4|\mathcal{A}|^2}{\kappa^2} + \frac{64||\bar{c}||_\infty \bar{B}|\mathcal{S}|^2|\mathcal{H}|^4|\mathcal{A}|^2}{(1-\gamma)^3\kappa^2} + \frac{16\kappa\bar{B}|\mathcal{S}||\mathcal{H}|^4|\mathcal{A}|^2}{\kappa^2} + \frac{16\kappa\bar{B}|\mathcal{S}||\mathcal{H}|^3|\mathcal{A}|^2}{\kappa^2} \\
&\overset{(a)}{\le} \frac{(48(\frac{\lambda}{\alpha}+\lambda) + 64||\bar{c}||_\infty + 32)\bar{B}|\mathcal{S}|^2|\mathcal{H}|^4|\mathcal{A}|^2}{(1-\gamma)^3\kappa^2} \\
&\overset{(b)}{=} \frac{(48(\frac{\lambda}{\alpha}+\lambda) + 64||\bar{c}||_\infty + 32)\bar{B}|\mathcal{S}|^2|\mathcal{H}|^4|\mathcal{A}|^2}{(1-\gamma)^3\epsilon^2}\left(2||\frac{\rho}{\mu}||_\infty + \frac{2}{(1-\gamma)\pi_1^{LB}}||\frac{d_\rho^{\pi^*}}{\mu^P}||_\infty\right)^2
\end{aligned}$$

where (a) is true because $0<\gamma<1,\ \kappa\le 1,\ |\mathcal{S}|,|\mathcal{H}|\ge 1$ and (b) uses Eq. (17).

Therefore, choosing $T \ge \frac{64(3(\frac{\lambda}{\alpha}+\lambda)+4||\bar{c}||_\infty+2)\bar{B}|\mathcal{S}|^2|\mathcal{H}|^4|\mathcal{A}|^2}{(1-\gamma)^3\epsilon^2} D_2^2$ with $D_2^2 = (||\frac{\rho}{\mu}||_\infty + \frac{1}{(1-\gamma)\pi_1^{LB}}||\frac{d_\rho^{\pi^*}}{\mu^P}||_\infty)^2$ satisfies (18). This completes the proof. □

D. Smoothness Results

In this section, we present a helpful lemma, which is applicable to both direct and softmax parameterizations. This lemma helps us prove the smoothness properties of the objective functions under direct and softmax parameterizations.

Lemma D.1. *Consider a unit vector u and let $\pi^\alpha := \pi^{\theta+\alpha u}$, $\tilde{J}(\alpha) := J^{\pi^\alpha}(s_1)$. Assume that*

$$\sum_{a_1,\eta_2}\left|\frac{d\pi_1^\alpha(a_1,\eta_2|s_1)}{d\alpha}|_{\alpha=0}\right| \le C_1,\ \sum_{a_1,\eta_2}\left|\frac{d^2\pi_1^\alpha(a_1,\eta_2|s_1)}{(d\alpha)^2}|_{\alpha=0}\right| \le C_2,\ \forall s_1\in\mathcal{S}$$

$$\sum_{a_t,\eta_{t+1}}\left|\frac{d\pi_2^\alpha(a_t,\eta_{t+1}|s_t,\eta_t)}{d\alpha}|_{\alpha=0}\right| \le C_1,\ \sum_{a_t,\eta_{t+1}}\left|\frac{d^2\pi_2^\alpha(a_t,\eta_{t+1}|s_t,\eta_t)}{(d\alpha)^2}|_{\alpha=0}\right| \le C_2,\ \forall s_t\in\mathcal{S},\ \eta_t\in\mathcal{H}$$

Then

$$\max_{||u||_2=1}\left|\frac{d^2\tilde{J}(\alpha)}{(d\alpha)^2}|_{\alpha=0}\right| \leq C_2(\frac{\lambda}{\alpha}+\lambda+\frac{||\bar{c}||_\infty}{(1-\gamma)^2})+\frac{2\gamma C_1^2}{(1-\gamma)^3}||\bar{c}||_\infty.$$

Proof of Lemma D.1. Let $\tilde{P}(\alpha)$ be the state-action transition matrix under policy π^α, i.e.,

$$[\tilde{P}(\alpha)]_{(s_t,\eta_t,a_t,\eta_{t+1})\to(s'_t,\eta'_t,a'_t,\eta'_{t+1})} = \begin{cases} P(s'_t|s_t,a_t)\pi_2^\alpha(a'_t,\eta'_{t+1}|s'_t,\eta'_t), \text{ if } \eta'_t=\eta_{t+1} \\ 0, \text{ otherwise.}\end{cases}$$

We can differentiate $\tilde{P}(\alpha)$ with respect to α:

$$\left[\frac{d\tilde{P}(\alpha)}{d\alpha}|_{\alpha=0}\right]_{(s_t,\eta_t,a_t,\eta_{t+1})\to(s'_t,\eta'_t,a'_t,\eta'_{t+1})} = \begin{cases} \frac{d\pi_2^\alpha(a'_t,\eta'_{t+1}|s'_t,\eta'_t)}{d\alpha}|_{\alpha=0}P(s'_t|s_t,a_t), \text{ if } \eta'_t=\eta_{t+1} \\ 0, \text{ otherwise.}\end{cases}$$

For any arbitrary vector x, we have

$$\left[\frac{d\tilde{P}(\alpha)}{d\alpha}|_{\alpha=0}x\right]_{(s_t,\eta_t,a_t,\eta_{t+1})} = \sum_{s'_t,a'_t,\eta'_{t+1}}\frac{d\pi_2^\alpha(a'_t,\eta'_{t+1}|s'_t,\eta_{t+1})}{d\alpha}|_{\alpha=0}P(s'_t|s_t,a_t)x_{s'_t,\eta_{t+1},a'_t,\eta'_{t+1}}$$

and thus

$$\begin{aligned}
\max_{||u||_2=1}\left|\left[\frac{d\tilde{P}(\alpha)}{d\alpha}|_{\alpha=0}x\right]_{(s_t,\eta_t,a_t,\eta_{t+1})}\right| &= \max_{||u||_2=1}\left|\sum_{s'_t,a'_t,\eta'_{t+1}}\frac{d\pi_2^\alpha(a'_t,\eta'_{t+1}|s'_t,\eta_{t+1})}{d\alpha}|_{\alpha=0}P(s'_t|s_t,a_t)x_{s'_t,\eta_{t+1},a'_t,\eta'_{t+1}}\right| \\
&\leq \max_{||u||_2=1}\sum_{s'_t,a'_t,\eta'_{t+1}}\left|\frac{d\pi_2^\alpha(a'_t,\eta'_{t+1}|s'_t,\eta_{t+1})}{d\alpha}|_{\alpha=0}\right|P(s'_t|s_t,a_t)|x_{s'_t,\eta_{t+1},a'_t,\eta'_{t+1}}| \\
&\leq \max_{||u||_2=1}\sum_{s'_t}P(s'_t|s_t,a_t)||x||_\infty\sum_{a'_t,\eta'_{t+1}}\left|\frac{d\pi_2^\alpha(a'_t,\eta'_{t+1}|s'_t,\eta_{t+1})}{d\alpha}|_{\alpha=0}\right| \\
&\leq \sum_{s'_t}P(s'_t|s_t,a_t)||x||_\infty C_1 \\
&\leq C_1||x||_\infty
\end{aligned}$$

By the definition of ℓ_∞ norm, we have

$$\max_{||u||_2=1}||\frac{d\tilde{P}(\alpha)}{d\alpha}|_{\alpha=0}x||_\infty \leq C_1||x||_\infty$$

Similarly, differentiating $\tilde{P}(\alpha)$ twice with respect to α, we obtain

$$\left[\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}|_{\alpha=0}\right]_{(s_t,\eta_t,a_t,\eta_{t+1})\to(s'_t,\eta'_t,a'_t,\eta'_{t+1})} = \begin{cases} \frac{d^2\pi_2^\alpha(a'_t,\eta'_{t+1}|s'_t,\eta'_t)}{(d\alpha)^2}|_{\alpha=0}P(s'_t|s_t,a_t), \text{ if } \eta'_t=\eta_{t+1} \\ 0, \text{ otherwise.}\end{cases}$$

An identical argument leads to the following result: for arbitrary x,

$$\max_{||u||_2=1}||\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}|_{\alpha=0}x||_\infty \leq C_2||x||_\infty$$

Let $Q^\alpha(s_1,a_1,\eta_2)$ be the corresponding Q-function for policy π^α at state s_1 and action a_1,η_2 and denote $\bar{c}$ as a vector where $\bar{c}(s_t,\eta_t,a_t,\eta_{t+1})=\frac{\lambda}{\alpha}[C(s_t,a_t)-\eta_t]_+ +(1-\lambda)C(s_t,a_t)+\gamma\lambda\eta_{t+1}$. Observe that $Q^\alpha(s_1,a_1,\eta_2)$ can be written as

$$Q^\alpha(s_1,a_1,\eta_2) \overset{(a)}{=} C(s_1,a_1)+\gamma\lambda\eta_2+e^\intercal_{(s_1,\eta_1,a_1,\eta_2)}(\sum_{n=1}^\infty\gamma^n\tilde{P}(\alpha)^n)\bar{c}$$

$$
\begin{aligned}
&= C(s_1,a_1) + \gamma\lambda\eta_2 + e^{\mathsf{T}}_{(s_1,\eta_1,a_1,\eta_2)}(\sum_{n=0}^{\infty}\gamma^n \tilde{P}(\alpha)^n)\bar{c} - \bar{c}(s_1,\eta_1,a_1,\eta_2) \\
&\overset{(b)}{=} C(s_1,a_1) + \gamma\lambda\eta_2 + e^{\mathsf{T}}_{(s_1,\eta_1,a_1,\eta_2)} M(\alpha)\bar{c} - \bar{c}(s_1,\eta_1,a_1,\eta_2) \qquad (19)
\end{aligned}
$$

where in (a), η_1 is a dummy variable that can take any value in $\mathcal{H}$ and $e_{(s_1,\eta_1,a_1,\eta_2)}$ is a vector that takes value of 1 at (s_1,η_1,a_1,η_2) and 0 for all other entries; in (b), $M(\alpha) := (\mathbf{I} - \gamma\tilde{P}(\alpha))^{-1} = \sum_{n=0}^{\infty}\gamma^n\tilde{P}(\alpha)^n$ because of power series expansion of matrix inverse. Now differentiating twice with respect to α gives

$$
\begin{aligned}
\frac{dQ^{\alpha}(s_1,a_1,\eta_2)}{d\alpha} &= \gamma e^{\mathsf{T}}_{(s_1,\eta_1,a_1,\eta_2)} M(\alpha)\frac{d\tilde{P}(\alpha)}{d\alpha}M(\alpha)\bar{c}, \\
\frac{d^2Q^{\alpha}(s_1,a_1,\eta_2)}{(d\alpha)^2} &= 2\gamma^2 e^{\mathsf{T}}_{(s_1,\eta_1,a_1,\eta_2)} M(\alpha)\frac{d\tilde{P}(\alpha)}{d\alpha}M(\alpha)\frac{d\tilde{P}(\alpha)}{d\alpha}M(\alpha)\bar{c} + \gamma e^{\mathsf{T}}_{(s_1,\eta_1,a_1,\eta_2)} M(\alpha)\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}M(\alpha)\bar{c}.
\end{aligned}
$$

Because $M(\alpha) \geq 0$ (componentwise) and $M(\alpha)\mathbf{1} = \frac{1}{1-\gamma}\mathbf{1}$, i.e., each row of $M(\alpha)$ is positive and sums to $\frac{1}{1-\gamma}$, we have

$$
\max_{||u||_2=1} ||M(\alpha)x||_\infty \leq \frac{1}{1-\gamma}||x||_\infty
$$

According to Eq. (19), we have

$$
\begin{aligned}
|Q^{\alpha}(s_1,a_1,\eta_2)| &= \left| -\frac{\lambda}{\alpha}[C(s_1,a_1)-\eta_1]_+ + \lambda C(s_1,a_1) + e^{\mathsf{T}}_{(s_1,\eta_1,a_1,\eta_2)}M(\alpha)\bar{c} \right| \\
&\overset{(a)}{\leq} \frac{\lambda}{\alpha} + \lambda + \frac{1}{1-\gamma}||\bar{c}||_\infty \qquad (20)
\end{aligned}
$$

where (a) is true because $C(s_1,a_1) \in [0,1],\ \eta_1 \in [0,1]$. Furthermore, we obtain

$$
\begin{aligned}
\max_{||u||_2=1}\left|\frac{dQ^{\alpha}(s_1,a_1,\eta_2)}{d\alpha}|_{\alpha=0}\right| &\leq ||\gamma M(\alpha)\frac{d\tilde{P}(\alpha)}{d\alpha}M(\alpha)\bar{c}||_\infty \\
&\leq \frac{\gamma C_1}{(1-\gamma)^2}||\bar{c}||_\infty \\
\max_{||u||_2=1}\left|\frac{d^2Q^{\alpha}(s_1,a_1,\eta_2)}{(d\alpha)^2}|_{\alpha=0}\right| &\leq 2\gamma^2||M(\alpha)\frac{d\tilde{P}(\alpha)}{d\alpha}M(\alpha)\frac{d\tilde{P}(\alpha)}{d\alpha}M(\alpha)\bar{c}||_\infty + \gamma||M(\alpha)\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}M(\alpha)\bar{c}||_\infty \\
&\leq \left(\frac{2\gamma^2C_1^2}{(1-\gamma)^3} + \frac{\gamma C_2}{(1-\gamma)^2}\right)||\bar{c}||_\infty
\end{aligned}
$$

Consider the equation

$$
\tilde{J}(\alpha) = \sum_{a_1,\eta_2}\pi_1^{\alpha}(a_1,\eta_2|s_1)Q^{\alpha}(s_1,a_1,\eta_2)
$$

By differentiating $\tilde{J}(\alpha)$ twice with respect to α, we get

$$
\frac{d^2\tilde{J}(\alpha)}{(d\alpha)^2} = \sum_{a_1,\eta_2}\frac{d^2\pi_1^{\alpha}(a_1,\eta_2|s_1)}{(d\alpha)^2}Q^{\alpha}(s_1,a_1,\eta_2) + 2\sum_{a_1,\eta_2}\frac{d\pi_1^{\alpha}(a_1,\eta_2|s_1)}{d\alpha}\frac{dQ^{\alpha}(s_1,a_1,\eta_2)}{d\alpha} + \sum_{a_1,\eta_2}\pi_1^{\alpha}(a_1,\eta_2|s_1)\frac{d^2Q^{\alpha}(s_1,a_1,\eta_2)}{(d\alpha)^2}
$$

Thus,

$$
\begin{aligned}
\max_{||u||_2=1}\left|\frac{d^2\tilde{J}(\alpha)}{(d\alpha)^2}|_{\alpha=0}\right| &\leq C_2(\frac{\lambda}{\alpha} + \lambda + \frac{1}{1-\gamma}||\bar{c}||_\infty) + \frac{2\gamma C_1^2}{(1-\gamma)^2}||\bar{c}||_\infty + \frac{2\gamma^2C_1^2}{(1-\gamma)^3}||\bar{c}||_\infty + \frac{\gamma C_2}{(1-\gamma)^2}||\bar{c}||_\infty \\
&\leq C_2(\frac{\lambda}{\alpha} + \lambda + \frac{||\bar{c}||_\infty}{(1-\gamma)^2}) + \frac{2\gamma C_1^2}{(1-\gamma)^3}||\bar{c}||_\infty.
\end{aligned}
$$

This completes the proof. □

E. Standard optimization results

In this section, we present standard optimization results from Ghadimi & Lan (2016); Beck (2017). We consider solving the following optimization problem

$$\min_{x \in C} f(x)$$

where C is a nonempty closed and convex set, f is proper and closed, $dom(f)$ is convex, and f is σ-smooth over $int(dom(f))$.

Definition E.1. We define the gradient mapping $G^\beta(x)$ as

$$G^\beta(x) = \frac{1}{\beta}(x - P_C(x - \beta \nabla_x f(x)))$$

where P_C is the projection onto C. Note that when $C = \mathbb{R}^d$, the gradient mapping $G^\beta(x) = \nabla f(x)$.

Theorem E.2 (Theorem 10.15 in (Beck, 2017))**.** *Let $x^{(t)} = x^{(t-1)} - \beta G^\beta(x^{(t-1)})$ with the stepsize $\beta = 1/\sigma$. Then*

1. *The sequence $\{f(x^{(t)})\}_{t \geq 0}$ is non-increasing.*
2. *$G^\beta(x^{(t)}) \to 0$ as $t \to \infty$.*
3. *$\min_{t=0,1,\ldots,T-1} ||G^\beta(x^{(t)})||_2 \leq \frac{\sqrt{2\sigma(f(x^{(0)}) - f(x^*))}}{\sqrt{T}}$*

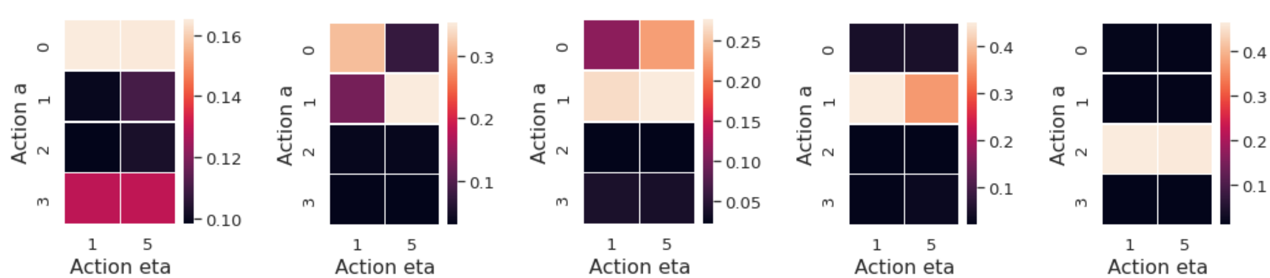
Lemma E.3 (Lemma 3 in (Ghadimi & Lan, 2016))**.** *Let $x^+ = x - \beta G^\beta(x)$. If $||G^\beta(x)||_2 \leq \epsilon$, then*

$$-\nabla f(x^+) \in N_C(x^+) + \epsilon(\beta\sigma + 1)B_2$$

where B_2 is the unit ℓ_2 ball, and N_C is the normal cone of the set C.

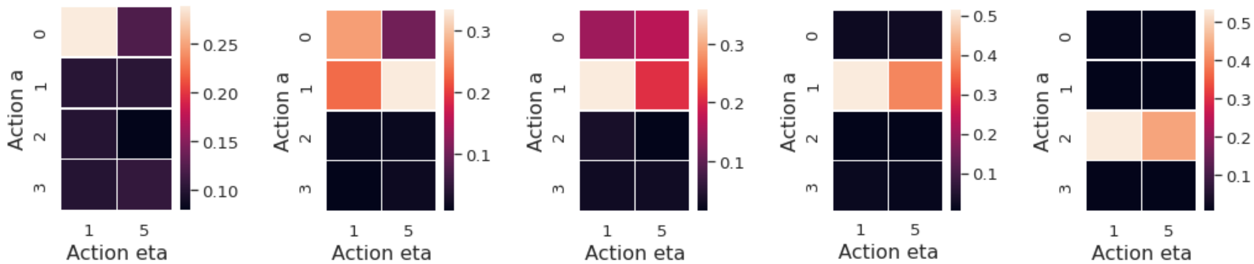
F. Additional computational results

In this section, we present the average action probabilities at each state over the 10 independent simulation runs with varying λ-value in Figures 7–11.



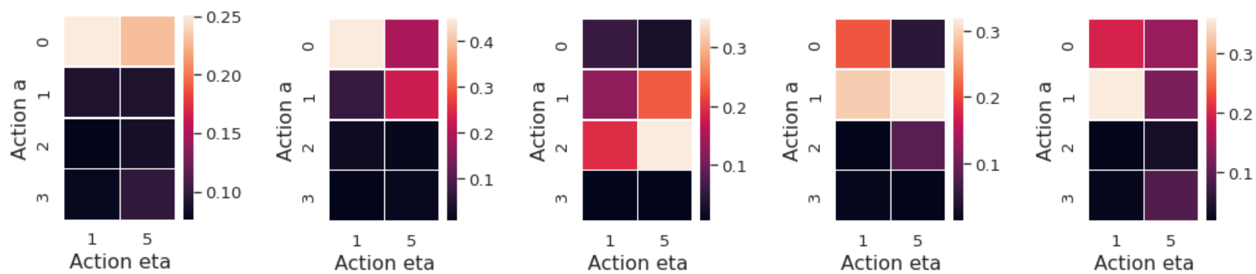
(a) $s_t = (3, 0)$ (b) $s_t = (2, 0),\ \eta_t = 1$ (c) $s_t = (2, 1),\ \eta_t = 5$ (d) $s_t = (2, 2),\ \eta_t = 5$ (e) $s_t = (2, 3),\ \eta_t = 1$

Figure 7. For $\lambda = 0$, the learned optimal path is $[3, 0] \to [2, 0] \to [2, 1] \to [2, 2] \to [2, 3] \to [3, 3]$.

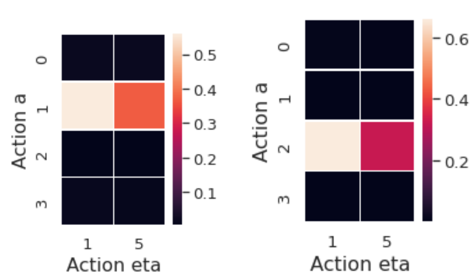


(a) $s_t = (3, 0)$ (b) $s_t = (2, 0),\ \eta_t = 1$ (c) $s_t = (2, 1),\ \eta_t = 5$ (d) $s_t = (2, 2),\ \eta_t = 1$ (e) $s_t = (2, 3),\ \eta_t = 1$

Figure 8. For $\lambda = 0.25$, the learned optimal path is $[3, 0] \rightarrow [2, 0] \rightarrow [2, 1] \rightarrow [2, 2] \rightarrow [2, 3] \rightarrow [3, 3]$.

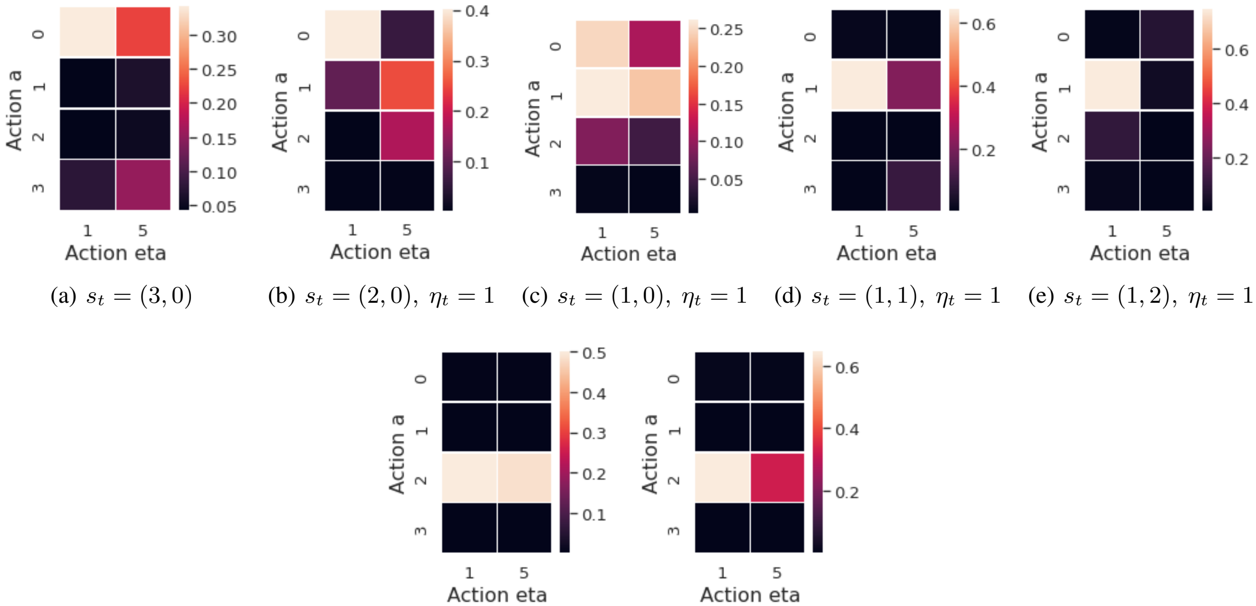


(a) $s_t = (3, 0)$ (b) $s_t = (2, 0),\ \eta_t = 1$ (c) $s_t = (1, 0),\ \eta_t = 1$ (d) $s_t = (2, 0),\ \eta_t = 5$ (e) $s_t = (2, 1),\ \eta_t = 5$



(f) $s_t = (2, 2),\ \eta_t = 1$ (g) $s_t = (2, 3),\ \eta_t = 1$

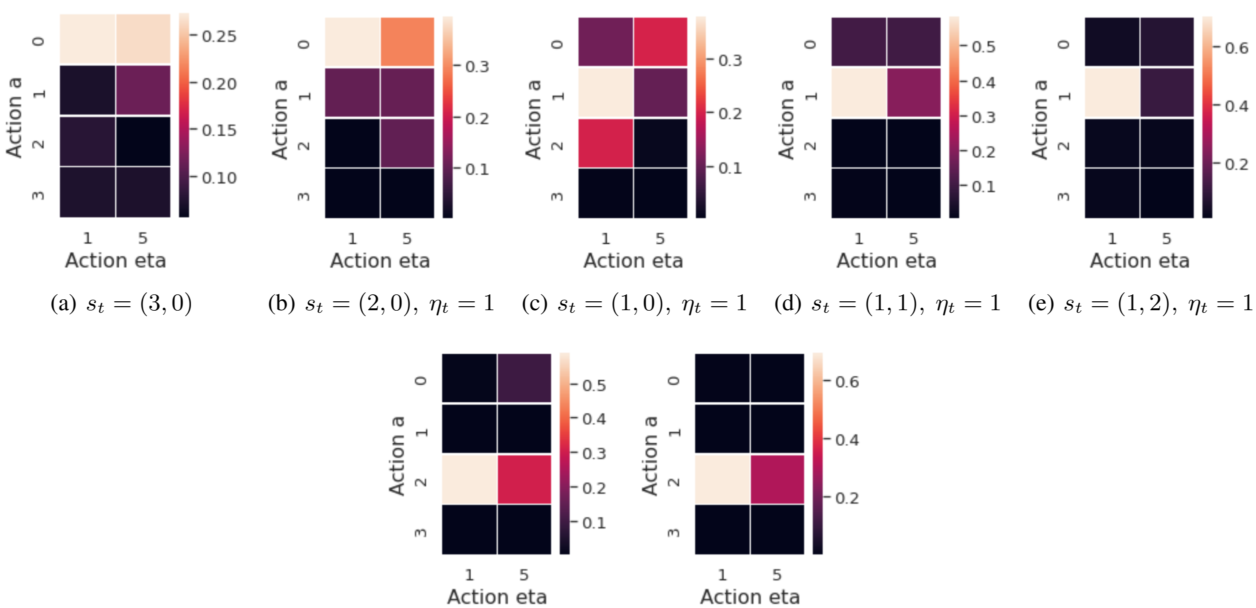
Figure 9. For $\lambda = 0.5$, the learned optimal path is $[3, 0] \rightarrow [2, 0] \rightarrow [1, 0] \rightarrow [2, 0] \rightarrow [2, 1] \rightarrow [2, 2] \rightarrow [2, 3] \rightarrow [3, 3]$.



(a) $s_t = (3, 0)$ (b) $s_t = (2, 0),\ \eta_t = 1$ (c) $s_t = (1, 0),\ \eta_t = 1$ (d) $s_t = (1, 1),\ \eta_t = 1$ (e) $s_t = (1, 2),\ \eta_t = 1$

(f) $s_t = (1, 3),\ \eta_t = 1$ (g) $s_t = (2, 3),\ \eta_t = 1$

Figure 10. For $\lambda = 0.75$, the learned optimal path is $[3, 0] \rightarrow [2, 0] \rightarrow [1, 0] \rightarrow [1, 1] \rightarrow [1, 2] \rightarrow [1, 3] \rightarrow [2, 3] \rightarrow [3, 3]$.



(a) $s_t = (3, 0)$ (b) $s_t = (2, 0),\ \eta_t = 1$ (c) $s_t = (1, 0),\ \eta_t = 1$ (d) $s_t = (1, 1),\ \eta_t = 1$ (e) $s_t = (1, 2),\ \eta_t = 1$

(f) $s_t = (1, 3),\ \eta_t = 1$ (g) $s_t = (2, 3),\ \eta_t = 1$

Figure 11. For $\lambda = 1$, the learned optimal path is $[3, 0] \rightarrow [2, 0] \rightarrow [1, 0] \rightarrow [1, 1] \rightarrow [1, 2] \rightarrow [1, 3] \rightarrow [2, 3] \rightarrow [3, 3]$.