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
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.

**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 (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 – ECPRMs, first introduced by Homem-de Mello & Pagnoncelli (2016). We would like to note that, although Yu & Shen (2022) have shown the ECRM-based riskaverse 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.

### Table 1. Iteration complexity comparison between the risk-neutral results in Agarwal et al. (2021) and our risk-averse setting, where $S, A$ are the state and action space, $H$ is the space of an auxiliary variable $\eta$, $\gamma \in (0, 1)$ is the discount factor, $\epsilon$ is the optimality gap, $D_{\infty} = \|\frac{d_{\text{stationary}}}{\mu}\|_{\infty}$ is used in Agarwal et al. (2021), and $D_1 = \max \{\|\frac{d_{\text{stationary}}}{\mu}\|_{\infty}, \frac{1}{1-\gamma}\|\frac{d_{\text{stationary}}}{\mu}\|_{\infty}\}$.

<table>
<thead>
<tr>
<th>Iteration Complexity</th>
<th>Direct Parameter</th>
<th>Softmax Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agarwal et al. (2021) (Risk-neutral)</td>
<td>$O\left(\frac{D^2_{\infty}</td>
<td>S</td>
</tr>
<tr>
<td>Our work (Risk-averse)</td>
<td>$O\left(\frac{D^2_{\infty}</td>
<td>S</td>
</tr>
</tbody>
</table>

### 2. Preliminaries

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

A stationary Markov policy $\pi^\theta : S \rightarrow \Delta(A)$ parameterized by $\theta$ specifies a probability distribution over the action space given each state $s \in S$, where $\Delta(\cdot)$ denotes the probability simplex, i.e., $0 \leq \pi^\theta(a|s) \leq 1$, $\sum_{a \in A} \pi^\theta(a|s) = 1$, $\forall s \in S$, $a \in A$. A policy induces a distribution over trajectories $\{s_1, a_1, C(s_1, a_1), s_2, a_2, \ldots\}$, where $s_1$ is drawn from the initial state distribution $\rho$, 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 $V^\pi^\theta : S \rightarrow \mathbb{R}$ is defined as the discounted sum of future costs starting at state $s$ and executing $\pi$, i.e., $V^\pi^\theta(s) = \mathbb{E}_s[\sum_{t=1}^{\infty} \gamma^{t-1} C(s_t, a_t) | \pi^\theta, s_1 = s]$. We overload the notation and define $V^\pi^\theta(\rho)$ as the expected value under initial state distribution $\rho$, i.e., $V^\pi^\theta(\rho) = \mathbb{E}_{s_1 \sim \rho} [V^\pi^\theta(s_1)]$. The action-value (or Q-value) function $Q^\pi^\theta : S \times A \rightarrow \mathbb{R}$ is defined as $Q^\pi^\theta(s, a) = \mathbb{E}_s[\sum_{t=1}^{\infty} \gamma^{t-1} C(s_t, a_t) | \pi^\theta, s_1 = s, a_1 = a]$.

In a risk-neutral RL framework, the goal of the agent is to find a policy $\pi^\theta$ that minimizes the expected total cost from the initial state, i.e., the agent seeks to solve $\min_{\theta \in \Theta} V^{\pi^\theta}$ 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 $\pi^{\ast}$ that simultaneously minimizes $V^{\pi^\theta}$ for all states $s_1 \in S$. It is worth noting that $V^{\pi^\theta}(s)$ is non-convex in $\theta$, 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 $\theta$ of the policy in the gradient descent direction. Before introducing the policy gradient methods, we first define the discounted state visitation distribution $d_{\pi}^s$ of a policy $\pi$ as $d_{\pi}^s(s) = (1 - \gamma) \sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{P}^{\pi}\left(s_t = s | s_1\right)$, where $\mathbb{P}^{\pi}\left(s_t = s | s_1\right)$ is the probability that $s_t = s$ after executing $\pi$ starting from state $s_1$. Correspondingly, we define the discounted state visitation distribution under initial distribution $\rho$ as $d_{\rho}^s(s) = \mathbb{E}_{s_1 \sim \rho}[d_{\pi}^s(s)]$.

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

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

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 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 $\mathcal{F}_1 = \{0, \Xi\}$, and thus $Z_1$ is in fact deterministic. Let $\mathcal{L}$ denote a space of $\mathcal{F}_t$-measurable functions from $\Xi$ to $\mathbb{R}$.

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

For ease of presentation, we rewrite $C(s_t, a_t) = c_t$ 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 risk measures $\rho_t^{[s_1, t-1]}$ mapping from $Z_t$ to $Z_{t-1}$, which is a convex combination of conditional expectation and Conditional Value-at-Risk (CVaR):

$$\rho_t^{[s_1, t-1]}(c_t) = (1 - \lambda) \mathbb{E}[c_t | s_1, t-1] + \lambda \text{CVaR}_\alpha[c_t | s_1, t-1],$$

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 $\alpha = 0$.

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

$$\text{CVaR}_\alpha[c_t | s_1, t-1] := \min_{\eta_t \in \mathcal{H}} \left\{ \eta_t + \frac{1}{\alpha} \mathbb{E}[|c_t - \eta_t| | s_1, t-1] \right\},$$

where $[\alpha] := \max\{a, 0\}$, and $\eta_t$ 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 $\alpha$-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 $\alpha$ value makes CVaR sensitive to rare but very high costs. Because $c_t \in [0, 1]$, we can restrict the $\eta$-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.](image)

2.3. Expected Conditional Risk Measures

We consider a class of multi-period risk function $\mathbb{F}$ mapping from $Z_{1, \infty} := Z_1 \times Z_2 \times \cdots \times Z_t \times \cdots$ to $\mathbb{R}$ as follows:

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

where $\rho_t^{[s_{1, t-1}]}$ is the coherent one-step conditional risk measure mapping from $Z_t$ to $Z_{t-1}$ 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 state $s_{t-1}$. This class of multi-period risk measures is called expected conditional risk measures (ECRMs), first introduced in (Homem-de Mello & Pagnottoni, 2016).

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

$$
\min_{a_{1:T}} \mathbb{E} \left\{ C_{[1, T]}(s_1) \right\} =
\min_{a_{1:T}} \left\{ C_{[1, T]}(s_1, a_1) + \gamma \mathbb{E}_{a_2} \left[ \min_{a_{3:T}} \left\{ \frac{\lambda}{\alpha} [C_{[1, T]}(s_2, a_2) - \eta_2] + (1 - \lambda) C_{[1, T]}(s_2, a_2) + \gamma \mathbb{E}_{a_3} \left[ \min_{a_{4:T}} \left\{ \frac{\lambda}{\alpha} [C_{[1, T]}(s_3, a_3) - \eta_3] + (1 - \lambda) C_{[1, T]}(s_3, a_3) + \gamma \mathbb{E}_{a_4} \left[ \min_{a_{5:T}} \left\{ \frac{\lambda}{\alpha} [C_{[1, T]}(s_4, a_4) - \eta_4] + \cdots \right\} \right] \right] \right] \right] \right] \right\},
$$

where $\mathbb{E}_{a_{s_{t-1}}} = \mathbb{E}_{s_{t-1}}[|s_{t-1}|]$ is the conditional expectation and we apply the Markov property to recast $\mathbb{E}_{a_{s_{t-1}}}^{s_{t-1}}$ as $\mathbb{E}_{s_{t-1}^{s_{t-1}}}$. The auxiliary variable $\eta_t$ from Eq. (2) is decided before taking conditional expectation $\mathbb{E}_{s_{t-1}^{s_{t-1}}}$ and thus it can be regarded as a $(t-1)$-stage action, similar to $a_{t-1}$. Here, $\eta_t$ denotes the tail information of state $s_{t}$’s immediate cost (i.e., $\eta_0 = \text{Var}_s[C(s_t, a_t)|s_{1:t-1}]$), 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 $a_t \in A$ to be $(a_t, \eta_{t+1}) \in A \times \mathcal{H}$ 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 $C(s_t, a_t)$ with $C_t = C(s_t, a_t, \eta_t) = C(s_t, a_t) + \gamma \mathbb{E}_{a_{t+1}}[C(s_{t+1}, a_{t+1})]$) and replacing $C(s_t, a_t, \eta_t, \eta_{t+1})$ with $C_t = C(s_t, a_t, \eta_t, \eta_{t+1}) = C(s_t, a_t) + \gamma \mathbb{E}_{a_{t+1}}[C(s_{t+1}, a_{t+1})]$. This leads to the following proposition.

**Proposition 2.2.** Define the augmented action space as $\mathcal{A} = A \times \mathcal{H}$, augmented state space as $\mathcal{S} = S \times \mathcal{H}$, and the state-action transition matrix under policy $\pi$ as $\bar{P}(s_t, a_t, \eta_{t+1}) = P(s_{t+1}|s_t, a_t)\pi(s_t, a_t, \eta_{t+1})$. If $s_t = s_{t+1}$; otherwise, $\bar{P}(s_t, a_t, \eta_{t+1}) = P(s_{t+1}|s_t, a_t)\pi(s_t, a_t, \eta_{t+1})$. Then the risk-averse RL with ECRM-based objective function (4) is equivalent to a risk-neutral RL with $\hat{M} = (\mathcal{S}, \mathcal{A}, \mathcal{C}, \bar{P}, \gamma, \rho)$. 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 $\hat{C}_t(s_t, a_t, \eta_t)$ in the first step has a different form than $C_t(s_t, a_t, \eta_t)$ 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 $\gamma \mathbb{E}_{a_{t+1}}[C(s_{t+1}, a_{t+1})]$ 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 $\pi^\rho = (\pi_1^\rho, \pi_2^\rho) \in \Delta(A \times H)|S| + |H|$ for $\pi_1^\rho(a_1, \eta_2|s_1, \eta_1)$ as the policy for the first time step parameterized by $\theta_1$ and $\pi_2^\rho(a_2, \eta_{t+1}|s_t, \eta_t)$, $\forall t \geq 2$ is the stationary policy for the following time steps parameterized by $\theta_2$. We omit the dependence of $\pi$ on $\theta$ in the following for ease of presentation. The goal is to solve the optimization problem below

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

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

$$
J^\rho(\pi) = \mathbb{E}_{s_1 \sim \rho} \left[ J^\rho(s_1) \right] = \mathbb{E}_{s_1 \sim \rho} \mathbb{E}_{(a_1, \eta_2) \sim \pi_1(.|s_1)} \left[ Q^{\pi_2}(s_1, a_1, \eta_2) \right]
$$

and

$$
Q^{\pi_2}(s_1, a_1, \eta_2) = C(s_1, a_1) + \gamma \mathbb{E}_{a_2} \left[ \min_{a_3} \left\{ \frac{\lambda}{\alpha} [C(s_2, a_2) - \eta_2] + (1 - \lambda) C(s_2, a_2) + \gamma \mathbb{E}_{a_3} \left[ \min_{a_4} \left\{ \frac{\lambda}{\alpha} [C(s_3, a_3) - \eta_3] + (1 - \lambda) C(s_3, a_3) + \gamma \mathbb{E}_{a_4} \left[ \min_{a_5} \left\{ \frac{\lambda}{\alpha} [C(s_4, a_4) - \eta_4] + \cdots \right\} \right] \right] \right] \right] \right].
$$

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

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

where

$$
\hat{Q}^{\pi_2}(s_t, \eta_t, a_t, \eta_{t+1}) = \frac{\lambda}{\alpha} [C(s_t, a_t) - \eta_t] + \cdots
$$
The policy gradients of risk-averse $\pi$ are given by:

$$
\nabla_{\theta} \log \pi_{\tau}^{\pi}(a_t | s_t, \eta_{t+1}) \frac{Q_{\pi}^{\tau_2}(s_{t+1}, \eta_{t+1})}{\rho_{\pi}(s, \eta)}
$$

where $\rho_{\pi}(s, \eta) = \sum_{s_1} \rho(s_1)Q_{\pi}^{\tau_1}(s_2 = s, \eta_2 = \eta | s_1)$.

The differences between risk-averse ECRM-based and risk-neutral policy gradients are twofold. First, we break the policy parameters into two parts, $\theta_1$ and $\theta_2$, and derive the gradient for each one separately. Second, as reflected in the state visitation distribution $Q_{\pi}^{\tau_1}(s_2 = s, \eta_2 = \eta | s_1)$, different from the risk-neutral gradients where the initial state is $s_1$ with distribution $\rho$, and the state visitation distribution is $\tilde{d}_{\rho}^{\tau}$.

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 $\nabla_{\pi} J_{\pi}^{\tau}(\rho)$ instead of $\nabla_{\theta} J_{\pi}^{\tau}(\rho)$, and the gradients are

$$
\frac{\partial J_{\pi}^{\tau}(\rho)}{\partial \pi_1(a, \eta | s)} = \rho(s)Q_{\pi}^{\tau}(s, a, \eta)
$$

$$
\frac{\partial J_{\pi}^{\tau}(\rho)}{\partial \pi_2(a, \eta' | s, \eta)} = \frac{\gamma}{1 - \gamma} \tilde{d}_{\rho}^{\tau}(s, \eta, a, \eta')
$$

using Theorem 3.2. Next, we show that the objective function $J_{\pi}^{\tau}(s_1)$ 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 $s_1$, $J_{\pi}^{\tau}(s_1)$ is $(\frac{2|\mathcal{A}| |\mathcal{H}|}{(1 - \gamma)^2} ||\bar{c}||_{\infty}$-smooth in $\pi$, i.e.,

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

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

However, smoothness alone can only guarantee the convergence of the gradient descent method to a stationary point (i.e., $\nabla_{\pi} J_{\pi}^{\tau}(s_1) = 0$). 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 $(\nu, \mu)$-gradient dominated over $\Theta$ if there exists constants $\nu > 0$ and $\mu \geq 0$ such that for all $\theta \in \Theta$,
\[
\min_{\theta' \in \Theta} f(\theta') \geq f(\theta) + \min_{\theta' \in \Theta} \left[ \nu \langle \nabla f(\theta), \theta' - \theta \rangle + \frac{\mu}{2} ||\theta - \theta'||^2 \right].
\]

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 $\theta$, 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 $\theta$, 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 $\mu$ to our Theorem 3.5, we not only need a strictly positive barrier regularizer in the next section, which ensures 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 $\pi^{LB}_1 \rightarrow 0$, the iteration bound $T \rightarrow +\infty$. To circumvent this issue, we consider a softmax parameterization with log barrier regularizer in the next section, which ensures that $\pi^{LB}_1 > 0$.

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
\[
\pi^{(t+1)} = P_{\Delta(A \times H)^{|S|} + |S||H|}(\pi^{(t)} - \beta \nabla \nabla J^{\pi(t)}(\mu))
\]
where $P_{\Delta(A \times H)^{|S|} + |S||H|}$ is the projection onto $\Delta(A \times H)^{|S|} + |S||H|$ in the Euclidean norm, and $\beta$ is the step size. Using Theorem 3.5, we now give an iteration complexity bound for projected gradient descent methods.

**Theorem 3.7.** Let $\pi^{LB}_1 = \inf_{1 \geq t} \min_{s,a,\eta} \pi^{(t)}_1(a, \eta|s)$, $D_1 = \max\{||\rho||_\infty, \frac{1}{(1-\gamma)\bar{\lambda}} ||\frac{\partial^2 \pi^{LB}_1}{\partial s \partial a}||_\infty \}$ and $\mu^B(s, \eta) = \sum_{s_1,a_1} \mu(s_1)P(s|s_1, a_1), \forall s \in S, \eta \in H$.

Note that the significance of Theorem 3.5 is that although the iteration bound
\[
T \geq D_1^2 \frac{128\gamma|A||H|^2}{\lambda} \frac{C||\bar{c}||_\infty}{(1-\gamma)\bar{\lambda}^2}
\]
with $\bar{C} = \left( \frac{\lambda}{\alpha} + \lambda + \frac{1}{1-\gamma} ||\bar{c}||_\infty \right)$, $||\bar{c}||_\infty = \frac{\lambda}{\alpha} + (1-\lambda) + \gamma \lambda$.

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 $\pi^{LB}_1 \rightarrow 0$, the iteration bound $T \rightarrow +\infty$. To circumvent this issue, we consider a softmax parameterization with log barrier regularizer in the next section, which ensures that $\pi^{LB}_1 > 0$.

### 3.2. Unconstrained Softmax Parameterization

In this section, we aim to solve the optimization problem (5) with the following softmax parameterization: for all $s_1, a_1, \eta_2$ and $s_1, \eta_1, a_t, \eta_{t+1}$, $t \geq 2$, we have
\[
\pi^{\theta_1}_1(a_1, \eta_2|s_1) = \frac{\exp(\theta_1(s_1, a_1, \eta_2))}{\sum_{a_1', \eta_2'} \exp(\theta_1(s_1, a_1', \eta_2'))}
\]
\[
\pi^{\theta_2}_t(a_t, \eta_{t+1}|s_1, \eta_t) = \frac{\exp(\theta_2(s_t, \eta_t, a_t, \eta_{t+1}))}{\sum_{a_t', \eta_{t+1}'} \exp(\theta_2(s_t, \eta_t, a_t', \eta_{t+1}'))}
\]

Note that the softmax parameterization is preferable to the direct parameterization, since the parameters $\theta$ are uncon-
strained (i.e., $\pi_1^0, \pi_2^0$ 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:

$$\frac{\partial J^\pi_\theta}{\partial \theta_1(s_t, a_t, \eta_t)} = \mu(s_t) \pi_\theta(a_t | s_t, \eta_t)^2 \left(-A^\pi_\theta(s_t, a_t, \eta_t)\right)$$

$$\frac{\partial J^\pi_\theta}{\partial \theta_2(s_t, \eta_t, a_t, \eta_{t+1})} = \frac{\gamma}{1 - \gamma} d_{\mu}^\pi(s_t, \eta_t)^2 \pi_\theta(a_t, \eta_{t+1} | s_t, \eta_t)$$

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

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

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

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 $\kappa$ is small enough.

**Theorem 3.10.** Let $\pi_1^{LB} = \min_{s,a,\eta} \pi_\theta^0(a, \eta | s)$. Suppose

$$\|\nabla_\theta L_\kappa(\theta)\|_2 \leq \epsilon_1 \leq \frac{\kappa}{2|S||A||H|},$$

$$\|\nabla_\theta^2 L_\kappa(\theta)\|_2 \leq \epsilon_2 \leq \frac{\kappa}{2|S||H||A||H|},$$

then we have that for all starting state distributions $\rho$:

$$J^\pi(\rho) - J^*(\rho) \leq 2\kappa \|\rho\|_\infty + \frac{2\epsilon}{1-\gamma} \|\pi_1^{LB}\|_\infty \|\rho\|_\infty.$$
We train the model over 10000 episodes where each episode All the hyperparameters are kept the same across different
different runs and plot the test cost in the last 1000 episodes
values 1 or 5, we set the
test cost in the first 1500 episodes, after which the policy
in Figure 3(b).

![Image](image1.png)

```
Figure 4. 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.

```
Algorithm 1 Risk-averse REINFORCE with softmax parameterization
1: Initialize \( \theta_1(s_1, a_1, \eta_2) \), \( \forall s_1 \in S \), \( a_1 \in A \), \( \eta_2 \in H \) and set softmax policy \( \pi_1^\theta(a_1, \eta_2|s_1) = \frac{\exp(\theta_1(s_1, a_1, \eta_2))}{\sum_{a_1'} \exp(\theta_1(s_1, a_1', \eta_2))} \).
2: Initialize \( \theta_2(s_t, \eta_t, a_t, \eta_{t+1}), \forall s_t \in S \), \( a_t \in A \), \( \eta_t \in H \) and set softmax policy \( \pi_2^\theta(a_t, \eta_{t+1}|s_t, \eta_t) = \frac{\exp(\theta_2(s_t, a_t, \eta_t, \eta_{t+1}))}{\sum_{a_t'} \exp(\theta_2(s_t, a_t', \eta_t, \eta_{t+1}))} \).
3: while not converged do
4: Generate one trajectory on policy \( \pi^\theta = (\pi_1^\theta, \pi_2^\theta): s_1, a_1, \eta_2, s_1, a_2, \ldots, s_{T-1}, a_{T-1}, \eta_{T-1}, s_T \), \( \eta_{T-1}, \eta_{T+1} \in \eta_{T-1}, \eta_{T+1} \in \eta \).
5: Modify immediate costs as \( c_1 = c_1 + \gamma \lambda \eta_{T-1}, c_1 = \lambda [c_1 - \eta_1] + (1 - \lambda) c_1 + \gamma \lambda \eta_{T+1}, \forall t \geq 2. \)
6: Update \( \theta_1 := \theta_1 - \beta \nabla_{\theta_1} \log \pi_1^\theta(a_1, \eta_2|s_1) \sum_{T=1}^{T-1} \gamma^{T-1} e_T. \)
7: Update \( \theta_2 := \theta_2 - \beta \sum_{t=2}^{T-1} \nabla_{\theta_2} \log \pi_2^\theta(a_t, \eta_{t+1}|s_t, \eta_t) \sum_{T=2}^{T-1} \gamma^{T-1} e_T. \)
8: end while
```

```
Figure 3. Test cost over 10 independent runs with varying \( \lambda \).
```

```
Figure 3. Test cost over 10 independent runs with varying \( \lambda \).
```

We present the average action probabilities at each state in
each state (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 \( \lambda \) 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 \( \lambda \) 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 \( \kappa \) from 0 to 0.5 in Figure 5. A higher \( \kappa \) helps speed up the convergence to a steady path while \( \kappa = 0 \) generates the highest variance of test cost after 2000 episodes.
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 $\pi_1$ and $\pi_2$ in Algorithm 1 with neural networks parameterized by $\theta_1$ and $\theta_2$, respectively. We also construct neural networks for value critics $v_1(s)$ with parameter $w_1$ and value critic $v_2(s, \eta)$ 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 $\lambda$ 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

1: Initialize policy $\pi_1(a_1, \eta_1|s_1)$ with parameter $\theta_1$ and policy $\pi_2(a_t, \eta_{t+1}|s_t, \eta_t)$ with parameter $\theta_2$.
2: Initialize value critic $v_1(s_1)$ with parameter $w_1$ and value critic $v_2(s, \eta)$ with parameter $w_2$.
3: while not converged do
4: Generate one trajectory on policy $\pi^\theta = (\pi_1^\theta_1, \pi_2^\theta_2)$: $s_1, a_1, \eta_1, c_1, s_2, \ldots, s_{T-1}, a_{T-1}, \eta_{T-1}, s_T$.
5: Modify immediate costs as $\bar{c}_t = c_t + \gamma \lambda \eta_t$: $\bar{c}_t = \frac{1}{\alpha}[c_t - (1 - \lambda)c_t + \gamma \lambda \eta_{t+1} + \forall t \geq 2$.
6: Compute discounted costs: $V_t = \sum_{t=1}^{T-1} \gamma^{t-1} \bar{c}_t$ for all $t = 1, \ldots, T - 1$.
7: Update $w_1$ to minimize $||v_1^{\theta_1}(s_1) - V_1||^2$ and update $w_2$ to minimize $\sum_{t=2}^{T-1} ||v_2^{\theta_2}(s_t, \eta_t) - V_t||^2$.
8: Update $\theta_1 := \theta_1 - \beta \nabla_{\theta_1} \log \pi_1^{\theta_1}(a_1, \eta_1|s_1)V_1$.
9: Update $\theta_2 := \theta_2 - \beta \sum_{t=2}^{T-1} \nabla_{\theta_2} \log \pi_2^{\theta_2}(a_t, \eta_{t+1}|s_t, \eta_t)V_t$.
10: end while

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


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

\[ J^\pi(s_1) - J^\pi'(s_1) = J^\pi(s_1) - \mathbb{E}_{\tau \sim P^\tau}[\sum_{t=1}^{\infty} \gamma^{t-1} \tilde{c}_t] \]

\[ = J^\pi(s_1) - \mathbb{E}_{\tau \sim P^\tau}[\sum_{t=1}^{\infty} \gamma^{t-1}(\epsilon_t + J^\pi(s_{t+1}) - J^\pi(s_t))] \]

\[ \overset{(a)}{=} - \mathbb{E}_{\tau \sim P^\tau}[\sum_{t=1}^{\infty} \gamma^{t-1}(\epsilon_t + \gamma J^\pi(s_{t+1}) - J^\pi(s_t))] \]

\[ \overset{(b)}{=} - \mathbb{E}_{\tau \sim P^\tau}[\sum_{t=1}^{\infty} \gamma^{t-1}(\epsilon_t + \gamma E_{s_{t+1}}[J^\pi(s_{t+1})] - J^\pi(s_t))], \]

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

\[ - \mathbb{E}_{(a_1, \eta_2) \sim \pi_1(s_1)}[C(s_1, a_1) + \gamma \lambda \eta_2 + \gamma E_{s_2}^{a_1}[J^\pi(s_2)] - J^\pi(s_1)] \]

\[ = - \mathbb{E}_{(a_1, \eta_2) \sim \pi_1(s_1)}[Q^T(s_1, a_1, \eta_2) + \gamma E_{s_2}^{a_1}[J^\pi(s_2) - \hat{J}^\pi(s_2, \eta_2)] - J^\pi(s_1)] \]

\[ = \mathbb{E}_{(a_1, \eta_2) \sim \pi_1(s_1)}[A^\pi(s_1, a_1, \eta_2) + \gamma E_{s_2}^{a_1}[J^\pi(s_2) - \hat{J}^\pi(s_2)]] \tag{12} \]

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

\[ - \mathbb{E}_{\tau \sim P^\tau}[\gamma^{t-1}\left(\frac{\Lambda}{Q}(C(s_t, a_t) - \eta_t) + (1 - \lambda)C(s_t, a_t) + \gamma \lambda \eta_{t+1} + \gamma E_{s_{t+1}}^{a_t}[J^\pi(s_{t+1})] - J^\pi(s_t)\right)] \]

\[ = - \mathbb{E}_{\tau \sim P^\tau}[\gamma^{t-1}\left(\hat{Q}^T(s_t, \eta_t, a_t, \eta_{t+1}) - \hat{J}^\pi(s_t, \eta_t) + \gamma E_{s_{t+1}}^{a_t}[J^\pi(s_{t+1}) - \hat{J}^\pi(s_{t+1}, \eta_{t+1})] + (\hat{J}^\pi(s_t, \eta_t) - J^\pi(s_t))\right)] \]

\[ = \mathbb{E}_{\tau \sim P^\tau}[\gamma^{t-1}\left(\hat{A}^T(s_t, \eta_t, a_t, \eta_{t+1}) + \gamma E_{s_{t+1}}^{a_t}[\hat{J}^\pi(s_{t+1}, \eta_{t+1}) - J^\pi(s_{t+1})] - (\hat{J}^\pi(s_t, \eta_t) - J^\pi(s_t))\right)]. \tag{13} \]

Observing that \( \mathbb{E}_{s_{t+1}, \eta_{t+1} \sim P^\tau}[J^\pi(s_{t+1}, \eta_{t+1}) - J^\pi(s_{t+1})] = \mathbb{E}_{s_{t+1}, \eta_{t+1} \sim P^\tau}[(\hat{J}^\pi(s_{t+1}, \eta_{t+1}) - J^\pi(s_{t+1})) \text{ for all } t \geq 1] \), 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

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

This completes the proof. \qed
Proof of Theorem 3.2. [Risk-averse policy gradients] According to Eq. (6), we have
\[
\nabla_{\theta_1} J^\pi(s_1) = \nabla_{\theta_1} \left( \sum_{a_1, \eta_2} \pi_1(a_1, \eta_2 | s_1) Q_{\pi_2}^\pi(s_1, a_1, \eta_2) \right)
\]
\[
= \sum_{a_1, \eta_2} \nabla_{\theta_1} \pi_1(a_1, \eta_2 | s_1) Q_{\pi_2}^\pi(s_1, a_1, \eta_2)
\]
\[
= \mathbb{E}_{(a_1, \eta_2) \sim \pi_1} \left[ \nabla_{\theta_1} \log \pi_1(a_1, \eta_2 | s_1) Q_{\pi_2}^\pi(s_1, a_1, \eta_2) \right]
\]
where (a) is true because \( \nabla_{\theta_1} Q_{\pi_2}^\pi(s_1, a_1, \eta_2) = 0 \). 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 \( Q_{\pi_2}^\pi(s_1, a_1, \eta_2) \), we have
\[
\nabla_{\theta_2} J^\pi(s_1) = \nabla_{\theta_2} \left( \sum_{a_1, \eta_2} \pi_1(a_1, \eta_2 | s_1) Q_{\pi_2}^\pi(s_1, a_1, \eta_2) \right)
\]
\[
= \sum_{a_1, \eta_2} \pi_1(a_1, \eta_2 | s_1) \nabla_{\theta_2} \left( c_1 + \gamma \lambda \eta_2 + \gamma \mathbb{E}_{s_2 \sim \pi_2} [J^\pi(s_2, \eta_2)] \right)
\]
\[
= \gamma \sum_{a_1, \eta_2} \pi_1(a_1, \eta_2 | s_1) \sum_{s_2} P(s_2 | s_1, a_1) \nabla_{\theta_2} J^\pi(s_2, \eta_2)
\]
\[
= \gamma \sum_{a_1, \eta_2} \mathbb{P}_{\pi_1}(s_2, \eta_2 | s_1) \nabla_{\theta_2} J^\pi(s_2, \eta_2)
\]

Now for \( \nabla_{\theta_2} J^\pi_2(s_2, \eta_2) \), we have
\[
\nabla_{\theta_2} J^\pi_2(s_2, \eta_2)
\]
\[
= \nabla_{\theta_2} \left( \sum_{a_1, \eta_3} \pi_2(a_2, \eta_3 | s_2, \eta_2) Q_{\pi_2}^\pi_2(s_2, \eta_2, a_2, \eta_3) \right)
\]
\[
= \sum_{a_2, \eta_3} \left( \nabla_{\theta_2} \pi_2(a_2, \eta_3 | s_2, \eta_2) Q_{\pi_2}^\pi_2(s_2, \eta_2, a_2, \eta_3) + \pi_2(a_2, \eta_3 | s_2, \eta_2) \nabla_{\theta_2} \hat{Q}_{\pi_2}^\pi_2(s_2, \eta_2, a_2, \eta_3) \right)
\]
\[
= \sum_{a_2, \eta_3} \left( \nabla_{\theta_2} \pi_2(a_2, \eta_3 | s_2, \eta_2) Q_{\pi_2}^\pi_2(s_2, \eta_2, a_2, \eta_3) + \gamma \sum_{a_2, \eta_3} \pi_2(a_2, \eta_3 | s_2, \eta_2) \sum_{s_3} P(s_3 | s_2, a_2) \nabla_{\theta_2} J^\pi_2(s_3, \eta_3) \right),
\]
where the last equation follows from the definition of \( \hat{Q}_{\pi_2}^\pi(s_2, \eta_2, a_2, \eta_3) \) in Eq. (7). Using a similar argument in risk-neutral policy gradient theorems (Williams, 1992; Sutton et al., 1999) and denoting \( \phi(s_t, \eta_t, a_{t+1}) := \nabla_{\theta_2} \pi_2(a_t, \eta_{t+1} | s_t, \eta_t) \hat{Q}_{\pi_2}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) \), we obtain
\[
\nabla_{\theta_2} J^\pi(s_2, \eta_2)
\]
\[
= \sum_{a_2, \eta_3} \phi(s_2, \eta_2, a_2, \eta_3) + \gamma \sum_{s_3, \eta_3} \mathbb{P}_{\pi_2}(s_3, \eta_3 | s_2, \eta_2) \sum_{a_3, \eta_4} \phi(s_3, \eta_3, a_3, \eta_4)
\]
\[
+ \gamma^2 \sum_{s_4, \eta_4} \mathbb{P}_{\pi_2}(s_4, \eta_4 | s_2, \eta_2) \sum_{a_4, \eta_5} \phi(s_4, \eta_4, a_4, \eta_5) + \cdots
\]
\[
= \frac{1}{1 - \gamma} \mathbb{E}_{(s_t, \eta_t) \sim \pi_{\pi_2}(s_2, \eta_2)} \mathbb{E}_{(s_{t+1}, \eta_{t+1}) \sim \pi_{\pi_2}(s_t, \eta_t)} \left[ \nabla_{\theta_2} \log \pi_2(a_t, \eta_{t+1} | s_t, \eta_t) \hat{Q}_{\pi_2}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) \right]
\]

As a result,
\[
\nabla_{\theta_2} J^\pi(\rho)
\]
\[
= \gamma \sum_{s_2, \eta_2} \sum_{s_1} \rho(s_1) \mathbb{P}_{\pi_1}(s_2, \eta_2 | s_1) \nabla_{\theta_2} J^\pi_2(s_2, \eta_2)
\]
\[
= \gamma \sum_{s_2, \eta_2} \rho(s_2, \eta_2) \nabla_{\theta_2} J^\pi_2(s_2, \eta_2)
\]
\[
= \frac{\gamma}{1 - \gamma} \mathbb{E}_{(s_t, \eta_t) \sim \rho^{s_2, \eta_2}} \mathbb{E}_{(s_{t+1}, \eta_{t+1}) \sim \pi_{\pi_2}(s_t, \eta_t)} \left[ \nabla_{\theta_2} \log \pi_2(a_t, \eta_{t+1} | s_t, \eta_t) \hat{Q}_{\pi_2}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) \right].
\]
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 \( \pi^\alpha := \pi^{\theta + \alpha u} \), \( J(\alpha) := J^{\pi^\alpha}(s_1) \). For direct parameterization, we have \( \pi^\alpha = \pi^{\theta + \alpha u} = \theta + \alpha u \). Differentiating \( \pi^\alpha \) with respect to \( \alpha \) gives

\[
\sum_{\alpha_1, \eta_2} \left| \frac{d\pi^\alpha_1(a_1, \eta_2|s_1)}{da_1} \right| \leq \sum_{\alpha_1, \eta_2} |u_1(s_1, a_1, \eta_2)| \leq \sqrt{\sum_{\alpha_1, \eta_2} |u_1(s_1, a_1, \eta_2)|^2} \sqrt{\sum_{\alpha_1, \eta_2} 1^2} \leq ||u||_2 \sqrt{||A||H} = ||\pi||_2
\]

\[
\sum_{\alpha_1, \eta_2} \left| \frac{d^2\pi^\alpha_1(a_1, \eta_2|s_1)}{(da_1)^2} \right| \leq \sum_{\alpha_1, \eta_2} |u_2(s_1, \eta_1, a_1, \eta_1+1)| \leq \sqrt{\sum_{\alpha_1, \eta_2} |u_2(s_1, \eta_1, a_1, \eta_1+1)|^2} \sqrt{\sum_{\alpha_1, \eta_2} 1^2} \leq ||\pi||_2 \sqrt{||A||H}
\]

\[
\sum_{\alpha_1, \eta_2} \left| \frac{d^2\pi^\alpha_1(a_1, \eta_2|s_1)}{(da_1)^2} \right| = 0, \quad \sum_{\alpha_1, \eta_2} \left| \frac{d^2\pi^\alpha_1(a_1, \eta_2|s_1)}{(da_1)^2} \right| = 0.
\]

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

\[
\max_{|u|_2 = 1} \left| \frac{d^2\tilde{J}(\alpha)}{(da_1)^2} \right| \leq C_2 \left( \frac{\lambda}{\alpha} + \frac{||\tilde{c}||_\infty}{(1-\gamma)^2} \right) + \frac{2\gamma C_2^2}{(1-\gamma)^3} ||\tilde{c}||_\infty
\]

\[
\leq \frac{2\gamma ||A||H}{(1-\gamma)^3} ||\tilde{c}||_\infty
\]

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

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

\[
J^\pi(\rho) - J^\pi(\rho) = \mathbb{E}_{s_1 \sim \rho} \mathbb{E}_{\pi \sim \pi^\rho(s_1)} \left[ \tilde{A}(s_1, a_1, \eta_2) \right] = \sum_{s_1, a_1, \eta_2} \rho(s_1) \pi^\rho(s_1, a_1, \eta_2) \pi^\rho(s_1, a_1, \eta_2)
\]

Then for the first term, we have

\[
\mathbb{E}_{s_1 \sim \rho} \mathbb{E}_{\pi \sim \pi^\rho(s_1)} \left[ A^\pi(s_1, a_1, \eta_2) \right]
\]

\[
= \sum_{s_1, a_1, \eta_2} \rho(s_1) \pi^\rho(s_1, a_1, \eta_2) \pi^\rho(s_1, a_1, \eta_2)
\]

\[
\leq \sum_{s_1} \rho(s_1) \max_{a_1, \eta_2} A^\pi(s_1, a_1, \eta_2)
\]

\[
\leq \frac{\rho(s_1)}{\mu(s_1)} \max_{a_1, \eta_2} A^\pi(s_1, a_1, \eta_2)
\]

\[
\leq ||\rho||_\infty \sum_{s_1} \mu(s_1) \max A^\pi(s_1, a_1, \eta_2)
\]

\[
\leq ||\rho||_\infty \max_{s_1} \sum_{a_1, \eta_2} \mu(s_1) \pi^\rho(a_1, \eta_2|s_1) A^\pi(s_1, a_1, \eta_2)
\]

\[
\leq ||\rho||_\infty \max_{\pi \in \Delta(A \times H)} \sum_{s_1, a_1, \eta_2} \mu(s_1) \pi_1(a_1, \eta_2|s_1) A^\pi(s_1, a_1, \eta_2)
\]

\[
\leq ||\rho||_\infty \max_{\pi \in \Delta(A \times H)} \sum_{s_1, a_1, \eta_2} \mu(s_1) \pi_1(a_1, \eta_2|s_1) A^\pi(s_1, a_1, \eta_2)
\]

\[
\leq ||\rho||_\infty \max_{\pi \in \Delta(A \times H)} \sum_{s_1, a_1, \eta_2} \mu(s_1) \pi_1(a_1, \eta_2|s_1) A^\pi(s_1, a_1, \eta_2)
\]

\[
\leq ||\rho||_\infty \max_{\pi \in \Delta(A \times H)} \sum_{s_1, a_1, \eta_2} \mu(s_1) \pi_1(a_1, \eta_2|s_1) A^\pi(s_1, a_1, \eta_2)
\]

\[
\leq ||\rho||_\infty \max_{\pi \in \Delta(A \times H)} \sum_{s_1, a_1, \eta_2} \mu(s_1) \pi_1(a_1, \eta_2|s_1) A^\pi(s_1, a_1, \eta_2)
\]

\[
\leq ||\rho||_\infty \max_{\pi \in \Delta(A \times H)} (\pi - \tilde{\pi})^T \nabla_\pi J^\pi(\mu)
\]

where (a) is true because \( \max_{s_1, \eta_2} A^\pi(s_1, a_1, \eta_2) = J^\pi(s_1) - \min_{s_1, \eta_2} Q^\pi(s_1, a_1, \eta_2) \geq 0 \); (b) is true because \( \max_{a_1, \eta_2} \) is attained at an action \( (a_1, \eta_2) \) that maximizes \( A^\pi(s_1, \cdot, \cdot) \) for each state \( s_1 \); (c) follows since
On the Global Convergence of Risk-Averse Policy Gradient

\[ \sum_{a_1, \eta_2} \pi_1(a_1, \eta_2 | s_1) A^T(s_1, a_1, \eta_2) = J^r(s_1) - \sum_{a_1, \eta_2} \pi_1(a_1, \eta_2 | s_1) Q^\pi^r(s_1, a_1, \eta_2) = 0; \]  
\[ (d) \text{ follows since } \sum_{a_1, \eta_2} (\tilde{\pi}_1(a_1, \eta_2 | s_1) - \pi_1(a_1, \eta_2 | s_1)) J^r(s_1) = 0; \text{ and } (e) \text{ follows from Theorem 3.2 and Eq. (9)}. \]

For the second term, we have

\[
\begin{align*}
& \mathbb{E}_{s_1 \sim \mu, \tau \sim P_{\pi, \eta}} \left[ \sum_{t=2}^{\infty} \gamma^{t-1} (\hat{A}^\pi(s_t, \eta_t, a_t, \eta_{t+1})) \right] \\
& = \gamma \mathbb{E}_{(s_1, \eta_2) \sim \pi, \eta} \left[ \sum_{t=0}^{\infty} \gamma^t \hat{A}^\pi(s_{t+2}, \eta_{t+2}, a_{t+2}, \eta_{t+3}) \right] \\
& = \gamma \sum_{s_1, \eta_2} \mathbb{E}_{(s_2, \eta_2) \sim \pi, \eta} [\hat{A}^\pi(s_t, \eta_t, a_t, \eta_{t+1})] \\
& \leq \gamma \sum_{s_1, \eta_2} d_{\mu}^\pi(s_t, \eta_t) \max_{a_t, \eta_{t+1}} \hat{A}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) \\
& = \gamma \sum_{s_1, \eta_2} d_{\mu}^\pi(s_t, \eta_t) \max_{a_t, \eta_{t+1}} \hat{A}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) \\
& \leq (a) \gamma \left\| \frac{d_{\mu}^\pi}{d\mu} \right\| \sum_{s_1, \eta_2} d_{\mu}^\pi(s_t, \eta_t) \max_{a_t, \eta_{t+1}} \hat{A}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) \\
& \leq (b) \gamma \left\| \frac{d_{\mu}^\pi}{d\mu} \right\| \sum_{s_1, \eta_2} \max_{\pi_2 \in \Delta(\mathcal{A} \times \mathcal{X})^{S||H}} d_{\mu}^\pi(s_t, \eta_t) \max_{a_t, \eta_{t+1}} \hat{A}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) \\
& \leq (c) \gamma \left\| \frac{d_{\mu}^\pi}{d\mu} \right\| \sum_{s_1, \eta_2} \max_{\pi_2 \in \Delta(\mathcal{A} \times \mathcal{X})^{S||H}} d_{\mu}^\pi(s_t, \eta_t) \left[ \pi_2(a_t, \eta_{t+1} | s_t, \eta_t) - \pi_2(a_t, \eta_{t+1} | s_t, \eta_t) \right] \hat{A}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) \\
& \leq (d) \gamma \left\| \frac{d_{\mu}^\pi}{d\mu} \right\| \sum_{s_1, \eta_2} \max_{\pi_2 \in \Delta(\mathcal{A} \times \mathcal{X})^{S||H}} d_{\mu}^\pi(s_t, \eta_t) \left[ \pi_2(a_t, \eta_{t+1} | s_t, \eta_t) - \pi_2(a_t, \eta_{t+1} | s_t, \eta_t) \right] \hat{A}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) \\
& \leq (e) \gamma \left\| \frac{d_{\mu}^\pi}{d\mu} \right\| \sum_{s_1, \eta_2} \max_{\pi_2 \in \Delta(\mathcal{A} \times \mathcal{X})^{S||H}} d_{\mu}^\pi(s_t, \eta_t) \left[ \pi_2 - \pi \right]^\top \nabla_{\pi_{s_2}} J^\pi(\mu) \\& \end{align*}
\]

where (a) is true because \( \max_{a_t, \eta_{t+1}} \hat{A}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) = \tilde{J}^\pi(s_t, \eta_t) - \gamma \tilde{J}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) \geq 0 \); (b) follows since \( \max_{a_t, \eta_{t+1}} \hat{A}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) = \tilde{J}^\pi(s_t, \eta_t) - \gamma \tilde{J}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) \); (c) follows since \( \sum_{a_t, \eta_{t+1}} \pi_2(a_t, \eta_{t+1} | s_t, \eta_t) \hat{A}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) = \tilde{J}^\pi(s_t, \eta_t) - \sum_{a_t, \eta_{t+1}} \pi_2(a_t, \eta_{t+1} | s_t, \eta_t) \tilde{J}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) = 0 \); (d) follows since \( \sum_{a_t, \eta_{t+1}} \left[ \tilde{\pi}_2(a_t, \eta_{t+1} | s_t, \eta_t) - \pi_2(a_t, \eta_{t+1} | s_t, \eta_t) \right] \tilde{J}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) = 0 \); and (e) follows from Theorem 3.2 and Eq. (10).

Combining these two terms, we obtain

\[
\begin{align*}
J^r(\rho) - J^\pi(\rho) & \leq \left\| \frac{d_{\mu}^\pi}{d\mu} \right\| \sum_{s_1, \eta_2} \max_{\pi_1 \in \Delta(\mathcal{A} \times \mathcal{X})^{S||H}} (\pi_1 - \tilde{\pi}_1)^\top \nabla_{\pi_1} J^\pi(\mu) + \left\| \frac{d_{\mu}^\pi}{d\mu} \right\| \sum_{s_1, \eta_2} \max_{\pi_2 \in \Delta(\mathcal{A} \times \mathcal{X})^{S||H}} (\pi_2 - \tilde{\pi}_2)^\top \nabla_{\pi_2} J^\pi(\mu) \\
& \leq \max \left\{ \left\| \frac{d_{\mu}^\pi}{d\mu} \right\|, \left\| \frac{d_{\mu}^\pi}{d\mu} \right\| \right\} \sum_{s_1, \eta_2} \max_{\pi_1, \eta_2} (\pi_1 - \tilde{\pi}_1)^\top \nabla_{\pi_1} J^\pi(\mu) \geq 0
\end{align*}
\]

where the last inequality is because \( \max_{\pi_1 \in \Delta(\mathcal{A} \times \mathcal{X})^{S||H}} (\pi_1 - \tilde{\pi}_1)^\top \nabla_{\pi_1} J^\pi(\mu) \geq 0 \), \( \max_{\pi_2 \in \Delta(\mathcal{A} \times \mathcal{X})^{S||H}} (\pi_2 - \tilde{\pi}_2)^\top \nabla_{\pi_2} J^\pi(\mu) \geq 0 \). Furthermore, we have

\[
\begin{align*}
d_{\mu}^\pi(s, \eta) & = (1 - \gamma) \sum_{a_2, \eta_2} \mu^\pi(s_2, \eta_2) \sum_{t=0}^{\infty} \gamma^t \Pr^\pi(s_{t+2} = s, \eta_{t+2} = \eta | s_2, \eta_2)
\end{align*}
\]
\[
(1 - \gamma)\mu^s(s, \eta) \\
= (1 - \gamma) \sum_{s_1} \mu(s_1) \sum_{a_1} \pi_1(a_1, \eta) P(s|s_1, a_1) \\
\geq (1 - \gamma) \pi_1^{LB} \sum_{s_1, a_1} \mu(s_1) P(s|s_1, a_1) = (1 - \gamma) \pi_1^{LB} \mu^s(s, \eta)
\]

Then

\[
J^\pi(\rho) - J^{\pi^*}(\rho) \\
\leq \max \{ \| \frac{\rho}{\mu} \|_\infty, \frac{1}{(1 - \gamma) \pi_1^{LB}} \| \frac{d\pi^{\pi^*}}{d\mu} \|_\infty \} \max(\pi - \bar{\pi})^T \nabla_{\pi} J^\pi(\mu)
\]

This concludes the proof. \(\square\)

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

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

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 \(\pi \in \Delta(A \times H)^{\mathcal{S} \times \mathcal{A} \times \mathcal{H}}\) is \(\epsilon\)-stationary with respect to the initial state distribution \(\mu\) if

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

where \(\Delta(A \times H)^{\mathcal{S} \times \mathcal{A} \times \mathcal{H}}\) 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 \(\pi\) is first-order stationary.

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

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

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

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

where \(B_2\) is the unit \(\ell_2\) ball, and \(N_{\mathcal{C}}\) is the normal cone of the set \(\mathcal{C}\). Since \(-\nabla_{\pi} J^\pi^+(\mu)\) is \(\epsilon(\beta \sigma + 1)\) distance from the normal cone \(N_{\Delta(A \times H)^{\mathcal{S} \times \mathcal{A} \times \mathcal{H}}}(\pi^+)\) and \(\delta\) is in the tangent cone of \(\Delta(A \times H)^{\mathcal{S} \times \mathcal{A} \times \mathcal{H}}\) at \(\pi^+\), we have \(-\delta^T \nabla_{\pi} J^\pi^+(\mu) \leq \epsilon(\beta \sigma + 1)\). Thus

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

This completes the proof. \(\square\)

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

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

16
From Proposition B.3, we have
\[
\max_{t=0,1,...,T-1} \min_{\pi(t+1)} \delta^T \nabla_\pi J^{t+1}(\mu) \geq -(\beta \sigma + 1) \sqrt{\frac{2\sigma(J^{(0)}(\mu) - J^*(\mu))}{T}} \]

Observe that
\[
\max_{\pi \in \Delta(A \times \mathcal{H})} (\pi - \bar{\pi})^T \nabla_\pi J^*(\mu) = -2 \sqrt{|S| + |S||H|} \min_{\pi \in \Delta(A \times \mathcal{H})} \frac{1}{2} \frac{(\bar{\pi} - \pi)^T \nabla_\pi J^*(\mu)}{\sqrt{|S| + |S||H|}} \leq -2 \sqrt{|S| + |S||H|} \min_{\pi + \delta \in \Delta(A \times \mathcal{H})} \min_{\delta \in \Delta(A \times \mathcal{H})} \delta^T \nabla_\pi J^*(\mu)
\]

where the last step follows as \( ||\pi - \bar{\pi}||^2 \leq ||\pi||^2 + ||\bar{\pi}||^2 \) and follows the convexity of the probability simplex.

Then using Theorem 3.5 and \( \beta \sigma = 1 \), we have
\[
\max_{t=0,1,...,T-1} \min_{\pi(t+1)} J^{t+1}(\rho) - J^*(\rho) \leq \min_{t=0,1,...,T-1} D_1 \max_{\pi(t+1)} \delta^T \nabla_\pi J^{t+1}(\mu) \leq -2D_1 \sqrt{|S| + |S||H|} \max_{t=0,1,...,T-1} \min_{\pi(t+1)} \delta^T \nabla_\pi J^{t+1}(\mu) \leq 4D_1 \sqrt{2|S||H|} \frac{2\sigma(J^{(0)}(\mu) - J^*(\mu))}{\sqrt{T}} \leq 4D_1 \sqrt{2|S||H|} \frac{2\sigma(\frac{\lambda}{\alpha} + \lambda + \epsilon)}{\sqrt{T}}
\]

where (a) is true because \( J^{(0)}(\mu) - J^*(\mu) \leq J^{(0)}(\mu) \leq \frac{\lambda}{\alpha} + \lambda + \frac{1}{1-\gamma} ||\bar{\pi}||_\infty \) from Eq. (20) and \( |H| \geq 1 \). If we set \( T \) such that
\[
4D_1 \sqrt{2|S||H|} \frac{2\sigma(\frac{\lambda}{\alpha} + \lambda + \epsilon)}{\sqrt{T}} \leq \epsilon
\]
or, equivalently,
\[
T \geq \frac{64|S||H|}{\epsilon^2} \frac{\sigma(\frac{\lambda}{\alpha} + \lambda + \frac{1}{1-\gamma} ||\bar{\pi}||_\infty)}{\sqrt{T}}
\]

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

\[\square\]

C. Proofs in Section 3.2

**Proof of Lemma 3.8.** [Gradients for softmax parameterization] According to Theorem 3.2, we have
\[
\nabla_{\theta_1} J^*(\mu) = \mathbb{E}_{a_1 \sim \mu} \mathbb{E}_{(s_1, a_1, \eta_1)} \pi_1(\cdot|s_1) \nabla_{\theta_1} \log \pi_1(s_1, a_1, \eta_1) Q^\pi(s_1, a_1, \eta_1)
\]
\[
\nabla_{\theta_2} J^*(\mu) = \frac{\gamma}{1 - \gamma} \mathbb{E}_{(s_1, a_1, \eta_1)} \nabla_{\theta_2} \log \pi_2(\cdot|s_1, \eta_1) \mathbb{E}_{a_2 \sim d_{\pi_2}} Q_{a_2}(s_1, a_2, \eta_1, \eta_2)
\]

17
Because of the softmax parameterization, we have $\sum_{a_t, \eta_{t+1}} \pi_1(a_1, \eta_2|s_1) = \sum_{a_t, \eta_{t+1}} \pi_2(a_t, \eta_{t+1}|s_t, \eta_t) = 1$ satisfied automatically for all $s_1, s_t \in S$, $\eta_t \in H$, $t \geq 2$. As a result, $\sum_{a_t, \eta_t} \nabla_{\theta_1} \pi_1(a_1, \eta_2|s_1) = \sum_{a_t, \eta_{t+1}} \nabla_{\theta_2} \pi_2(a_t, \eta_{t+1}|s_t, \eta_t) = 0$ and we have
\[
\begin{align*}
\nabla_{\theta_1} J^\pi(\mu) &= \mathbb{E}_{s_1 \sim \mu, (a_1, \eta_2) \sim \pi_1(\cdot|s_1)} \left[ \nabla_{\theta_1} \log \pi_1(a_1, \eta_2|s_1)(-A^\pi(s_1, a_1, \eta_2)) \right] \\
\nabla_{\theta_2} J^\pi(\mu) &= \frac{\gamma}{1 - \gamma} \mathbb{E}_{(s_t, \eta_t) \sim d^\pi_{\mu}, (a_t, \eta_{t+1}) \sim \pi_2(\cdot|s_t, \eta_t)} \left[ \nabla_{\theta_2} \log \pi_2(a_t, \eta_{t+1}|s_t, \eta_t)(-\hat{A}^\pi_2(s_t, \eta_t, a_t, \eta_{t+1})) \right].
\end{align*}
\]
Because $\log \pi_1(a_1, \eta_2|s_1) = \theta_1(s_1, a_1, \eta_2) - \log \sum_{a'_1, \eta'_2} \exp(\theta_1(s_1, a'_1, \eta'_2))$ and $\log \pi_2(a_t, \eta_{t+1}|s_t, \eta_t) = \theta_2(s_t, \eta_t, a_t, \eta_{t+1}) - \log \sum_{a'_t, \eta'_{t+1}} \exp(\theta_2(s_t, \eta_t, a'_t, \eta'_{t+1}))$, we have
\[
\begin{align*}
\frac{\partial \log \pi_1(a_1, \eta_2|s_1)}{\partial \theta_1(s, a, \eta)} &= \mathbb{I}(s_1 = s, a_1 = a, \eta_2 = \eta) - \frac{\exp(\theta_1(s, a, \eta))}{\sum_{a'_1, \eta'_2} \exp(\theta_1(s, a'_1, \eta'_2))} \mathbb{I}(s_1 = s) \\
&= \mathbb{I}(s_1 = s) \left( \mathbb{I}(a_1 = a, \eta_2 = \eta) - \pi_1(a, \eta|s) \right) \\
\frac{\partial \log \pi_2(a_t, \eta_{t+1}|s_t, \eta_t)}{\partial \theta_2(s, \eta, a, \eta')} &= \mathbb{I}(s_t = s, \eta_t = \eta, a_t = a, \eta_{t+1} = \eta') - \frac{\exp(\theta_2(s, \eta, a, \eta'))}{\sum_{a'_t, \eta'_{t+1}} \exp(\theta_2(s, \eta, a'_t, \eta'_{t+1}))} \mathbb{I}(s_t = s, \eta_t = \eta) \\
&= \mathbb{I}(s_t = s, \eta_t = \eta) \left( \mathbb{I}(a_t = a, \eta_{t+1} = \eta') - \pi_2(a, \eta'|s, \eta) \right)
\end{align*}
\]
and thus,\[
\begin{align*}
\frac{\partial J^\pi(\mu)}{\partial \theta_1(s, a, \eta)} &= \sum_{s_1} \mu(s_1) \sum_{a_1, \eta_{t+1}} \pi_1(a_1, \eta_2|s_1) \mathbb{I}(s_1 = s) \left( \mathbb{I}(a_1 = a, \eta_2 = \eta) - \pi_1(a, \eta|s) \right)(-A^\pi(s_1, a_1, \eta_2)) \\
&= \sum_{s_1} \mu(s_1) \sum_{a_1, \eta_{t+1}} \pi_1(a_1, \eta_2|s_1) \mathbb{I}(s_1 = s, a = a, \eta = \eta) \left( -\pi_1(a, \eta|s) \right)(-A^\pi(s_1, a_1, \eta_2)) \\
&\quad - \pi_1(a, \eta|s) \sum_{s_1} \mu(s_1) \sum_{a_1, \eta_{t+1}} \pi_1(a_1, \eta_2|s_1) \mathbb{I}(s_1 = s) \left( -\pi_1(a_1, \eta_2|s_1) \right)(-A^\pi(s_1, a_1, \eta_2)) \\
&\quad - (a) \mu(s) \pi_1(a, \eta|s)(-A^\pi(s, a, \eta))
\end{align*}
\]
and \[
\begin{align*}
\frac{\partial J^\pi(\mu)}{\partial \theta_2(s, \eta, a, \eta')} &= \frac{\gamma}{1 - \gamma} \sum_{s_t, \eta_t} d^\pi_{\mu}(s_t, \eta_t) \sum_{a_t, \eta_{t+1}} \pi_2(a_t, \eta_{t+1}|s_t, \eta_t) \mathbb{I}(s_t = s, \eta_t = \eta) \left( \mathbb{I}(a_t = a, \eta_{t+1} = \eta') - \pi_2(a, \eta'|s, \eta) \right)(-\hat{A}^\pi_2(s_t, \eta_t, a_t, \eta_{t+1})) \\
&= \frac{\gamma}{1 - \gamma} \sum_{s_t, \eta_t} d^\pi_{\mu}(s_t, \eta_t) \sum_{a_t, \eta_{t+1}} \pi_2(a_t, \eta_{t+1}|s_t, \eta_t) \mathbb{I}(s_t = s, \eta_t = \eta) \left( -\pi_2(a, \eta'|s, \eta) \right)(-\hat{A}^\pi_2(s_t, \eta_t, a_t, \eta_{t+1})) \\
&\quad - \pi_2(a, \eta'|s, \eta) \sum_{s_t, \eta_t} \frac{\gamma}{1 - \gamma} d^\pi_{\mu}(s_t, \eta_t) \sum_{a_t, \eta_{t+1}} \pi_2(a_t, \eta_{t+1}|s_t, \eta_t) \mathbb{I}(s_t, \eta_t = \eta) \left( -\hat{A}^\pi_2(s_t, \eta_t, a_t, \eta_{t+1}) \right) \\
&= \left( a \right) \frac{\gamma}{1 - \gamma} \sum_{s_t, \eta_t} d^\pi_{\mu}(s_t, \eta_t) \pi_2(a, \eta'|s, \eta)(-\hat{A}^\pi_2(s, \eta, a, \eta'))
\end{align*}
\]
where (a) is true because $\sum_{a_1, \eta_{t+1}} \pi_1(a_1, \eta_2|s)A^\pi(s_1, a_1, \eta_2) = 0$ 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 $J^\pi(\mu)$, let $\theta_{s, \eta} \in \mathbb{R}^{|A||H|}$ denote the parameters associated with a given state $s$, $\eta$. We have
\[
\nabla_{\theta_{s, \eta}} J^\pi(\mu) = \pi_1^\theta(a, \eta|s)(e_{a, \eta} - \pi_1^\theta(\cdot|s)) \\
\\nabla_{\theta_{s, \eta}} J^\pi(\mu) = \pi_2^\theta(a_t, \eta_{t+1}|s_t, \eta_t)(e_{a_t, \eta_{t+1}} - \pi_2^\theta(\cdot|s_t, \eta_t))
\]
where $e_{a, \eta}$ is a basis vector that only equals 1 at $(a, \eta)$-location and 0 elsewhere. Differentiating them with respect to $\theta$ again, we get

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

The same arguments apply to $\pi_2^\alpha$.

Define $\pi_1^\alpha := \pi_1^{\alpha + u}$ where $u \in \mathbb{R}^{|\mathcal{A}||\mathcal{H}|}$ 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 $\pi_1^\alpha$ with respect to $\alpha$, we obtain

$$
\sum_{a_1, \eta_2} \left| \frac{d \pi_1^\alpha(a_1, \eta_2|s_1)}{d \alpha} \right|_{\alpha=0} \leq \sum_{a_1, \eta_2} \left| u^T \nabla_{\theta + \alpha u} \pi_1^\alpha(a_1, \eta_2|s_1) \right|_{\alpha=0} u
$$

Now differentiating once again with respect to $\alpha$, we get

$$
\sum_{a_1, \eta_2} \left| \frac{d^2 \pi_1^\alpha(a_1, \eta_2|s_1)}{(d \alpha)^2} \right|_{\alpha=0} \leq \sum_{a_1, \eta_2} \left| u^T \nabla_{\theta + \alpha u}^2 \pi_1^\alpha(a_1, \eta_2|s_1) \right|_{\alpha=0} u
$$

where we get $\max_{a_1, \eta_2} \left| u^T e_{a_1, \eta_2} e_{a_1, \eta_2}^T u_s \right| + \left| u^T e_{a_1, \eta_2} \pi_1^\alpha(\cdot, |s_1)^T u_s \right| + \left| u^T \pi_1^\alpha(\cdot, |s_1) e_{a_1, \eta_2} u_s \right|

\leq \left| u^T \pi_1^\alpha(\cdot, |s) \pi_1^\alpha(\cdot, |s)^T u_s \right| + \left| u^T \text{diag}(\pi_1^\alpha(\cdot, |s_1) u_s \right) \leq 6

The same arguments apply to $\pi_2^\alpha$, where we get $\sum_{a_1, \eta_1+1} \left| \frac{d \pi_1^\alpha(a_1, \eta_1+1|s_1)}{d \alpha} \right|_{\alpha=0} \leq 2$ and

$$
\sum_{a_1, \eta_1+1} \left| \frac{d^2 \pi_1^\alpha(a_1, \eta_1+1|s_1)}{(d \alpha)^2} \right|_{\alpha=0} \leq 6
$$

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

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

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

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

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

We have

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

Equivalently,

$$
\nabla_{\theta_{s_1}} R_1(\theta) = \frac{1}{|\mathcal{A}||\mathcal{H}|} - \pi_1^\alpha(\cdot, |s_1)$$
On the Global Convergence of Risk-AVERSE Policy Gradient

As a result,
\[ \nabla_{\theta_s}^2 R_1(\theta) = -\text{diag}(\pi^\theta_1(\cdot, \cdot | s_1)) + \pi^\theta_1(\cdot, \cdot | s_1)\pi^\theta_1(\cdot, \cdot | s_1)^T \]

and for any vector \( u_{s_1} \),
\[ |u^T_s \nabla_{\theta_s}^2 R_1(\theta) u_s| = |u^T_s \text{diag}(\pi^\theta_1(\cdot, \cdot | s_1))u_s - (u_{s_1} \pi^\theta_1(\cdot, \cdot | s_1))^2| \leq 2\|u_{s_1}\|^2 \]

Because \( \nabla_{\theta_s} \nabla_{\theta_s} R_1(\theta) = 0 \) for \( s \neq s' \), we have
\[ |u^T \nabla_{\theta_s} R_1(\theta)| = |\sum_s u^T_s \nabla_{\theta_s} R_1(\theta)| \leq 2 \sum_s \|u_s\|^2 \leq 2 \|u\|^2 \]

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 \( R_2(\theta) = \frac{1}{|\mathcal{A}|} \sum_{s, \eta_t, a_t, \eta_{t+1}} \log \pi^\theta_2(a_t, \eta_{t+1} | s_t, \eta_t) \), where we get \( -\frac{\kappa}{|\mathcal{S}||\mathcal{A}|} R_2(\theta) \) is \( \frac{2\kappa}{|\mathcal{S}||\mathcal{A}|} \)-smooth.

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

**Proof of Theorem 3.10.** [Suboptimality for softmax parameterization] We only need to show that \( \max_{a_1, \eta_2} A^\pi(s_1, a_1, \eta_2) \leq \frac{2\kappa}{\mu(s_1)|\mathcal{S}|}, \forall s_1 \in \mathcal{S} \) and \( \max_{a_1, \eta_{t+1}} A^\pi(s_t, a_t, \eta_t, \eta_{t+1}) \leq \frac{2\kappa}{\mu(s_t)|\mathcal{S}||\mathcal{A}|}, \forall s_t \in \mathcal{S}, \eta_t \in \mathcal{H} \). To see why this is sufficient, observe that by the performance difference lemma (Lemma 3.1),
\[ J^\pi(\rho) - J^{\pi^*}(\rho) = \mathbb{E}_{s_1,...,s_T \sim \pi_\rho(\cdot | s_1)} \left[ A^\pi(s_1, a_1, \eta_2) + \sum_{t=2}^{\infty} \gamma^{t-1} A^{\pi^*}(s_t, \eta_t, a_t, \eta_{t+1}) \right] \]
\[ = \sum_{s_1, a_1, \eta_2} \rho(s_1) \pi^\theta_1(a_1, \eta_2 | s_1) A^\pi(s_1, a_1, \eta_2) + \frac{\gamma}{1-\gamma} \sum_{s_t, \eta_t} d^\pi_{\rho^*} \pi^\theta_2(s_t, \eta_t) \sum_{a_t, \eta_{t+1}} \pi^\theta_2(a_t, \eta_{t+1} | s_t, \eta_t) A^{\pi^*}(s_t, \eta_t, a_t, \eta_{t+1}) \]
\[ \leq \sum_{s_1} \rho(s_1) \max_{a_1, \eta_2} A^\pi(s_1, a_1, \eta_2) + \frac{\gamma}{1-\gamma} \sum_{s_t, \eta_t} d^\pi_{\rho^*} \pi^\theta_2(s_t, \eta_t) \max_{a_t, \eta_{t+1}} A^{\pi^*}(s_t, \eta_t, a_t, \eta_{t+1}) \]
\[ \leq \sum_{s_1} \rho(s_1) - \frac{2\kappa}{\mu(s_1)|\mathcal{S}|} + \frac{\gamma}{1-\gamma} \sum_{s_t, \eta_t} d^\pi_{\rho^*} \pi^\theta_2(s_t, \eta_t) \frac{2\kappa}{\gamma \pi^B \mu |\mathcal{H}||\mathcal{A}|} \]
\[ \leq 2\kappa \frac{\|\rho\|_\infty}{\mu} + \frac{2\kappa}{(1-\gamma)^{\pi^B}} \|\mu^{-\pi^B}\|_\infty. \]

Now to prove \( \max_{a_1, \eta_2} A^\pi(s_1, a_1, \eta_2) \leq \frac{2\kappa}{\mu(s_1)|\mathcal{S}|} \), it suffices to bound \( A^\pi(s_1, a_1, \eta_2) \) for any state-action pair \( s_1, a_1, \eta_2 \) where \( A^\pi(s_1, a_1, \eta_2) \geq 0 \) otherwise the claim holds trivially. Consider an \( (s_1, a_1, \eta_2) \) pair such that \( A^\pi(s_1, a_1, \eta_2) \geq 0 \). Using Lemma 3.8,
\[ \frac{\partial L_\kappa(\theta)}{\partial \theta_1(s_1, a_1, \eta_2)} = \mu(s_1) \pi^\theta_1(a_1, \eta_2 | s_1)(-A^{\pi^*}(s_1, a_1, \eta_2)) - \frac{\kappa}{|\mathcal{S}|} \left( \frac{1}{|\mathcal{A}|} - \pi^\theta_2(a_1, \eta_2 | s_1) \right) \]

The gradient norm assumption \( \|\nabla_{\theta_1} L_\kappa(\theta)\|_2 \leq \epsilon_1 \leq \frac{\kappa}{2|\mathcal{S}||\mathcal{A}||\mathcal{H}|} \) implies that
\[ -\frac{\kappa}{2|\mathcal{S}||\mathcal{A}||\mathcal{H}|} \leq -\epsilon_1 \leq \frac{\partial L_\kappa(\theta)}{\partial \theta_1(s_1, a_1, \eta_2)} = \mu(s_1) \pi^\theta_1(a_1, \eta_2 | s_1)(-A^{\pi^*}(s_1, a_1, \eta_2)) - \frac{\kappa}{|\mathcal{S}|} \left( \frac{1}{|\mathcal{A}|} - \pi^\theta_2(a_1, \eta_2 | s_1) \right) \]

where (a) is due to \( \frac{\partial L_\kappa(\theta)}{\partial \theta_1(s_1, a_1, \eta_2)} \), \( \|\nabla_{\theta_1} L_\kappa(\theta)\|_2 \leq \epsilon_1 \) and (b) uses the fact that \( -A^\pi(s_1, a_1, \eta_2) \leq 0 \). Rearranging the terms, we get
\[ \pi^\theta_1(a_1, \eta_2 | s_1) \geq \frac{1}{|\mathcal{A}|} \frac{1}{|\mathcal{H}|} - \frac{1}{2|\mathcal{A}||\mathcal{H}|} = \frac{1}{2|\mathcal{A}||\mathcal{H}|}. \]

20
From (14), we have
\[ A^{\pi^0}(s_1, a_1, \eta_2) = \frac{1}{\mu(s_1)} \left( \frac{1}{\pi^0_1(a_1, \eta_2 | s_1)} - \frac{-\partial L_{\kappa}(\theta)}{\mu(s, a, \eta)} \right) \leq \frac{1}{\mu(s_1)} \left( \frac{1}{\pi^0_1(a_1, \eta_2 | s_1)} - \frac{-\partial L_{\kappa}(\theta)}{\mu(s, a, \eta)} \right) + \frac{\kappa}{|S|} \left( 1 - \frac{1}{\pi^0_1(a_1, \eta_2 | s_1)} \right) \]

To prove \( \max_{a_t, \eta_{t+1}} \hat{A}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) \leq \frac{2\kappa}{\gamma |\pi^0_1(s_t, \eta_t) \mu(s_t, \eta_t)|} \), \( \forall s_t, \eta_t \), 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 \( \hat{A}^\pi(s_t, \eta_t, a_t, \eta_{t+1}) \geq 0 \). Using Lemma 3.8, we have

\[
\frac{\partial L_{\kappa}(\theta)}{\partial \theta_2(s_t, \eta_t, a_t, \eta_{t+1})} \leq -\frac{\kappa}{|S||A|} + \frac{1}{|A|} - \frac{1}{2|A|} \leq -\frac{\kappa}{2|S||H||A|} \leq -\frac{\kappa}{2|S||H||A|} \leq -\frac{\kappa}{2|S||H||A|}
\]

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

\[
\pi_2(a_t, \eta_{t+1} | s_t, \eta_t) \geq \frac{1}{|A|} - \frac{1}{2|A|} = \frac{1}{2|A|}.
\]

From (16), we have

\[
\hat{A}^{\pi^2}(s_t, \eta_t, a_t, \eta_{t+1}) = \frac{1}{\gamma} \frac{1}{d^p_{\mu}(s_t, \eta_t)} \left( \frac{1}{\pi_2(a_t, \eta_{t+1} | s_t, \eta_t)} - \frac{-\partial L_{\kappa}(\theta)}{\mu(s_t, \eta_t, a_t, \eta_{t+1})} \right) \leq \frac{1}{\gamma} \frac{1}{d^p_{\mu}(s_t, \eta_t)} \left( \frac{1}{\pi_2(a_t, \eta_{t+1} | s_t, \eta_t)} - \frac{-\partial L_{\kappa}(\theta)}{\mu(s_t, \eta_t, a_t, \eta_{t+1})} \right) + \frac{\kappa}{|S||H|} \left( 1 - \frac{1}{\pi_2(a_t, \eta_{t+1} | s_t, \eta_t)} \right)
\]

where the last inequality uses the fact that \( d^p_{\mu}(s_t, \eta_t) \geq (1 - \gamma) \mu^\pi(s_t, \eta_t) \geq (1 - \gamma) \pi_2^{B}(s_t, \eta_t) \) because \( \pi_1(a, \eta_t | s) \geq \pi_2^{B}(s_t, \eta_t) \). This completes the proof. \( \square \)

**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 \( \pi_1^{(0)} \) and \( \pi_2^{(0)} \) will be bounded away from 0, i.e., \( \pi_1^{(0)}, \pi_2^{(0)} > 0 \). As a result, the initial regularized objective function can be upper bounded, i.e., \( L_{\kappa}(\theta^{(0)}) < +\infty \). Indeed, \( L_{\kappa}(\theta^{(0)}) \leq \frac{C}{|S||A|} - \frac{\kappa}{|S||H||A|} \sum_{s_1,a_1,\eta_2} \log \pi_1^{(0)}(a_1, \eta_2 | s_1) - \frac{\kappa}{|S||H||A|} \sum_{s_1,a_1,\eta_2,\eta_3} \log \pi_2^{(0)}(a_2, \eta_3 | s_1, \eta_2) - 2\kappa \log |A||H| < +\infty. \) Because \( L_{\kappa}(\theta) \) is \( \sigma_{\kappa} \)-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 \( t \in \mathbb{Z}_+ \cup \{ +\infty \} \) such that \( \pi_1^{(t)}(a, \eta | s) = 0 \). Then we have \( -\log \pi_1^{(t)}(a, \eta | s) = +\infty \) and \( L_{\kappa}(\theta^{(t)}) = +\infty \), which contradicts with \( L_{\kappa}(\theta^{(t)}) < +\infty \). Similar arguments can also be applied to \( \pi_2^{(t)}(a, \eta | s, \eta) \) where we conclude that \( \pi_2^{B} := \inf_{t \geq 1} \min_{a, \eta, \eta'} \pi_2^{(0)}(a, \eta | s, \eta) > 0 \). This completes the proof. \( \square \)
On the Global Convergence of Risk-Averse Policy Gradient

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

\[
\kappa = \frac{\epsilon}{2 \| \mu \|_{\infty} + \frac{2}{(1 - \gamma) \pi_{1}^{B} \| \mu \|_{\infty}} \| d^{\pi^{*}}_{\mu} \|_{\infty}}
\]

and if \( \| \nabla_{\theta} L_{\kappa}(\theta) \|_{2} \leq \frac{\kappa}{2 \| S \| | H | | A | | H |} \) (then we have \( \| \nabla_{\theta_{1}} L_{\kappa}(\theta) \|_{2} \leq \frac{\kappa}{2 \| S \| | H | | A | | H |} \leq \frac{\kappa}{2 \| S \| | A | | H |} \) and \( \| \nabla_{\theta_{2}} L_{\kappa}(\theta) \|_{2} \leq \frac{\kappa}{2 \| S \| | H | | A | | H |} \)). 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

\[
\min_{t \leq T} \| \nabla_{\theta} L_{\kappa}(\theta(t)) \|_{2} \leq \frac{2\sigma_{\kappa}(L_{\kappa}(\theta(0)) - L_{\kappa}(\theta^{*}))}{T} \leq \frac{2\sigma_{\kappa}(\bar{C} - \kappa \log \pi_{1}^{L_{B}} - \kappa \log \pi_{2}^{L_{B}})}{T}
\]

where (a) is true because \( L_{\kappa}(\theta(0)) - L_{\kappa}(\theta^{*}) = J^{(0)}(\mu) - J^{*}(\mu) - \kappa \sum_{s, a, r, \pi} \log \frac{\pi_{1}^{0}(a, \pi, r, s)}{\pi_{1}(a, \pi, r, s)} \leq \bar{C} - \kappa \log \pi_{1}^{L_{B}} - \kappa \log \pi_{2}^{L_{B}} \); (b) uses the fact that \( \kappa \leq 1 \). Denoting \( B = \bar{C} - \log \pi_{1}^{L_{B}} - \log \pi_{2}^{L_{B}} \), we seek to ensure

\[
\sqrt{\frac{2\sigma_{\kappa} B}{T}} \leq \frac{\kappa}{2 \| S \| | H | | A | | H |}
\]

Choosing \( T \geq \frac{8 \sigma_{\kappa} B \| S \| | H | | A |^{2}}{\kappa^{2}} \) satisfies the above inequality. By Lemma 3.9, we can plug in \( \sigma_{\kappa} = 6(\frac{1}{\sigma} + \lambda) + \frac{8}{(1 - \gamma)^{3}} \| \bar{c} \|_{\infty} + \frac{2\rho}{5 | S | | H |} \), which gives us

\[
\frac{8 \sigma_{\kappa} B \| S \| | H | | A |^{2}}{\kappa^{2}} = \frac{48(\frac{1}{\sigma} + \lambda) B \| S \| | H | | A |^{2}}{\kappa^{2}} + \frac{64 \| \bar{c} \|_{\infty} B \| S \| | H | | A |^{2}}{\kappa^{2}} + \frac{16 \kappa B \| S \| | H | | A |^{2}}{\kappa^{2}} + \frac{16 \kappa B \| S \| | H |^{3} | A |^{2}}{\kappa^{2}}
\]

\[
\leq a \leq \frac{48(\frac{1}{\sigma} + \lambda) B \| S \| | H |^{4} | A |^{2}}{(1 - \gamma)^{3} \kappa^{2}} + \frac{64 \| \bar{c} \|_{\infty} B \| S \| | H |^{4} | A |^{2}}{(1 - \gamma)^{3} \kappa^{2}} + \frac{32 \| \bar{c} \|_{\infty} B \| S \| | H |^{4} | A |^{2}}{(1 - \gamma)^{3} \kappa^{2}} + \frac{32 \| \bar{c} \|_{\infty} B \| S \| | H |^{4} | A |^{2}}{(1 - \gamma)^{3} \kappa^{2}}
\]

\[
\leq b \leq \frac{48(\frac{1}{\sigma} + \lambda) B \| S \| | H |^{4} | A |^{2}}{(1 - \gamma)^{3} \kappa^{2}} + \frac{64 \| \bar{c} \|_{\infty} B \| S \| | H |^{4} | A |^{2}}{(1 - \gamma)^{3} \kappa^{2}} + \frac{32 \| \bar{c} \|_{\infty} B \| S \| | H |^{4} | A |^{2}}{(1 - \gamma)^{3} \kappa^{2}} + \frac{32 \| \bar{c} \|_{\infty} B \| S \| | H |^{4} | A |^{2}}{(1 - \gamma)^{3} \kappa^{2}}
\]

where (a) is true because \( 0 < \gamma < 1 \), \( \kappa \leq 1 \), \( | S |, | H | \geq 1 \) and (b) uses Eq. (17).

Therefore, choosing \( T \geq \frac{64(3(\frac{1}{\sigma} + \lambda) + 4 \| \bar{c} \|_{\infty} + 2) B \| S \| | H |^{4} | A |^{2}}{(1 - \gamma)^{3} \kappa^{2}} \) 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 \( \pi^{a} := \pi^{\theta + \alpha u}, J(\alpha) := J^{\pi^{a}}(s_{1}) \). Assume that

\[
\sum_{a_{1}, a_{2}} \frac{d^{2} \pi^{a}(a_{1}, a_{2})}{d\alpha^{2}} \bigg|_{\alpha=0} \leq C_{1}, \sum_{a_{1}, a_{2}} \frac{d^{2} \pi^{a}(a_{1}, a_{2}|s_{1})}{d\alpha^{2}} \bigg|_{\alpha=0} \leq C_{2}, \forall s_{1} \in S
\]

\[
\sum_{a_{1}, a_{2}} \frac{d^{2} \pi^{a}(a_{1}, a_{2}|s_{1})}{d\alpha^{2}} \bigg|_{\alpha=0} \leq C_{1}, \sum_{a_{1}, a_{2}} \frac{d^{2} \pi^{a}(a_{1}, a_{2}|s_{1})}{d\alpha^{2}} \bigg|_{\alpha=0} \leq C_{2}, \forall s_{1} \in S, \eta \in H
\]
Then
\[ \max_{||u||_2=1} \left| \frac{d^2 \tilde{J}(\alpha)}{(da)^2} \right|_{\alpha=0} \leq C_2 \left( \frac{\lambda}{\alpha} + \gamma \bar{c} + \frac{||\bar{c}||_\infty}{(1-\gamma)^2} + \frac{2\gamma C_1^2}{(1-\gamma)^3} ||\bar{c}||_\infty \right). \]

**Proof of Lemma D.1.** Let \( \tilde{P}(\alpha) \) be the state-action transition matrix under policy \( \pi^\alpha \), i.e.,
\[
[\tilde{P}(\alpha)]_{(s_t, \eta_t, a_t, \eta_{t+1}) \rightarrow (s'_t, \eta'_t, a'_t, \eta'_{t+1})} = \begin{cases} 
P(s'_t | s_t, a_t) \pi^\alpha_2(a'_t, \eta'_{t+1} | s'_t, \eta'_t), & \text{if } \eta'_t = \eta_{t+1} \\
0, & \text{otherwise.} \end{cases}
\]
We can differentiate \( \tilde{P}(\alpha) \) with respect to \( \alpha \):
\[
\left[ \frac{d\tilde{P}(\alpha)}{d\alpha} \right]_{\alpha=0} = \sum_{s'_t, a'_t, \eta'_{t+1}} \frac{d\pi^\alpha_2(a'_t, \eta'_{t+1} | s'_t, \eta'_t)}{d\alpha} |_{\alpha=0} P(s'_t | s_t, a_t) x_{s'_t, \eta_{t+1}, a'_t, \eta'_{t+1}}
\]
and thus
\[
\max_{||u||_2=1} \left| \frac{d\tilde{P}(\alpha)}{d\alpha} \right|_{\alpha=0} x = \max_{||u||_2=1} \left| \sum_{s'_t, a'_t, \eta'_{t+1}} \frac{d\pi^\alpha_2(a'_t, \eta'_{t+1} | s'_t, \eta'_t)}{d\alpha} |_{\alpha=0} P(s'_t | s_t, a_t) x_{s'_t, \eta_{t+1}, a'_t, \eta'_{t+1}} \right|
\]
\[
\leq \max_{||u||_2=1} \sum_{s'_t, a'_t, \eta'_{t+1}} \left| \frac{d\pi^\alpha_2(a'_t, \eta'_{t+1} | s'_t, \eta'_t)}{d\alpha} \right|_{\alpha=0} P(s'_t | s_t, a_t) ||x||_\infty \sum_{s'_t, a'_t, \eta'_{t+1}} \left| \frac{d\pi^\alpha_2(a'_t, \eta'_{t+1} | s'_t, \eta'_t)}{d\alpha} \right|_{\alpha=0} \leq \sum_{s'_t} P(s'_t | s_t, a_t) ||x||_\infty C_1 \leq C_1 ||x||_\infty
\]

By the definition of \( \ell_\infty \) norm, we have
\[
\max_{||u||_2=1} \left| \frac{d\tilde{P}(\alpha)}{d\alpha} \right|_{\alpha=0} x \leq C_1 ||x||_\infty
\]

Similarly, differentiating \( \tilde{P}(\alpha) \) twice with respect to \( \alpha \), we obtain
\[
\left[ \frac{d^2 \tilde{P}(\alpha)}{(da)^2} \right]_{\alpha=0} = \frac{d^2 \pi^\alpha_2(a'_t, \eta'_{t+1} | s'_t, \eta'_t)}{(da)^2} |_{\alpha=0} P(s'_t | s_t, a_t), \text{ if } \eta'_t = \eta_{t+1} \quad 0, \text{ otherwise.}
\]
An identical argument leads to the following result: for arbitrary \( x \),
\[
\max_{||u||_2=1} \left| \frac{d^2 \tilde{P}(\alpha)}{(da)^2} \right|_{\alpha=0} x \leq C_2 ||x||_\infty
\]

Let \( Q^\alpha(s_1, a_1, \eta_2) \) be the corresponding \( Q \)-function for policy \( \pi^\alpha \) at state \( s_1 \) and action \( a_1, \eta_2 \) and denote \( \bar{c} \) as a vector where \( \bar{c}(s_t, \eta_t, a_t, \eta_{t+1}) = \frac{\lambda}{\alpha} [C(s_t, a_t) - \eta_{t+1} + (1-\lambda)C(s_t, a_t) + \gamma \lambda \eta_{t+1}]. \) Observe that \( Q^\alpha(s_1, a_1, \eta_2) \) can be written as
\[
Q^\alpha(s_1, a_1, \eta_2) = (a) C(s_1, a_1) + \gamma \lambda \eta_2 + e^T(s_1, \eta_1, a_1, \eta_2) \sum_{n=1}^{\infty} \gamma^n \tilde{P}(\alpha)^n \bar{c}
\]
According to Eq. (19), we have
\[
\alpha \text{ expansion of matrix inverse. Now differentiating twice with respect to }
\]
\[
\frac{dQ^n(s_1, a_1, \eta_2)}{d\alpha} = \gamma e^{T(s_1, \eta_1, a_1, \eta_2)} M(\alpha) \frac{d^2 \hat{P}(\alpha)}{d\alpha} M(\alpha) \tilde{c},
\]
\[
\frac{d^2Q^n(s_1, a_1, \eta_2)}{(d\alpha)^2} = 2\gamma^2 e^{T(s_1, \eta_1, a_1, \eta_2)} M(\alpha) \frac{d^2 \hat{P}(\alpha)}{d\alpha} M(\alpha) \tilde{c} + \gamma e^{T(s_1, \eta_1, a_1, \eta_2)} M(\alpha) \frac{d^2 \hat{P}(\alpha)}{(d\alpha)^2} M(\alpha) \tilde{c}.
\]
Because \( M(\alpha) \geq 0 \) (componentwise) and \( M(\alpha) 1 = \frac{1}{1-\gamma} 1 \), i.e., each row of \( M(\alpha) \) is positive and sums to \( \frac{1}{1-\gamma} \), we have
\[
\max_{||u||_2=1} ||M(\alpha)u||_\infty \leq \frac{1}{1-\gamma} ||x||_\infty
\]
According to Eq. (19), we have
\[
|Q^n(s_1, a_1, \eta_2)| = \left| -\frac{\lambda}{\alpha} [C(s_1, a_1) - \eta_1] + \lambda C(s_1, a_1) + e^{T(s_1, \eta_1, a_1, \eta_2)} M(\alpha) \tilde{c} \right|
\]
\[
\leq \lambda \alpha + \lambda + \frac{1}{1-\gamma} ||\tilde{c}||_\infty
\]
where (a) is true because \( C(s_1, a_1) \in [0, 1], \ \eta_1 \in [0, 1] \). Furthermore, we obtain
\[
\max_{||u||_2=1} \left| \frac{dQ^n(s_1, a_1, \eta_2)}{d\alpha} \right|_{\alpha=0} \leq ||M(\alpha)\frac{d^2 \hat{P}(\alpha)}{d\alpha} M(\alpha) \tilde{c}| |_{\alpha=0} \leq \gamma C_1 \frac{(1-\gamma)^2}{(1-\gamma)^2} ||\tilde{c}||_\infty
\]
\[
\max_{||u||_2=1} \left| \frac{d^2Q^n(s_1, a_1, \eta_2)}{(d\alpha)^2} \right|_{\alpha=0} \leq 2\gamma^2 \max_{||u||_2=1} ||M(\alpha)\frac{d^2 \hat{P}(\alpha)}{d\alpha} M(\alpha) \tilde{c}||_\infty + \gamma \max_{||u||_2=1} ||M(\alpha)\frac{d^2 \hat{P}(\alpha)}{(d\alpha)^2} M(\alpha) \tilde{c}||_\infty \leq \left( \frac{2\gamma^2 C_1^2}{(1-\gamma)^3} + \frac{\gamma C_2}{(1-\gamma)^2} \right) ||\tilde{c}||_\infty
\]
Consider the equation
\[
\tilde{J}(\alpha) = \sum_{a_1, \eta_2} \pi^0_{a_1}(a_1, \eta_2|s_1) Q^n(s_1, a_1, \eta_2)
\]
By differentiating \( \tilde{J}(\alpha) \) twice with respect to \( \alpha \), we get
\[
\frac{d^2 \tilde{J}(\alpha)}{(d\alpha)^2} = \sum_{a_1, \eta_2} d^2 \pi^0_{a_1}(a_1, \eta_2|s_1) Q^n(s_1, a_1, \eta_2) + 2 \sum_{a_1, \eta_2} \frac{d \pi^0_{a_1}(a_1, \eta_2|s_1)}{d\alpha} \frac{dQ^n(s_1, a_1, \eta_2)}{d\alpha} + \sum_{a_1, \eta_2} \pi^0_{a_1}(a_1, \eta_2|s_1) \frac{d^2 Q^n(s_1, a_1, \eta_2)}{(d\alpha)^2}
\]
Thus,
\[
\max_{||u||_2=1} \left| \frac{d^2 \tilde{J}(\alpha)}{(d\alpha)^2} \right|_{\alpha=0} \leq C_2 \left( \frac{\lambda}{\alpha} + \lambda + \frac{1}{1-\gamma} ||\tilde{c}||_\infty \right) + 2\gamma C_1^2 \frac{(1-\gamma)^2}{(1-\gamma)^2} ||\tilde{c}||_\infty + \frac{2\gamma^2 C_1^2}{(1-\gamma)^3} ||\tilde{c}||_\infty + \frac{\gamma C_2}{(1-\gamma)^2} ||\tilde{c}||_\infty \leq C_2 \left( \frac{\lambda}{\alpha} + \lambda + \frac{1}{1-\gamma} ||\tilde{c}||_\infty \right) + 2\gamma C_1^2 \frac{(1-\gamma)^2}{(1-\gamma)^2} ||\tilde{c}||_\infty.
\]
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, \( \text{dom}(f) \) is convex, and \( f \) is \( \sigma \)-smooth over \( \text{int}(\text{dom}(f)) \).

**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. \( \min_{t=0,1,\ldots,T-1} \|G^\beta(x(t))\|_2 \leq \sqrt{\frac{2\sigma(f(x^0) - f(x^*))}{T}} \)

**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 \( \ell_2 \) 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 \( \lambda \)-value in Figures 7–11.

![Figure 7](image)

*Figure 7.* For \( \lambda = 0 \), the learned optimal path is \([3, 0] \to [2, 0] \to [2, 1] \to [2, 2] \to [2, 3] \to [3, 3] \).
Figure 8. For $\lambda = 0.25$, the learned optimal path is $[3, 0] \rightarrow [2, 0] \rightarrow [2, 1] \rightarrow [2, 2] \rightarrow [2, 3] \rightarrow [3, 3]$.

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

On the Global Convergence of Risk-Averse Policy Gradient
Figure 10. For $\lambda = 0.75$, the learned optimal path is $[3, 0] \rightarrow [2, 0] \rightarrow [1, 0] \rightarrow [1, 1] \rightarrow [1, 2] \rightarrow [1, 3] \rightarrow [2, 3] \rightarrow [3, 3]$.

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