
Robust Entropy-regularized Markov Decision Processes

Tien Mai

School of Computing and Information Systems
Singapore Management University
atmai@smu.edu.sg

Patrick Jaillet

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
jaillet@mit.edu

Abstract

Stochastic and soft optimal policies resulting from entropy-regularized Markov decision processes (ER-MDP) are desirable for exploration and imitation learning applications. Motivated by the fact that such policies are sensitive with respect to the state transition probabilities, and the estimation of these probabilities may be inaccurate, we study a robust version of the ER-MDP model, where the stochastic optimal policies are required to be robust with respect to the ambiguity in the underlying transition probabilities. Our work is at the crossroads of two important schemes in reinforcement learning (RL), namely, robust MDP and entropy-regularized MDP. We show that essential properties that hold for the non-robust ER-MDP and robust unregularized MDP models also hold in our settings, making the robust ER-MDP problem tractable. We show how our framework and results can be integrated into different algorithmic schemes including value or (modified) policy iteration, which would lead to new robust RL and inverse RL algorithms to handle uncertainties. Analyses on computational complexity and error propagation under conventional uncertainty settings are also provided.

1 Introduction

This paper is focused on a robust approach for entropy-regularized Markov Decision Processes (ER-MDP) when the transition probabilities (or dynamics) are themselves uncertain. By studying the robust ER-MDP framework, we aim at providing a theoretical basis to develop new robust reinforcement learning (RL)/planning algorithms to make decisions under dynamics uncertainty, and more accurate inverse reinforcement learning (IRL) algorithms for solving the problem of reward learning when the experts are conservative with respect to the dynamics uncertainty.

Robust MDP is an important framework in the machine learning and operations research literature. The framework is motivated by the fact that, in many practical RL/planning problems, the estimation of dynamics might be far from accurate and optimal policies to Markov Decision problems would be very sensitive with respect to these probabilities [23]. The MDP/RL literature has seen a number of solution methods on how to make robust policies in this uncertainty setting [27, 21]. On the other hand, ER-MDP is another important scheme in the RL/IRL literature with a different motivation. The framework was first proposed by [39] in the context of IRL, i.e., the problem of recovering an expert's reward function from demonstrations, with the advantage of removing ambiguity between demonstrations and the expert policy, and casting the reward learning as a maximum likelihood

estimation problem. This framework then became popular in the IRL literature with many successful algorithms [20, 16, 8]. To the best of our knowledge, existing IRL frameworks/algorithms all assume that the expert knows the dynamics with certainty. Since it might be not the case and the expert would adapt their decisions with respect to the dynamic uncertainty, e.g., being conservative when making decisions, ignoring this uncertainty issue in expert’s demonstrations would lead to inaccurate reward structures. In addition, the ER-MDP has also been popular in the (deep) reinforcement learning (RL) literature with many state-of-the-art *soft*-RL algorithms, e.g. Soft-Actor-Critic [13, 14], with various motivations such as improving exploration, compositionality and robustness in RL. In fact, since the estimation of the dynamics might not be accurate, robust versions of the ER-MDP framework and soft-RL algorithms are relevant and worth exploring, noting that such a robust framework has never been formally studied before.

Given the importance of the ER-MDP framework in the RL/IRL literature and the issue of facing uncertainties when making policies, a robust approach for ER-MDP would provide principled answer to the questions of how to recover an accurate reward function from expert’s robust/conservative demonstrations, and how to be robust in *soft*-RL algorithms when the dynamics are themselves uncertain. This motivates us to introduce and study the robust ER-MDP framework, aiming at proving a complete and rigorous theoretical basic for developing new RL/IRL algorithm that is robust with respect to dynamics uncertainty. More specifically, we explore several aspects such as the duality properties of the robust problem, the complexity of the resulting algorithms and the complexity of the adversary’s problem under different uncertainty settings. We then use these results to design new and efficient robust soft-RL and IRL algorithms.

Our main contributions in this paper are to show that the estimation of the robust optimal policies in robust ER-MDP can be done efficiently ¹, under conventional uncertainty settings, and the complexity is similar to the case of the robust (unregularized) MDP and only modestly larger than the case of non-robust ER-MDP model. More specifically, we consider two conventional uncertainty settings, i.e., (s, a) - and (s) -rectangularity [17, 35] and show that some essential properties such as contraction and Markov optimality hold for the robust ER-MDP model. These properties are important to design tractable algorithms to solve our robust Markov problems. We also point out that the *perfect duality* (or *minimax equality*) that holds for the classical robust MDP model also holds in our setting, noting that our robust ER-MDP problem is more challenging to handle and requires new proofs, as many results that hold for the unregularized MDP do not hold for the ER-MDP, e.g. the Markov problem no-longer can be formulated as a linear program. From these basic properties, we analyze and provide bounds for the computational complexity and error propagation of value function when the adversary’s minimization problems are only solved approximately. We further show how our framework can be used to develop new RL/IRL algorithms. Moreover, since solving the adversary’s minimization problem is a key issue in robust MDP, we extend the results from previous studies [17, 27] by considering uncertainty sets based on several KL divergence bounds and show that the resulting minimization problem can be solved efficiently as well. We also provide numerical results to demonstrate applications of our framework/algorithms in some IRL tasks.

Related work: The machine learning and operations research literatures have seen a number of studies on robust Markov decision processes (MDP) and reinforcement learning in robust MDP [34, 17, 27, 21, 35]. Existing work mostly relies on unregularized MDP, thus makes use of some results that only hold for the unregularized model, e.g., the Markov problem can be formulated as a linear program. On the other hand, the ER-MDP framework has become popular in both RL and IRL literature. In RL, [31] propose a policy iteration scheme, called Trust Region Policy Optimization, in which entropy terms are added to the greedy step to penalize the Kullback-Leibler (KL) divergence between two consecutive policies. The idea of using entropy regularizers to penalize the divergence between consecutive policies has been also used in Dynamic Policy Programming (DPP) [3], Maximum A Posteriori Policy Optimization (MPO) [2, 1], and robust MPO [22]. Some recent RL algorithms have been developed to take advantage of soft value function and soft policies resulting from the ER-MDP scheme, for example, Soft-Q learning [7, 32, 12] and Soft-Actor-Critic [13, 14]. [22] have added robustness to a *soft*-RL (i.e., MPO) algorithm, but their work is experimentally focused and their theoretical explorations are limited, in the sense that many important aspects such as duality properties, complexity and other uncertainty settings were not investigated. In IRL, many state-of-the-art algorithms are based on the ER-MDP framework, e.g., Gaussian Process IRL [20] and generative adversarial IRL [11, 8, 37]. [33] propose a robust IRL algorithm under a

¹Here, “efficiency” means that “the worst-case complexity is polynomial time.”

dynamic mismatch between the expert and learner, but their settings are different, as they assume that the learner does not know the expert’s dynamics with certainty, while in our context the expert is unsure about the dynamics and this information is revealed to the learner. Other types of regularizers have been also studied. For example, [19] propose to use a Tsallis entropy with the motivation of having sparse policies. [9] propose a general MDP framework regularized by any concave function. We will show that our theoretical results can also be applied to these general settings.

Our paper is structured as follows. Section 2 describes our problem setting and Section 3 presents theoretical properties of the robust ER-MDP model. We discuss related algorithms and frameworks in Section 4. Section 5 analyses the computational complexity of the adversary’s problems. Section 6 provides experiments for robust IRL, and finally, Section 7 concludes. We provide all the proofs and relevant discussions in the supplementary material. We use $|\mathcal{S}|$ to denote the cardinality of set \mathcal{S} . Boldface characters represent matrices (or vectors) or a collection of values.

2 Problem description

Consider an infinite-horizon Markov decision process (MDP) for an agent with finite states and actions, defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathbf{Q}, \mathbf{r}, \gamma)$, where \mathcal{S} is a set of states $\mathcal{S} = \{1, 2, \dots, |\mathcal{S}|\}$, \mathcal{A} is a finite set of actions, $\mathbf{Q} = \{\mathbf{q}^0, \dots, \mathbf{q}^\infty\}$ are transition probabilities where $\mathbf{q}^t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is a transition probability function at time t , i.e., $q^t(s_{t+1}|a_t, s_t)$ is the probability of moving to state $s_{t+1} \in \mathcal{S}$ from $s_t \in \mathcal{S}$ by performing action $a_t \in \mathcal{A}$ at time step t , $\mathbf{r} = \{r(a|s), a \in \mathcal{A}, s \in \mathcal{S}\}$ is a reward function, and $\gamma \in [0, 1]$ is a discount factor.

Let $\mathbf{\Pi} = \{\boldsymbol{\pi}^0, \dots, \boldsymbol{\pi}^\infty\}$ be a policy function where $\pi^t(a_t|s_t)$ is the probability of making action $a_t \in \mathcal{A}$ at state $s_t \in \mathcal{S}$ at time $t \in \{0, 1, \dots\}$, the goal of (forward) reinforcement learning under maximum causal entropy principle is to find an optimal policy $\mathbf{\Pi}$ that maximizes the expected entropy-regularized discounted reward [38, 4, 32, 12]

$$\max_{\substack{\boldsymbol{\pi}^t \in \Delta^\pi \\ t=0,1,\dots}} \left\{ F_\infty(\mathbf{\Pi}, \mathbf{Q}) = \mathbb{E}_{\tau \sim (\mathbf{\Pi}, \mathbf{Q})} \left[\sum_{t=0}^{\infty} \gamma^{[t]} r(a_t|s_t) - \gamma^{[t]} \eta \ln \pi^t(a_t|s_t) \right] \right\}, \quad (1)$$

where $\gamma^{[t]}$ refers to “ γ to the power of t ” (we use $[\cdot]$ to distinguish it from a superscript t), $\tau = \{(s_0, a_0), \dots, (s_\infty, a_\infty)\}$ is a strategy in the infinite-horizon case, Δ^π is the set of policies $\Delta^\pi = \{\pi(a|s) \in [0, 1] \mid \sum_{a \in \mathcal{A}} \pi(a|s) = 1, \forall s \in \mathcal{S}\}$, and $\eta \geq 0$ is a regularization coefficient. The term $\mathcal{H}(\boldsymbol{\pi}) = -\mathbb{E}_{\mathbf{\Pi}, \mathbf{Q}}[\sum_{t=0}^{\infty} \gamma^{[t]} \eta \ln \pi^t(a_t|s_t)]$ is referred to as a γ -discounted causal entropy, distinguishing the entropy regularized with the standard MDP one. This term makes the expected discount rewards no-longer linear in $\boldsymbol{\pi}^t$, for $t = 0, \dots, \infty$.

In our problem, we assume that the dynamics (i.e., transition probabilities) are uncertain and the robust model aims at finding a robust policy that maximizes the worst-case expected entropy-regularized reward function. The robust problem can be formulated as

$$\max_{\substack{\boldsymbol{\pi}^t \in \Delta^\pi \\ t=0,1,\dots}} \min_{\substack{\mathbf{q}^t \in \mathcal{Q} \\ t=0,1,\dots}} \left\{ F_\infty(\mathbf{\Pi}, \mathbf{Q}) \right\}, \quad (2)$$

where \mathcal{Q} is an uncertainty set for the dynamics, defined as $\mathcal{Q} \subset \Delta^q = \{\mathbf{q} \mid \sum_{s' \in \mathcal{S}} q(s'|a, s) = 1, \forall (s', s, a)\}$, and \mathbf{q}^t is a vector of transition probabilities chosen by the adversary at time step $t = 0, 1, \dots, \infty$. Here, the uncertainty set \mathcal{Q} are assumed to be (state,action)-wise or (state)-wise decomposable, i.e.. the uncertainty set \mathcal{Q} has the form $\mathcal{Q} = \otimes_{(s,a)} \mathcal{Q}_{sa}$ or $\mathcal{Q} = \otimes_{(s)} \mathcal{Q}_s$, where \mathcal{Q}_{sa} and \mathcal{Q}_s are uncertainty sets for the transition probabilities $\mathbf{q}_{sa} = \{q(s'|s, a), \forall s'\}$ and $\mathbf{q}_s = \{q(s'|s, a) \mid \forall s, a\}$, respectively, for all $a \in \mathcal{A}, s \in \mathcal{S}$. We call these assumptions as (s, a) - and (s) -rectangularity. These assumptions have been widely used to derive tractable solutions for robust MDP problems [27, 17, 35].

3 Theoretical Properties and Algorithms

We present essential theoretical results for the ER-MDP model. These results are critical for the tractability of the robust problems. We then also discuss how to compute robust optimal policies and provide complexity analyses.

3.1 Theoretical Properties

We will show that some basic results holding for the robust unregularized MDP and non-robust ER-MDP models are also valid for our robust one, making the computation of robust optimal policies tractable. To facilitate our exposition, let us first consider the mapping $\mathcal{T}[V] : \mathbb{R}^{|\mathcal{S}|} \rightarrow \mathbb{R}^{|\mathcal{S}|}$

$$\mathcal{T}[V](s) = \max_{\boldsymbol{\pi} \in \Delta^\pi} \min_{\mathbf{q} \in \mathcal{Q}} \left\{ \psi_s(\boldsymbol{\pi}, \mathbf{q}, V) \right\}, \quad \forall s \in \mathcal{S} \quad (3)$$

where $\psi_s(\boldsymbol{\pi}, \mathbf{q}, V) = \mathbb{E}_{\boldsymbol{\pi}_s} [r(a|s) - \eta \ln \pi(a|s) + \gamma \mathbb{E}_{s' \sim \mathbf{q}_{sa}} [V(s')]]$. We also define $\mathcal{T}^\pi[V] = \min_{\mathbf{q} \in \mathcal{Q}} \{ \psi_s(\boldsymbol{\pi}, \mathbf{q}, V) \}$ for a fixed policy $\boldsymbol{\pi}$. On the other hand, let $V^*, V^\pi : \mathcal{S} \rightarrow \mathbb{R}$ be the expected worst-case accumulated rewards under uncertain transition probabilities (value functions)

$$V^\pi(s) = \min_{\substack{\mathbf{q}^t \in \mathcal{Q} \\ t=0,1,\dots}} \left\{ \mathbb{E}_{\boldsymbol{\tau} \sim (\boldsymbol{\Pi}, \mathcal{Q})} \left[\sum_{t=0}^{\infty} \gamma^{[t]} \left(r(a_t|s_t) - \eta \ln \pi^t(a_t|s_t) \right) \middle| s_0 = s \right] \right\} \quad (4)$$

and $V^*(s) = \max_{\boldsymbol{\pi}^t \in \Delta^\pi, t=0,1,\dots} V^\pi(s)$, $\forall s \in \mathcal{S}$. The following theorem focuses on the (s, a) -rectangularity case and shows some main properties of the robust problem. For notational brevity, let us first denote $h(a, s|V) = r(a|s) + \gamma \min_{\mathbf{q}_{sa} \in \mathcal{Q}_{sa}} \{ \mathbb{E}[V(s')] \}$ for any $a \in \mathcal{A}, s \in \mathcal{S}$.

Theorem 3.1 ((s, a)-rectangularity) *Assume that the uncertainty set \mathcal{Q} is (s, a) -rectangular, $\mathcal{T}[V]$ and $\mathcal{T}^\pi[V]$ are contraction mappings of parameters γ and V^* , and V^π are unique solutions to the contraction systems $\mathcal{T}[V] = V$ and $\mathcal{T}^\pi[V] = V$, and the mapping $\mathcal{T}[V]$ can be updated as $\mathcal{T}[V] = \eta \ln \left(\sum_{a \in \mathcal{A}} \exp \left(h(a, s|V)/\eta \right) \right)$, and the policy $\boldsymbol{\pi}^*$ defined as $\pi^*(a|s) = \left(\exp \left(h(a, s|V^*)/\eta \right) \right) / \left(\sum_{a'} \exp \left(h(a', s|V^*)/\eta \right) \right)$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$ is optimal to (2). Moreover, the perfect duality holds for both the mapping $\mathcal{T}[V]$ and $\mathcal{T}^\pi[V]$ and robust expected reward function, i.e., $\max_{\boldsymbol{\pi} \in \Delta^\pi} \min_{\mathbf{q} \in \mathcal{Q}} \{ \psi_s(\boldsymbol{\pi}, \mathbf{q}, V) \} = \min_{\mathbf{q} \in \mathcal{Q}} \max_{\boldsymbol{\pi} \in \Delta^\pi} \{ \psi_s(\boldsymbol{\pi}, \mathbf{q}, V) \}$ and $\max_{\boldsymbol{\pi}^0, \dots} \min_{\mathbf{q}^0, \dots} F_\infty(\boldsymbol{\Pi}, \mathcal{Q}) = \min_{\mathbf{q}^0, \dots} \max_{\boldsymbol{\pi}^0, \dots} F_\infty(\boldsymbol{\Pi}, \mathcal{Q})$.*

The detailed proof is provided in the supplementary (Section A.1). The proof for the contraction property of \mathcal{T} and \mathcal{T}^π shares the same spirit as in the standard robust MDP model [17]. The main difference here is the inclusion of the nonlinear entropy term in the Bellman update. The formulation for the robust optimal policy has a similar form as those from the non-robust ER-MDP model, except that instead of performing the Bellman update with known transition probabilities, we need to compute a minimization value $\min_{\mathbf{q}_{sa} \in \mathcal{Q}_{sa}} \{ \mathbb{E}_{\mathbf{q}_{sa}} [V(s')] \}$. The proof can be done using the fact that the minimization problem $\min_{\mathbf{q}_{sa}}$ can be put inside the expectation in such a way that it does not depend on the policy $\boldsymbol{\pi}$, i.e.,

$$\min_{\mathbf{q} \in \mathcal{Q}} \left\{ \psi_s(\boldsymbol{\pi}, \mathbf{q}, V) \right\} = \mathbb{E}_{\boldsymbol{\pi}} \left[r(a|s) - \eta \ln \pi(a|s) + \gamma \min_{\mathbf{q}_{sa}} \mathbb{E}_{s' \sim \mathbf{q}_{sa}} [V(s')] \right], \quad (5)$$

noting that it is only valid if the uncertainty set \mathcal{Q} are (s, a) -rectangular. [22] give the same formulations, but they do not explicitly show that this solution is optimal to the infinite-horizon problem (2) and corresponds to a saddle point of the *max-min* problem. The *perfect duality* property is interesting in the sense that we can swap the *max-min* order even-though the objective functions $\psi_s(\boldsymbol{\pi}, \mathbf{q}, V)$ and $F_\infty(\cdot)$ are not linear in $\boldsymbol{\pi}$ and the uncertainty set \mathcal{Q} is not necessarily compact and/or convex. We note that in [27], the perfect duality is proved for standard robust MDP using linear programming, which is not applicable in our context.

We now discuss how our results can be extended to the (s) -rectangularity case, which allows the transition probabilities \mathbf{q}_{sa} to be dependent over actions $a \in \mathcal{A}$. The main difference and challenge lie in the fact that adversary's minimization problem involve the policy variable $\boldsymbol{\pi}$, making (5) no-longer valid. The following theorem shows how a robust optimal policy in this case can be efficiently computed. First, let $z(a, s|V, \mathbf{q}) = r(a|s) + \gamma \sum_{s' \in \mathcal{S}} q(s'|a, s)V(s')$ for notational simplicity.

Theorem 3.2 ((s)-rectangularity) *Assume that the uncertainty set \mathcal{Q} is (s) -rectangular and \mathcal{Q}_s are compact and convex, for all $s \in \mathcal{S}$, a robust optimal policy $\boldsymbol{\pi}_s^*$ to (2) can be computed as $\pi^*(a|s) = \left(\exp \left(z(a, s|V^*, \mathbf{q}^*)/\eta \right) \right) / \left(\sum_{a'} \exp \left(z(a', s|V^*, \mathbf{q}^*)/\eta \right) \right)$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$, where V^* is the unique fixed point solution to the contraction mapping $\mathcal{T}[V] = V$, where $\mathcal{T}[V](s) =$*

$\eta \left\{ \ln \left(\sum_{a \in \mathcal{A}} \exp(z(a, s|V, \mathbf{q}^*)/\eta) \right) \right\}$, where and \mathbf{q}_s^* is an optimal solution to the convex optimization problem

$$\min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \sum_{a \in \mathcal{A}} \exp(z(a, s|V^*, \mathbf{q})/\eta) \right\}, \forall s \in \mathcal{S}. \quad (6)$$

Moreover, the perfect duality holds.

The proof can be found in the supplementary material. In this setting, we assume that the uncertainty set is convex and compact in order to use the Von Neumann's minimax theorem to swap the *max-min* order. Note that the perfect duality for the robust unregularized MDP model under (s) -rectangular sets has been proved in [35] using the result that the value function can be expressed as $V^\pi = \sum_{t=0}^{\infty} [\lambda \hat{\mathbf{P}}^\pi]^t \hat{\mathbf{r}}^\pi$, where $\hat{\mathbf{P}}^\pi$ is of size $|\mathcal{S}| \times |\mathcal{S}|$ with entries $\hat{\mathbf{P}}_{ss'}^\pi = \sum_a \pi(a|s)q(s'|a, s)$ and $\hat{\mathbf{r}}_s^\pi = \sum_a \pi(a|s)r(a|s)$ [28]. This result does not apply in our context due to the inclusion of the (nonlinear) entropy terms. Our idea to derive the formulation for the optimal policy is that if we swap the *min-max* order, then the inner maximization problem $\max_{\pi} \{\psi_s(\pi, \mathbf{q}, V)\}$ will yield a closed-form solution and the corresponding *min-max* problem can be converted into a *min* problem with a (strictly) convex objective function, which is way easier to solve than the *max-min* counterpart. For this reason, we assume that the uncertainty set \mathcal{Q} is convex, which is a typical assumption in the robust optimization literature, and show that a solution to the *min-max* problem is also optimal to the *max-min* one (i.e., saddle point). This greatly simplifies the computation.

In the (s) -rectangularity case, the Bellman update requires to solve an exponential convex optimization problem, instead of a linear one. Under the same uncertainty settings as in the (s, a) -rectangularity case (e.g. uncertainty sets based on KL divergence), it seems not possible to efficiently solve the inner problem by bisection. However, it is still possible to solve these problems in polynomial time. We discuss this in detail in Section 5.

3.2 Approximate Robust Value Iteration and Complexity Analysis

In this section we discuss the computation of optimal policies under our robust settings. We focus on value iteration and will talk about other algorithms, e.g., IRL, policy iteration, in the next section. We first consider the (s, a) -rectangularity case and note that the analysis can be further extended to the (s) -rectangularity case.

The contraction mapping implies that the value iteration method converges to a unique fixed point when the number of iterations tends to infinity. Let us define $\mathcal{T}^n[V] = \mathcal{T}[\mathcal{T}^{n-1}[V]]$ for $n = 1, 2, \dots$ and $\mathcal{T}^0[V] = V$, for any $V \in \mathbb{R}^{|\mathcal{S}|}$, and let V^* is the unique fixed point to the mapping $\mathcal{T}[V] = V$. From the contraction property, it is well-known that, to obtain an ϵ -approximation of the fixed point solution, one would need $(\ln \epsilon^{-1} - \ln \|V^*\|_\infty) / \ln \gamma \in \mathcal{O}(\ln \epsilon^{-1})$ iterations.

The mapping $\mathcal{T}[V]$ involves a minimization problem $\min_{\mathbf{q}_{sa}} \mathbb{E}[V(s')]$, which can be solved approximately by bisection [17, 27]. In this section we study the complexity to compute ϵ -approximations of the value function and optimal policy under *soft* Bellman updates. First, to facilitate our exposition, given $\xi > 0$, assume that there is an algorithm of complexity $C(\xi)$ that allows to compute a solution $\bar{\mathbf{q}}_{sa}$ such that, $\forall a \in \mathcal{A}, s \in \mathcal{S}, \min_{\mathbf{q}_{sa}} \mathbb{E}_{\mathbf{q}_{sa}}[V(s')] \geq \mathbb{E}_{\bar{\mathbf{q}}_{sa}}[V(s')] - \xi$. We will examine $C(\xi)$ under different uncertainty sets later in Section 5. Since the adversary's problem can only be solved approximately, $\mathcal{T}[V]$ and optimal policies are approximated, for any $s \in \mathcal{S}$, as

$$\tilde{\mathcal{T}}[V](s) = \eta \ln \left(\sum_{a \in \mathcal{A}} \exp \left(\left(z(a, s|V, \bar{\mathbf{q}}) \right) / \eta \right) \right), \quad \tilde{\pi}_s = \frac{\exp(z(a, s|\tilde{V}, \bar{\mathbf{q}})/\eta)}{\sum_{a'} \exp(z(a', s|\tilde{V}, \bar{\mathbf{q}})/\eta)}, \quad (7)$$

where \tilde{V} is an approximate value function. Theorem 3.3 below examines approximation errors in (7). The results differ from standard robust MDP because we have *soft* approximate optimal policies and $\tilde{\mathcal{T}}[V]$ involves *log* and *exp* functions.

Theorem 3.3 *The approximate Bellman update and policy can be bounded as follows*

$$(i) \quad \|\tilde{\mathcal{T}}^n[V] - \mathcal{T}^n[V]\|_\infty \leq \xi \gamma (1 - \gamma^{[n]}) / (1 - \gamma)$$

- (ii) For any $\epsilon > 0$, if $\xi \leq \epsilon(1 - \gamma)^2/(4\gamma)$ and $\|\tilde{\mathcal{T}}^{n+1}[V] - \tilde{\mathcal{T}}^n[V]\|_\infty \leq 3\epsilon(1 - \gamma)/4$, then $\|\tilde{\mathcal{T}}^n[V] - V^*\|_\infty \leq \epsilon$
- (iii) If we compute a soft policy $\tilde{\pi}$ by an approximate value function \tilde{V} such that $\|\tilde{V} - V^*\|_\infty \leq \epsilon$, for an $\epsilon > 0$, then $\|\tilde{\pi} - \pi^*\|_\infty \leq (e^{2(\epsilon+\xi)/\eta} - 1)$.

Theorem 3.3-(i) is useful to analyze the error propagation of value iteration after a certain number of iterations. This bound also holds for policy evaluation. Theorem 3.3-(ii) answers the questions of when we should stop the value iteration algorithm to achieve a certain level of accuracy, and Theorem 3.3-(iii) shows an approximation error of the approximate optimal policy. We also see that one needs $\mathcal{O}(\ln \epsilon^{-1})$ iterations to get an approximation in (ii). We now analyze the computational complexity to get ϵ -approximations of the value function V^* and the optimal policy π^* . According to Theorem 3.3-(ii) and analyses from Section 5, to get an ϵ -approximation of the value function, it would require a worst-case complexity of $\mathcal{O}(|\mathcal{S}||\mathcal{A}| \max_s \{N_s\} (\ln \epsilon^{-1})^2)$ for uncertainty sets based on a single KL divergence bound, and $\mathcal{O}(|\mathcal{S}||\mathcal{A}| (\max_s \{N_s\})^{7/2} (\ln \epsilon^{-1})^2)$ for the case of several KL divergence bounds, where N_s is the number of states that can be reached from state $s \in \mathcal{S}$, which is typically much smaller than $|\mathcal{S}|$. On the other hand, to get an ϵ -approximation of π^* , using Theorem 3.3-(ii)-(iii), the computation would need complexities of $\mathcal{O}(|\mathcal{S}||\mathcal{A}| \max_s \{N_s\} \ln \epsilon^{-1} \ln((\ln(\epsilon + 1))^{-1}))$ and $\mathcal{O}(|\mathcal{S}||\mathcal{A}| (\max_s \{N_s\})^{7/2} \ln \epsilon^{-1} \ln((\ln(\epsilon + 1))^{-1}))$ for the cases of single KL bound and several KL bounds, respectively. Note that the robust (unregularized) MDP problem has the same complexity bounds, and under non-robust ER-MDP the complexity of getting an ϵ -approximation of the value function becomes $\mathcal{O}(|\mathcal{S}||\mathcal{A}| (\max_s \{N_s\}) \ln \epsilon^{-1})$. Thus, adding robustness to ER-MDP yields an extra computational cost of $\mathcal{O}(\ln \epsilon^{-1})$ for the case of single KL bound and $\mathcal{O}((\max_s N_s)^{5/2} \ln \epsilon^{-1})$ in the case of several KL bounds.

In the (s)-rectangularity case, we perform the Bellman update by solving (6). This is a convex optimization problem and, under some conventional settings, can be solved in polynomial time. Section 5 below shows that a ξ -approximation of the inner minimization problem (with uncertainty sets of several KL bounds) can be achieved with complexity $\mathcal{O}(N_s^{7/2} \ln \xi^{-1})$. As a result, it would require a complexity of $\mathcal{O}(|\mathcal{S}| (\max_s \{|\mathcal{A}| N_s\})^{7/2} (\ln \epsilon^{-1})^2)$ to have an ϵ -approximation of V^* , and a complexity of $\mathcal{O}(|\mathcal{S}| (|\mathcal{A}| \max_s \{N_s\})^{7/2} \ln \epsilon^{-1} \ln((\ln(\epsilon + 1))^{-1}))$ to have an ϵ -approximation of the optimal policy.

4 Applications

We discuss relevant frameworks and algorithms that would make use of our results. This shows broad benefits of using the robust ER-MDP formulations in different contexts.

IRL/Imitation Learning under Uncertainty. The (robust) ER-MDP framework yields soft/randomized optimal policies, making it appealing for imitation learning/IRL [38, 16, 37], as one can conveniently formulate the reward learning problem as maximum likelihood estimation. In general, the robust model will be useful under the assumption that the experts who give demonstrated decisions do not know the transition probabilities with certainty, and make *robust* decisions (looking at worst-case scenarios), noting that the problem of how to make robust decisions in uncertain environments has been widely investigated in the literature [34, 17, 27, 21]. So, it would be valuable to have an IRL algorithm that is able to learn a reward function from expert's robust decisions. Our results allow to cast the reward learning problem as maximum likelihood estimation as in the standard case. A detailed robust IRL algorithm with a complexity analysis are provided in the supplementary material. We provide numerical experiments in Section 6 below to demonstrate the benefits of having such a robust IRL algorithm to recover reward functions from robust policies.

RL with KL Divergence Penalties. Solving an ER-MDP problem can provide an optimal policy that is not *too far* from a given policy. In a planning context, one might be interested in finding a policy that is not far from a given pre-computed policy $\bar{\pi}$. This policy $\bar{\pi}$ may be an outcome of a robust (unregularized) MDP model, but due to some changes to the system (e.g. reward function or uncertainty set \mathcal{Q}), one might need to recompute the robust optimal policy without ending up with a completely new one. To this end, we can penalize the reward function by a KL divergence between the old policy $\bar{\pi}$ and the new one π_s $\text{KL}(\pi_s || \bar{\pi}_s) = \sum_a \pi(a|s) \ln \frac{\pi(a|s)}{\bar{\pi}(a|s)}$, or solve a Markov problem with constraints $\text{KL}(\pi_s || \bar{\pi}_s) \leq \beta$, for a scalar $\beta \geq 0$. The uses of KL divergence penalties

or constraints are generally equivalent due to the fact that the function $\text{KL}(\pi_s || \bar{\pi}_s)$ can be moved to the objective function using Lagrange duality. We then can solve following robust ER-MDP problem to obtain a new optimal policy

$$\max_{\substack{\pi^t \in \Delta^\pi \\ t=0, \dots}} \min_{\substack{\mathbf{q}^t \in \mathcal{Q} \\ t=0, \dots}} \left\{ \mathbb{E}_{\mathbf{n}, \mathbf{Q}} \left[\sum_{t=0}^{\infty} \gamma^{[t]} (r(a_t | s_t) - \eta \text{KL}(\pi_{s_t}^t || \bar{\pi}_{s_t})) \right] \right\}, \quad (8)$$

which yields robust regularized Bellman equation $\mathcal{T}[V] = \max_{\pi_s} \min_{\mathbf{q}_s} \{ \mathbb{E}[r(a|s) - \eta \ln \frac{\pi(a|s)}{\bar{\pi}(a|s)} + \gamma \mathbb{E}_{\mathbf{q}_{sa}} [V(s')]] \}$. Clearly, if we let $\eta \rightarrow \infty$ then the optimal solution to (8) should approaches $\bar{\pi}$ and if $\eta \rightarrow 0$ then we retrieve the robust unregularized MDP model.

ER-MDP has been also used in policy iteration to prevent early convergence to sub-optimal policies, e.g., [2, 1, 22]. So, it would be useful to use it to compute an optimal policy of a robust (unregularized) MDP while solving a robust regularized Bellman equation at each greedy step of a policy iteration algorithm. To facilitate the idea, let consider the modified policy iteration (MPI) approach [29]. Under our uncertainty settings, at an iteration k of the robust MPI algorithm, we need to perform

$$\begin{aligned} (i) \quad \pi_s^{k+1} &= \operatorname{argmax}_{\pi_s} \left\{ \min_{\mathbf{q}_s} \left\{ \mathbb{E}_{\pi_s} [r(a|s) + \mathbb{E}[V^k(s')]] \right\} - \eta \text{KL}(\pi_s || \pi_s^k) \right\} \\ (ii) \quad V^{k+1} &= (\mathcal{T}^{\text{UR}, \pi^{k+1}})^m [V^k], \end{aligned}$$

where $\mathcal{T}^{\text{UR}, \pi^{k+1}} = \min_{\mathbf{q}_s} \mathbb{E}_{\pi_s^{k+1}, \mathbf{q}_s} [r(a|s) + \gamma V(s')]$. Here, $\mathcal{T}^{\text{UR}, \pi^k}$ is the robust (unregularized) Bellman update under policy π^k and the entropy term $\eta \text{KL}(\pi_s || \pi_s^k)$ is used to control the distance between π^k and π^{k+1} . Note that robust policy iteration algorithms without the KL entropy terms have been studied in some previous work [18, 15]. Now, from an initial policy π^0 , the algorithm iteratively find new policy by performing the robust regularized step (i) and robust (unregularized) policy evaluation step (ii). Clearly, if $m = 1$ we retrieve the robust value iteration considered in Section 3.2, and with a sufficiently large m we retrieve a policy evaluation step. Since we only solve the inner minimization approximately, it is important to look at the approximation errors of Steps (i) and (ii). Theorem 3.3-(iii) tells us that if we solve the inner minimization with approximation error $\epsilon > 0$, then we can obtain a policy $\tilde{\pi}^{k+1}$ with approximation error $e^{2\epsilon/\eta} - 1$. For Step (ii), the bound in Theorem 3.3-(i) applies to an approximate Bellman update $\tilde{\mathcal{T}}^{\text{UR}, \pi^k}$ where the *min* problem is solved approximately. Hence, the approximation error for Step (ii) is $\epsilon\gamma(1 - \gamma^m)/(1 - \gamma)$. These bounds allow to analyze the error propagation of the robust MPI (and thus, its convergence and rate of convergence), analogously to non-robust MPI algorithms [30, 9], and bound the complexity required for the two steps to get a certain level of accuracy.

Robust General Regularized MDP. Beside entropy-regularized models, other types of regularizers have been considered. For example, [19] propose to use Tsallis entropy with the motivation of having sparse optimal policies. [9] study a general version of the ER-MDP model by replacing the entropy terms by any convex functions of π_s . It is possible to show that the basic properties mentioned in Theorem 3.3 still hold for the robust version of that general model, but the robust Bellman update might have no closed-forms and would be more difficult to perform. In some cases, one may only do it approximately, thus producing an additional level of approximation to value/policy iteration. We briefly discuss these in the supplementary (Section B.5).

5 Uncertainty Models

A key issue when solving robust MDP problems is to efficiently solve the adversary's minimization problems. We discuss this under both (s, a) - and (s) -rectangularity cases, noting that the objective function in the latter case is exponential. We focus on uncertainty sets based on KL divergence (relative entropy or likelihood models) due to their appealing statistical properties in modeling uncertainties [27, 17]. Previous studies show that if the uncertainty set involves only one KL divergence bound, then in the (s, a) -rectangularity case, the inner minimization can be solved efficiently by bisection. We further extend these results by examining the (s) -rectangularity case and uncertainty sets based on several KL divergence bounds, with the motivation of better use the availability of historical data and migrate the conservativeness of the uncertainty sets. We present our main results below and refer the reader to the supplementary (Section B.6) for detailed proofs.

(s, a) -rectangular uncertainty sets. In this setting, the objective function of the inner minimization problem is linear in \mathbf{q}_{sa} . In a likelihood model or relative entropy model, the uncertainty sets \mathcal{Q}_s are determined by KL divergence constraints of the forms $\sum_{s'} \hat{q}(s'|s, a) \ln q(s'|s, a) \geq \beta$ or $\sum_{s'} q(s'|s, a) \ln(q(s'|s, a)/\hat{q}(s'|s, a)) \leq \beta$, respectively, where $\hat{q}(s'|s, a)$ is an empirical estimate of the transition probability associated with states $s, s' \in \mathcal{S}$ and action $a \in \mathcal{A}$. In [17] the authors show that if \mathcal{Q}_{sa} is defined based on only one KL bound, then the inner problem can be solved by bisection with complexity $C(\xi) = \mathcal{O}(N_s \ln \xi^{-1})$. If we define uncertainty sets using several KL bounds, the bisection no-longer works. However, using some results from convex programming [25], we can show that if the number of KL bounds is much less than N_s (which is typically the case), then the inner problem can be solved by interior-point with complexity $\mathcal{O}(N_s^{7/2} \ln \xi^{-1})$.

(s) -rectangular uncertainty sets. In this case, the inner minimization problems involve exponential (convex) objective function: $\min_{\mathbf{q}_s} \sum_a \exp(r(a|s) + \mathbb{E}_{\mathbf{q}_s}[V(s')])$, making it not solvable by bisection, even with uncertainty sets of only one KL bound. In this context, the problem still can be solved efficiently by interior-point and it is possible to show that, if the number of KL bounds is much smaller than N_s , then the complexity can be bounded by $\mathcal{O}((|\mathcal{A}|N_s)^{7/2} \ln \xi^{-1})$, for which we provide a detailed proof in the supplementary. If the number of KL bounds is significant, then we also provide a detailed bound for the complexity in the supplementary material. It is worth noting that in a general regularized MDP model [9], there might be no closed-form for $\max_{\pi_s} \{\cdot\}$ problems, thus one needs to solve $\max_{\pi_s} \min_{\mathbf{q}_s} \{\cdot\}$ to perform the Bellman update, for which a saddle-point algorithm would be useful [e.g. 10], but the complexity is not easy to bound.

6 Experiments with Robust IRL

We provide numerical experiments to demonstrate the application of our robust ER-MDP models and algorithms in IRL. Here we focus on IRL, noting that extensive experiments for a robust soft-RL were provided in [22]. We employ the MaxEnt algorithm [39], one of the most popular IRL algorithms in the literature. We assume that the experts are uncertain about the dynamics and make robust decisions and our aim is to recover the experts' reward function from such robust decisions. In this context, the standard MaxEnt algorithm will ignore the uncertainties and tries to learn the reward function using a fixed vector of transition probabilities, and our robust version (named as Robust MaxEnt) will explicitly take the uncertainty issue into consideration.

To evaluate how each algorithm performs, in analogy to prior IRL work [20], we use the “*expected value difference*” score, which measures how a learned policy performs under the true rewards. We will evaluate IRL outputs on both environments on which they were learned and random environments (denoted by “*transfer*”). For the latter, we bring the learned parameters of the rewards to compute rewards and optimal policies in the new environments. We will test our robust IRL algorithm using two simulated environments, i.e., Objectworld and Highway Driving Behavior. Brief descriptions are given below.

The **Objectworld** is an $N \times N$ grid of states in which objects are randomly placed. Each object is assigned one of C inner and outer colors. At each state, there are five possible actions corresponding to staying at the same place or stepping to four different directions (up, down, left, right). For the **Highway Driving Behavior** environment, the task is to navigate a vehicle in a highway of three lanes with all vehicles moving with the same speed. The agent's vehicle can switch lanes and drive at up to four times speed of the traffic. Other vehicles (motorcycle or car) are civilian or police, and are placed randomly on the three lanes. The agent can make five different actions of changing lanes (left or right), speeding or slowing down, or staying at the same lane and same speed. Demonstrations are paths of length 8 generated by the true rewards, true dynamics and robust behavior. We generate 128 samples for each score measures and we repeat the training and evaluation 8 times to compute the means and standard errors of the scores. We test the algorithms with two ways of generating expert trajectories, that is, standard unregularized MDP (*deterministic policy*) and ER-MDP (*stochastic policy*). We use the code and data used in [20] and keep the the same settings. The experiments were conducted using a PC running Window 10 with Intel(R) CoreTM i7-7700HQ CPU (2.80Hz) and 16 GB RAM.

We define the uncertainty sets as $\mathcal{Q}_{sa} = \{\mathbf{q}_{sa} \mid \text{KL}(\mathbf{q}_{sa} \parallel \mathbf{q}_{sa}^0) \leq \epsilon\}$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$ where \mathbf{q}_{sa}^0 are the “*true*” transition probabilities, noting that these true values are not known by the experts and we used \mathbf{q}_{sa}^0 to compute the *expected value difference* scores. In this context, ϵ represents an uncertainty

level. That is, larger ϵ will correspond to more uncertain expert behavior. We vary ϵ from 0 to 0.1 to show the performance of the robust IRL algorithm with different ϵ .

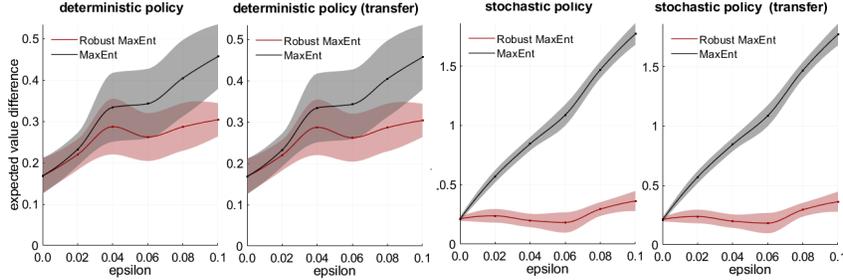


Figure 1: Experiments with objectworld, solid curves show the mean and the shading shows the standard errors.

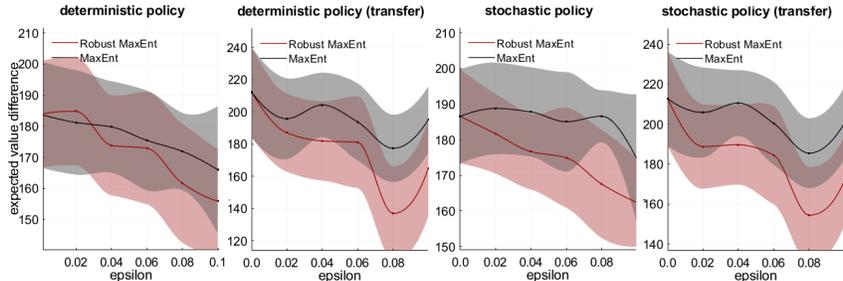


Figure 2: Highway driving behavior experiments, solid curves shows the means and shading shows standard errors.

The means and standard errors of the expected value difference scores for the Objectworld and Highway Driving Behavior environments are plotted in Fig. 1 and Fig. 2, in which the lower the better. It is clear that the Robust MaxEnt constantly outperforms the standard MaxEnt for all the tests, especially for the Objectworld example. The performance gap also increases when ϵ grows, demonstrating the consequences of ignoring the uncertainty issues in IRL.

7 Conclusion

We study a robust ER-MDP model, aiming at taking the advantages of both robust MDP and ER-MDP schemes to develop new algorithms for robust decision-making and learning. We show that several properties that hold in the robust- and ER-MDP models also hold in the robust ER-MDP one. From that, we look at the computation of robust optimal policies and providing computational complexity and error propagation analyses. We show how our robust framework can be used to design robust IRL and robust policy iteration algorithms under dynamics uncertainty. We provide numerical experiments to demonstrate the application of our frameworks/algorithms in the context of IRL.

In this paper we focus on planing settings, i.e., all the information of the MDP is given, except that the dynamics are only known partially. In the context that the environment is unknown and one needs to interact with it to make policies, the algorithms and approximation bounds would need additional work and we keep this for future work. Our robust model might be conservative, in the sense that we assume the dynamics can take any values in the uncertainty set and the uncertainty set needs to satisfy some rectangularity assumptions. Some ways to relax these assumptions are to use distributionally robust approaches [36] and/or k-rectangular robust MDP [24], which would be promising for future work.

References

[1] Abbas Abdolmaleki, Jost Tobias Springenberg, Jonas Degraeve, Steven Bohez, Yuval Tassa, Dan Belov, Nicolas Heess, and Martin Riedmiller. Relative entropy regularized policy iteration. *arXiv preprint arXiv:1812.02256*, 2018.

- [2] Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*, 2018.
- [3] Mohammad Gheshlaghi Azar, Vicenç Gómez, and Hilbert J Kappen. Dynamic policy programming. *Journal of Machine Learning Research*, 13(Nov):3207–3245, 2012.
- [4] Michael Bloem and Nicholas Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. In *53rd IEEE Conference on Decision and Control*, pages 4911–4916. IEEE, 2014.
- [5] Felix Brandt, Markus Brill, and Warut Suksompong. An ordinal minimax theorem. *Games and Economic Behavior*, 95:107–112, 2016.
- [6] Benjamin Eysenbach and Sergey Levine. If MaxEnt RL is the answer, what is the question? *arXiv preprint arXiv:1910.01913*, 2019.
- [7] Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. *arXiv preprint arXiv:1512.08562*, 2015.
- [8] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.
- [9] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. *arXiv preprint arXiv:1901.11275*, 2019.
- [10] Gauthier Gidel, Tony Jebara, and Simon Lacoste-Julien. Frank-wolfe algorithms for saddle point problems. *arXiv preprint arXiv:1610.07797*, 2016.
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [12] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1352–1361. JMLR. org, 2017.
- [13] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *International conference on machine learning*, 2018.
- [14] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.
- [15] Chin Pang Ho and Marek Petrik. Fast bellman updates for robust MDPs. In *ICML*, 2018.
- [16] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573, 2016.
- [17] Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.
- [18] David L Kaufman and Andrew J Schaefer. Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410, 2013.
- [19] Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Sparse markov decision processes with causal sparse tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018.
- [20] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Nonlinear inverse reinforcement learning with gaussian processes. In *Advances in Neural Information Processing Systems*, pages 19–27, 2011.

- [21] Shiau Hong Lim, Huan Xu, and Shie Mannor. Reinforcement learning in robust markov decision processes. In *Advances in Neural Information Processing Systems*, pages 701–709, 2013.
- [22] Daniel J Mankowitz, Nir Levine, Rae Jeong, Abbas Abdolmaleki, Jost Tobias Springenberg, Timothy Mann, Todd Hester, and Martin Riedmiller. Robust reinforcement learning for continuous control with model misspecification. *arXiv preprint arXiv:1906.07516*, 2019.
- [23] Shie Mannor, Duncan Simester, Peng Sun, and John N Tsitsiklis. Bias and variance in value function estimation. In *Proceedings of the twenty-first international conference on Machine learning*, page 72, 2004.
- [24] Shie Mannor, Ofir Mebel, and Huan Xu. Lightning does not strike twice: Robust mdps with coupled uncertainty. *arXiv preprint arXiv:1206.4643*, 2012.
- [25] Arkadi Nemirovski. Interior point polynomial time methods in convex programming. *Lecture notes*, 2004.
- [26] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. Siam, 1994.
- [27] Arnab Nilim and Laurent El Ghaoui. Robustness in markov decision problems with uncertain transition matrices. In *Advances in Neural Information Processing Systems*, pages 839–846, 2004.
- [28] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- [29] Martin L Puterman and Moon Chirl Shin. Modified policy iteration algorithms for discounted markov decision problems. *Management Science*, 24(11):1127–1137, 1978.
- [30] Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon, Boris Lesner, and Matthieu Geist. Approximate modified policy iteration and its application to the game of tetris. *Journal of Machine Learning Research*, 16(49):1629–1676, 2015.
- [31] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.
- [32] John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft Q-learning. *arXiv preprint arXiv:1704.06440*, 2017.
- [33] Luca Viano, Yu-Ting Huang, Parameswaran Kamalaruban, Adrian Weller, and Volkan Cevher. Robust inverse reinforcement learning under transition dynamics mismatch. *arXiv preprint arXiv:2007.01174*, 2020.
- [34] Chelsea C White III and Hany K Eldeib. Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749, 1994.
- [35] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.
- [36] Huan Xu and Shie Mannor. Distributionally robust markov decision processes. In *NIPS*, pages 2505–2513, 2010.
- [37] Lantao Yu, Jiaming Song, and Stefano Ermon. Multi-agent adversarial inverse reinforcement learning. In *International Conference on Machine Learning*, pages 7194–7201. PMLR, 2019.
- [38] Brian Ziebart. Modeling interaction via the principle of maximum causal entropy. *PhD thesis, Carnegie Mellon University*, 2010.
- [39] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

Supplementary Material

A Proofs of Main Results

A.1 Proof of Theorem 3.1

Contraction property. We prove the *Contraction* property. For any $s \in \mathcal{S}$, let us consider two cases $\mathcal{T}[V](s) \geq \mathcal{T}[V'](s)$ or $\mathcal{T}[V](s) < \mathcal{T}[V'](s)$. If $\mathcal{T}[V](s) \geq \mathcal{T}[V'](s)$. For any $\epsilon > 0$, let $\pi^* \in \Delta^\pi$ be a solution such that

$$\mathcal{T}[V](s) \leq \min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \mathbb{E}_{\pi^*} [r(a|s) - \eta \ln \pi^*(a|s) + \gamma \mathbb{E}_{s' \sim \mathbf{q}_{sa}} [V(s')]] \right\} + \epsilon.$$

Since, $\mathcal{T}[V](s) \geq \mathcal{T}[V'](s)$, we have

$$\begin{aligned} |\mathcal{T}[V](s) - \mathcal{T}[V'](s)| &\leq \min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \mathbb{E}_{\pi^*} [r(a|s) - \eta \ln \pi^*(a|s) + \gamma \mathbb{E}_{s' \sim \mathbf{q}_s} V(s')] \right\} + \epsilon \\ &\quad - \min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \mathbb{E}_{\pi^*} [r(a|s) - \eta \ln \pi^*(a|s) + \gamma \mathbb{E}_{s' \sim \mathbf{q}_s} V'(s')] \right\} \\ &= \min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \gamma \mathbb{E}_{\pi^*, s' \sim \mathbf{q}_s} [V(s')] \right\} - \min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \gamma \mathbb{E}_{\pi^*, s' \sim \mathbf{q}_s} [V'(s')] \right\} + \epsilon. \quad (9) \end{aligned}$$

We see that $\min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \gamma \mathbb{E}_{\pi^*, s' \sim \mathbf{q}_s} V(s') \right\} \leq \min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \gamma \mathbb{E}_{\pi^*, s' \sim \mathbf{q}_s} V'(s') \right\}$, so if we denote by $\mathbf{q}_s^* \in \mathcal{Q}_s$ a solution such that

$$\min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \gamma \mathbb{E}_{\pi^*, s' \sim \mathbf{q}_s} [V'(s')] \right\} \geq \mathbb{E}_{\pi^*, s' \sim \mathbf{q}_s^*} [V'(s')] - \epsilon,$$

then from (9) we have

$$|\mathcal{T}[V](s) - \mathcal{T}[V'](s)| \leq \gamma \mathbb{E}_{\pi^*, s' \sim \mathbf{q}_s^*} (V(s') - V'(s')) + 2\epsilon \leq \gamma \|V - V'\|_\infty + 2\epsilon, \forall s \in \mathcal{S}.$$

Let $\epsilon \rightarrow \infty$ we obtain $\|\mathcal{T}[V] - \mathcal{T}[V']\|_\infty \leq \|V - V'\|_\infty$. The case $\mathcal{T}[V](s) < \mathcal{T}[V'](s)$ can be done in a similar way. So $\mathcal{T}[V]$ is a contraction. The contraction property of $\mathcal{T}^\pi[V]$ can be proved in a similar way.

Markov optimality. We will now prove that if V^* is a unique solution to the contraction mapping $\mathcal{T}[V] = V$, then V^* will satisfies $V^*(s) = \max_{\pi^t \in \Delta^\pi, t=0,1,\dots} V^\pi(s)$, $\forall s \in \mathcal{S}$. To this end, we first denote $h(a_t, s_t) = r(a_t|s_t) - \eta \ln \pi^t(a_t|s_t)$, $s \in \mathcal{S}, a \in \mathcal{A}$. For any policy $\pi^0, \pi^1, \dots \in \Delta^\pi$, from the definition of $\mathcal{T}[V]$ we have

$$\begin{aligned} V^*(s) &\geq \min_{\mathbf{q}_s^0} \left\{ \mathbb{E}_{\pi^0} \left[h(a_0, s_0) + \gamma \mathbb{E}_{s_1 \sim \mathbf{q}_{s_0 a_0}^0} [V^*(s_1)] \right] \right\} \\ &\geq \min_{\mathbf{q}^0, \mathbf{q}^1} \left\{ \mathbb{E}_{\pi^0, \pi^1} [h(a_0, s_0) + \gamma h(a_1, s_1)] + \gamma^2 \mathbb{E}_{\mathbf{q}^1} [V^*(s_2) | s_0 = s] \right\}. \end{aligned}$$

This leads to, for any $n \in \mathbb{N}$,

$$\begin{aligned} V^*(s) &\geq \min_{\mathbf{q}^0, \dots, \mathbf{q}^n} \left\{ \mathbb{E}_{\pi^0, \dots, \pi^n} \left[\sum_{t=0}^n \gamma^t h(a_t, s_t) \mid s_0 = s \right] + \mathbb{E}_{\pi^t, \mathbf{q}^t} \left[\gamma^{[n+1]} V^*(s_{n+1}) \mid s_0 = s \right] \right\} \\ &= \min_{\mathbf{q}^0, \dots} \left\{ \mathbb{E}_{\pi^t, \mathbf{q}^t} \left[\sum_{t=0}^{\infty} \gamma^t h(a_t, s_t) \mid s_0 = s \right] + \mathbb{E}_{\pi^t, \mathbf{q}^t} \left[\gamma^{[n+1]} V^*(s_{n+1}) \mid s_0 = s \right] \right. \\ &\quad \left. - \mathbb{E}_{\pi^{n+1}, \dots} \left[\sum_{t=n+1}^{\infty} \gamma^t h(a_t, s_t) \right] \right\} \\ &\geq \min_{\mathbf{q}^0, \dots} \left\{ \mathbb{E}_{\pi^0, \dots} \left[\sum_{t=0}^{\infty} \gamma^t h(a_t, s_t) \mid s_0 = s \right] \right\} - \gamma^{[n+1]} \|V^*\|_\infty - \frac{\gamma^{[n+1]} H}{1 - \gamma}, \end{aligned}$$

where $\|V^*\|_\infty = \max_s V^*(s)$ and

$$H = \max_{\substack{a_t \in \mathcal{A}, s_t \in \mathcal{S} \\ \mathbf{q}^0, \dots, \mathbf{q}^t}} \mathbb{E}_{\substack{\boldsymbol{\pi}^0, \dots, \boldsymbol{\pi}^t \\ \mathbf{q}^0, \dots, \mathbf{q}^t}} [r(a_t | s_t) - \eta \ln \pi(a_t | s_t) | s_0 = s].$$

We can show that $H < \infty$ because

$$\begin{aligned} H &\leq \max_{\substack{a_t \in \mathcal{A}, s_t \in \mathcal{S} \\ \mathbf{q}^0, \dots, \mathbf{q}^t}} \mathbb{E}_{\substack{\boldsymbol{\pi}^0, \dots, \boldsymbol{\pi}^t \\ \mathbf{q}^0, \dots, \mathbf{q}^t}} [r(a_t | s_t) | s_0 = s] + \mathbb{E}_{\substack{\boldsymbol{\pi}^0, \dots, \boldsymbol{\pi}^t \\ \mathbf{q}^0, \dots, \mathbf{q}^t}} [-\eta \ln \pi(a_t | s_t) | s_0 = s] \\ &\leq R + \max_{\substack{a_t \in \mathcal{A}, s_t \in \mathcal{S} \\ \mathbf{q}^0, \dots, \mathbf{q}^t}} -\eta \pi(a_t | s_t) \ln \pi(a_t | s_t) \\ &\leq R + \eta/e, \end{aligned} \tag{10}$$

where $R = \max_{a,s} r(a|s)$ and e the base of the natural logarithm ($e \approx 2.7828$). Inequality (10) is due to the fact that $-x \ln x \leq e^{-1}$ for all $x \in [0, 1]$. So we have

$$V^*(s) \geq \max_{\boldsymbol{\pi}^0, \dots, \mathbf{q}^0, \dots} \min_{\mathbf{q}^0, \dots} \left\{ \mathbb{E}_{\substack{\boldsymbol{\pi}^0, \dots \\ \mathbf{q}^0, \dots}} \left[\sum_{t=0}^{\infty} \gamma^{[t]} h(a_t, s_t) \middle| s_0 = s \right] \right\} - \gamma^{[n+1]} \|V^*\|_\infty - \frac{\gamma^{[n+1]} H}{1 - \gamma}, \tag{11}$$

noting that $\|V^*\|_\infty$ is also finite. On the other hand, there is a policy $\bar{\boldsymbol{\pi}}^0, \bar{\boldsymbol{\pi}}^1, \dots$ such that, for any $\epsilon > 0$ and any $s \in \mathcal{S}$

$$V^*(s) \leq \min_{\mathbf{q}^0} \left\{ \mathbb{E}_{\bar{\boldsymbol{\pi}}^0} [h(a_0, s) + \gamma \mathbb{E}_{s' \sim \mathbf{q}_{s_0}^0} [V(s')]] \right\} + \epsilon.$$

We can expand the Bellman equation to obtain

$$\begin{aligned} V^*(s) &\leq \min_{\mathbf{q}^0} \left\{ \mathbb{E}_{\bar{\boldsymbol{\pi}}^0} [h(a_0, s) + \gamma \mathbb{E}_{s_1 \sim \mathbf{q}_{s_0}^0} [V(s_1)]] \right\} + \epsilon \\ &\leq \min_{\substack{\mathbf{q}^0, \mathbf{q}^1 \\ \bar{\boldsymbol{\pi}}^0, \bar{\boldsymbol{\pi}}^1}} \left\{ \mathbb{E}_{\substack{\mathbf{q}^0, \mathbf{q}^1 \\ \bar{\boldsymbol{\pi}}^0, \bar{\boldsymbol{\pi}}^1}} [h(a_0, s_0) + \gamma h(a_1, s_1) | s_0 = s] + \gamma^2 \mathbb{E}_{\substack{\mathbf{q}^0, \mathbf{q}^1 \\ \bar{\boldsymbol{\pi}}^0, \bar{\boldsymbol{\pi}}^1}} [V(s_2) | s_0 = s] \right\} + (1 + \gamma)\epsilon. \end{aligned}$$

Thus, by continuing expanding the inequality, we have, for any $n \in \mathbb{N}$,

$$\begin{aligned} V^*(s) &\leq \min_{\mathbf{q}^0, \dots, \mathbf{q}^n} \left\{ \mathbb{E}_{\substack{\mathbf{q}^0, \dots, \mathbf{q}^n \\ \bar{\boldsymbol{\pi}}^0, \dots, \bar{\boldsymbol{\pi}}^n}} \left[\sum_{t=0}^n \gamma^{[t]} h(a_t, s_t) \middle| s_0 = s \right] + \gamma^{[n+1]} \mathbb{E}_{\substack{\mathbf{q}^0, \dots, \mathbf{q}^n \\ \bar{\boldsymbol{\pi}}^0, \dots, \bar{\boldsymbol{\pi}}^n}} [V^*(s_{n+1}) | s_0 = s] \right\} \\ &\quad + \frac{1 - \gamma^{[n+1]}}{1 - \gamma} \epsilon \\ &\leq \min_{\substack{\mathbf{q}^0, \dots \\ \bar{\boldsymbol{\pi}}^0, \dots}} \left\{ \mathbb{E}_{\substack{\mathbf{q}^0, \dots \\ \bar{\boldsymbol{\pi}}^0, \dots}} \left[\sum_{t=0}^{\infty} \gamma^{[t]} h(a_t, s_t) \middle| s_0 = s \right] + \gamma^{[n+1]} \|V^*\|_\infty - \right. \\ &\quad \left. \mathbb{E}_{\substack{\mathbf{q}^0, \dots \\ \bar{\boldsymbol{\pi}}^0, \dots}} \left[\sum_{t=n+1}^{\infty} \gamma^{[t]} h(a_t, s_t) \middle| s_0 = s \right] \right\} + \frac{1 - \gamma^{[n+1]}}{1 - \gamma} \epsilon \\ &\leq \min_{\substack{\mathbf{q}^0, \dots \\ \bar{\boldsymbol{\pi}}^0, \dots}} \left\{ \mathbb{E}_{\substack{\mathbf{q}^0, \dots \\ \bar{\boldsymbol{\pi}}^0, \dots}} \left[\sum_{t=0}^{\infty} \gamma^{[t]} h(a_t, s_t) \middle| s_0 = s \right] \right\} + \gamma^{[n+1]} \|V^*\|_\infty + \frac{\gamma^{[n+1]} H}{1 - \gamma} + \frac{1 - \gamma^{[n+1]}}{1 - \gamma} \epsilon. \end{aligned}$$

So we have

$$V^*(s) \leq \max_{\boldsymbol{\pi}^0, \dots} \min_{\mathbf{q}^0, \dots} \left\{ \mathbb{E}_{\substack{\boldsymbol{\pi}^0, \dots \\ \mathbf{q}^0, \dots}} \left[\sum_{t=0}^{\infty} \gamma^{[t]} h(a_t, s_t) \middle| s_0 = s \right] \right\} + \gamma^{[n+1]} \|V^*\|_\infty + \frac{\gamma^{[n+1]} H}{1 - \gamma} + \frac{1 - \gamma^{[n+1]}}{1 - \gamma} \epsilon \tag{12}$$

From (11) and (12), we can take $n \rightarrow \infty$ and ϵ arbitrarily small, we have

$$V^*(s) = \max_{\boldsymbol{\pi}^0, \dots} \min_{\mathbf{q}^0, \dots} \left\{ \mathbb{E}_{\substack{\boldsymbol{\pi}^0, \dots \\ \mathbf{q}^0, \dots}} \left[\sum_{t=0}^{\infty} \gamma^{[t]} h(a_t, s_t) \middle| s_0 = s \right] \right\},$$

as desired. From the proof, we can also validate see that if V^π is a solution to the system $\mathcal{T}^\pi[V] = V$, then it also satisfies Eq. 4.

Perfect duality. We move to the *Perfect Duality* property. We first prove that this property holds for the Bellman update. The weak duality implies that

$$\min_{\mathbf{q}_s \in \mathcal{Q}_s} \max_{\boldsymbol{\pi}_s \in \Delta_s^\pi} \psi_s(\boldsymbol{\pi}, \mathbf{q}, V) \geq \max_{\boldsymbol{\pi}_s \in \Delta_s^\pi} \min_{\mathbf{q}_s \in \mathcal{Q}_s} \psi_s(\boldsymbol{\pi}, \mathbf{q}, V)$$

To prove the opposite inequality, for any $\epsilon > 0$, let $\bar{\mathbf{q}}_s \in \mathcal{Q}_s$ be a transition solution such that

$$\min_{\mathbf{q}_{s\alpha} \in \mathcal{Q}_{s\alpha}} \left\{ \mathbb{E}_{s' \sim \mathbf{q}_{s\alpha}} [V(s')] \right\} \geq \mathbb{E}_{s' \sim \bar{\mathbf{q}}_{s\alpha}} [V(s')] - \epsilon.$$

We have the chain

$$\begin{aligned} & \min_{\mathbf{q}_s \in \mathcal{Q}_s} \max_{\boldsymbol{\pi}_s \in \Delta_s^\pi} \psi_s(\boldsymbol{\pi}, \mathbf{q}, V) \leq \max_{\boldsymbol{\pi}_s \in \Delta_s^\pi} \psi_s(\boldsymbol{\pi}, \bar{\mathbf{q}}, V) \\ & = \max_{\boldsymbol{\pi}_s \in \Delta_s^\pi} \left\{ \mathbb{E}_{\boldsymbol{\pi}_s} \left[r(a|s) - \eta \ln \pi(a|s) + \mathbb{E}_{s' \sim \bar{\mathbf{q}}_{s\alpha}} [V(s')] \right] \right\} \\ & \leq \max_{\boldsymbol{\pi}_s \in \Delta_s^\pi} \left\{ \mathbb{E}_{\boldsymbol{\pi}_s} \left[r(a|s) - \eta \ln \pi(a|s) + \min_{\mathbf{q}_{s\alpha}} \mathbb{E}_{s' \sim \mathbf{q}_{s\alpha}} [\bar{V}(s')] \right] \right\} + \epsilon \\ & = \max_{\boldsymbol{\pi}_s \in \Delta_s^\pi} \min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \mathbb{E}_{\boldsymbol{\pi}_s} \left[r(a|s) - \eta \ln \pi(a|s) + \mathbb{E}_{s' \sim \mathbf{q}_{s\alpha}} [V(s')] \right] \right\} + \epsilon. \end{aligned}$$

Let $\epsilon \rightarrow 0$ we obtain the opposite-side of the inequality, which implies the perfect duality for the Bellman update. To prove the perfect duality for $\max_{\boldsymbol{\pi}} \min_{\mathbf{q}} F_\infty(\boldsymbol{\pi}, \mathbf{q})$ we define the dual of the mapping $\mathcal{T}[V]$ as $\hat{\mathcal{T}}[V] = \max_{\boldsymbol{\pi}} \min_{\mathbf{q}} \psi_s(\boldsymbol{\pi}, \mathbf{q}, V)$. Since $\mathcal{T}[V] = \hat{\mathcal{T}}[V]$ for any $V \in \mathbb{R}^{|\mathcal{S}|}$, they yield the same fixed point solution V^* . Now, similarly to the proof of the Markovian Optimality property, we can also show that V^* will satisfy

$$V^*(s) = \min_{\mathbf{q}^0, \dots, \boldsymbol{\pi}^0, \dots} \max_{\mathbf{q}^0, \dots} \left\{ \mathbb{E}_{\boldsymbol{\pi}^0, \dots} \left[\sum_{t=0}^{\infty} \gamma^t h(a_t, s_t) \mid s_0 = s \right] \right\}.$$

Combining with the *Markov Optimality* shown in point (iv), we obtain the minimax equality

$$\max_{\boldsymbol{\pi}^0, \dots, \mathbf{q}^0, \dots} \min_{\boldsymbol{\Pi}, \mathbf{Q}} F_\infty(\boldsymbol{\Pi}, \mathbf{Q}) = \min_{\boldsymbol{\Pi}, \mathbf{Q}} \max_{\mathbf{q}^0, \dots, \boldsymbol{\pi}^0, \dots} F_\infty(\boldsymbol{\Pi}, \mathbf{Q}),$$

which completes the proof.

To compute an optimal policy to the Markov problem, according to the Markov Optimality property of Theorem 3.1, we just need to find a solution by solving the Bellman equation $\mathcal{T}[V] = V$. Given any state $s \in \mathcal{S}$, we have

$$\begin{aligned} & \max_{\boldsymbol{\pi}_s} \min_{\mathbf{q}_s} \left\{ \mathbb{E}_{\boldsymbol{\pi}_s} \left[r(a|s) - \eta \ln \pi(a|s) + \mathbb{E}_{\mathbf{q}_{s\alpha}} [V(s')] \right] \right\} \\ & = \max_{\boldsymbol{\pi}_s} \left\{ \mathbb{E}_{\boldsymbol{\pi}_s} \left[r(a|s) - \eta \ln \pi(a|s) + \gamma \min_{\mathbf{q}_{s\alpha}} \mathbb{E}_{\mathbf{q}_{s\alpha}} [V(s')] \right] \right\}. \end{aligned}$$

So we can write

$$V(s) = \max_{\boldsymbol{\pi}_s} \left\{ \sum_{a \in \mathcal{A}} \pi(a|s) \left(r(a|s) - \eta \ln \pi(a|s) + \delta(s, a) \right) \right\},$$

where $\delta(s, a) = \min_{\mathbf{q}_{s\alpha}} \sum_{s' \in \mathcal{S}} q(s'|s, a) V(s')$ for notational brevity. Let consider the maximization problem

$$\begin{aligned} J(s) &= \max_{\boldsymbol{\pi}_s} \sum_{a \in \mathcal{A}} \pi(a|s) \left(r(a|s) - \eta \ln \pi(a|s) + \delta(s, a) \right) & (13) \\ \text{subject to} & \sum_{a \in \mathcal{A}} \pi(a|s) = 1 \\ & \pi(a|s) \geq 0, \forall a \in \mathcal{A}. \end{aligned}$$

We consider the Lagrange function

$$L(\boldsymbol{\pi}_s, \beta) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(r(a|s) - \eta \ln \pi(a|s) + \delta(s, a) \right) - \beta \left(\sum_a \pi(a|s) - 1 \right).$$

We see that if $\boldsymbol{\pi}_s^*$ is optimal to (13), then $(\partial L(\pi(\cdot|s), \beta))/\partial \pi(a|s) = 0$ at $\pi^*(a|s)$ for all $a \in \mathcal{A}$, leading to the following equations

$$\begin{cases} \eta \ln \pi^*(a|s) = r(a|s) + \delta(s, a) - (\eta + \beta) \\ \sum_a \pi^*(a|s) = 1. \end{cases} \quad (14)$$

Hence, we have

$$\begin{cases} \pi^*(a|s) = (\exp(r(a|s)/\eta + \delta(s, a)/\eta)) / \exp(1 + \beta/\eta) \\ \exp(1 + \beta/\eta) = \sum_a \exp(r(a|s)/\eta + \delta(s, a)/\eta) \\ J(s) = (1 + \beta/\eta) = \ln(\sum_a \exp(r(a|s)/\eta + \delta(s, a)/\eta)). \end{cases}$$

This leads to a closed form to compute the objective of the maximization problem. The value of V^t becomes

$$\begin{aligned} \mathcal{T}[V](s) &= \eta \ln \left(\sum_a \exp \left(r(a|s)/\eta + \frac{1}{\eta} \min_{\mathbf{q}_{sa}} \sum_{s' \in \mathcal{S}} q(s'|s, a) V(s') \right) \right) \\ &= \eta \ln \left(\sum_a e^{V(a|s)} \right), \end{aligned}$$

where

$$V(a|s) = r(a|s)/\eta + \frac{1}{\eta} \min_{\mathbf{q}_{sa}} \sum_{s' \in \mathcal{S}} q(s'|s, a) V(s').$$

The optimal policies $\pi^*(a|s)$ then becomes $\exp(V(a|s))/\exp(\mathcal{T}[V](s)/\eta)$, according to (14). We obtain the desired equations for both $\mathcal{T}[V]$ and an optimal policy $\boldsymbol{\pi}^*$.

A.2 Proof of Theorem 3.2

Contraction property. We also consider two cases. If $\mathcal{T}[V](s) \geq \mathcal{T}[V'](s)$. For any $\epsilon > 0$, let $\boldsymbol{\pi}^* \in \Delta^\pi$ be a solution such that

$$\mathcal{T}[V](s) \leq \min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \psi_s(\boldsymbol{\pi}^*, \mathbf{q}, V) \right\} + \epsilon.$$

We have

$$0 \leq \mathcal{T}[V](s) - \mathcal{T}[V'](s) \leq \min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \psi_s(\boldsymbol{\pi}^*, \mathbf{q}, V) \right\} + \epsilon - \min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \psi_s(\boldsymbol{\pi}^*, \mathbf{q}, V') \right\}. \quad (15)$$

So if we denote by $\mathbf{q}_s^* \in \mathcal{Q}_s$ a solution such that

$$\min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \psi_s(\boldsymbol{\pi}^*, \mathbf{q}, V') \right\} \geq \psi_s(\boldsymbol{\pi}^*, \mathbf{q}^*, V') - \epsilon,$$

then from (15) we have

$$|\mathcal{T}[V](s) - \mathcal{T}[V'](s)| \leq \psi_s(\boldsymbol{\pi}^*, \mathbf{q}^*, V) - \psi_s(\boldsymbol{\pi}^*, \mathbf{q}^*, V') + 2\epsilon \leq \gamma \|V - V'\|_\infty + 2\epsilon, \forall s \in \mathcal{S}.$$

Let $\epsilon \rightarrow \infty$ we obtain $\|\mathcal{T}[V] - \mathcal{T}[V']\|_\infty \leq \|V - V'\|_\infty$. The case $\mathcal{T}[V](s) < \mathcal{T}[V'](s)$ is similarly proved.

Given the contraction property, the Markov Optimality (iv) can be validated similarly as in the (s, a) -rectangularity case.

Perfect duality. For the perfect duality property, noting that the variables $\boldsymbol{\pi}_s$ in the adversary's problem $\min_{\mathbf{q}_s} \{ \cdot \}$ cannot be eliminated as in the (s, a) -rectangularity case. However, with the assumption that the uncertainty set is *convex* and *compact*, we can make use of the *von Neumann's*

minimax theorem [5] and the fact that function $\psi_s(\boldsymbol{\pi}, \mathbf{q}, V)$ is linear in $\boldsymbol{\pi}_s$ and convex in \mathbf{q}_s , to see that

$$\max_{\boldsymbol{\pi}_s} \min_{\mathbf{q}_s} \left\{ \psi_s(\boldsymbol{\pi}, \mathbf{q}, V) \right\} = \min_{\mathbf{q}_s} \max_{\boldsymbol{\pi}_s} \left\{ \psi_s(\boldsymbol{\pi}, \mathbf{q}, V) \right\},$$

Thus the Perfect Duality holds for Bellman equation $\mathcal{T}[V]$, $\mathcal{T}[V] = \widehat{\mathcal{T}}[V]$ for all $V \in \mathbb{R}^{|S|}$. Using the contraction property, let V^* be a unique fixed point solution to the systems $\mathcal{T}[V] = V$ and $\widehat{\mathcal{T}}[V] = V$, then similarly to proof of Theorem 3.1-(v), we can show that the Perfect Duality also holds for $\max_{\boldsymbol{\pi}^0, \dots} \min_{\mathbf{q}^0, \dots} F_\infty(\boldsymbol{\Pi}, \mathbf{Q})$.

Optimal policy. The key issue/challenge when proving the formulation for the optimal policy is to show that if $(\boldsymbol{\pi}_s^*, \mathbf{q}_s^*)$ is a solution to the Bellman update $\max_{\boldsymbol{\pi}_s} \min_{\mathbf{q}_s} \left\{ \psi_s(\boldsymbol{\pi}, \mathbf{q}, V) \right\}$, then is also solution to the *min-max* counterpart. It is not always the case even if the perfect duality (or minimax equality) holds. In this proof we show that this is actually the case.

First, to simplify the notations, let

$$g(\boldsymbol{\pi}_s, \mathbf{q}_s) = \mathbb{E}_{\boldsymbol{\pi}_s} \left[r(a|s) - \ln \pi(a|s) + \gamma \mathbb{E}_{\mathbf{q}_s} [V^*(s')] \right], \quad v(a, s|\mathbf{q}_s) = \exp \left(r(a|s) + \gamma \mathbb{E}_{\mathbf{q}_s} [V^*(s')] \right).$$

We see that $(\mathbf{q}_s^*, \boldsymbol{\pi}_s^*)$ specified in Theorem 3.2 is an optimal solution to the minimax problem $\min_{\mathbf{q}_s \in \mathcal{Q}_s} \max_{\boldsymbol{\pi}_s} \{g(\boldsymbol{\pi}_s, \mathbf{q}_s)\}$. We will show that $(\mathbf{q}_s^*, \boldsymbol{\pi}_s^*)$ is also a saddle point of the minimax problem, thus also a solution to the *max-min* counterpart. From the definition of $(\mathbf{q}_s^*, \boldsymbol{\pi}_s^*)$ we have $\boldsymbol{\pi}_s^* = \operatorname{argmax}_{\boldsymbol{\pi}_s} g(\boldsymbol{\pi}_s, \mathbf{q}_s^*)$. Now we prove that $\mathbf{q}_s^* = \operatorname{argmin}_{\mathbf{q}_s} g(\boldsymbol{\pi}_s^*, \mathbf{q}_s)$. From the definition of the optimal policy $\boldsymbol{\pi}_s^*$ we write

$$g(\boldsymbol{\pi}_s^*, \mathbf{q}_s) = \sum_a \frac{v(a, s|\mathbf{q}_s^*) (r(a|s) - \ln \pi^*(a|s))}{\sum_{a'} v(a', s|\mathbf{q}_s^*)} + \frac{\gamma \sum_a \sum_{s'} v(a, s|\mathbf{q}_s^*) q(s'|a, s) V(s')}{\sum_{a'} v(a', s|\mathbf{q}_s^*)}. \quad (16)$$

Recall that \mathbf{q}_s^* is an optimal solution to $\min_{\mathbf{q}_s} \{h(\mathbf{q}_s) = \sum_a v(a, s|\mathbf{q}_s)\}$. Now, consider any point $\mathbf{q}'_s \in \mathcal{Q}_s$ and denote $\boldsymbol{\delta}_s = \mathbf{q}'_s - \mathbf{q}_s^*$. The convexity of \mathcal{Q}_s implies that $\mathbf{q}_s^* + \alpha \boldsymbol{\delta}_s \in \mathcal{Q}_s$, for any $\alpha \in [0, 1]$, and there exists $\beta \in [0, 1]$ such that

$$h(\mathbf{q}_s^* + \alpha \boldsymbol{\delta}_s) - h(\mathbf{q}_s^*) = \nabla_{\mathbf{q}_s} h(\mathbf{q}_s^* + \alpha \beta \boldsymbol{\delta})^\top (\alpha \boldsymbol{\delta}_s),$$

where the equality is due to the fact that $h(\mathbf{q}_s)$ is differentiable and convex. This is also equivalent to

$$\nabla_{\mathbf{q}_s} h(\mathbf{q}_s^* + \alpha \beta \boldsymbol{\delta})^\top \boldsymbol{\delta}_s = \frac{h(\mathbf{q}_s^* + \alpha \boldsymbol{\delta}_s) - h(\mathbf{q}_s^*)}{\alpha}. \quad (17)$$

Let $\alpha \rightarrow 0$, the left side of (17) converges to $\nabla_{\mathbf{q}_s} h(\mathbf{q}_s^*)^\top \boldsymbol{\delta}_s$ while the right side is always non-negative. As a result, we need to have $\nabla_{\mathbf{q}_s} h(\mathbf{q}_s^*)^\top \boldsymbol{\delta}_s \geq 0$. To show this more precisely, assume that $\nabla_{\mathbf{q}_s} h(\mathbf{q}_s^*)^\top \boldsymbol{\delta}_s < 0$, then the continuity of the left side of (17) implies that there exists α small enough such that $\nabla_{\mathbf{q}_s} h(\mathbf{q}_s^* + \alpha \beta \boldsymbol{\delta})^\top \boldsymbol{\delta}_s < 0$, which means $h(\mathbf{q}_s^* + \alpha \boldsymbol{\delta}_s) < h(\mathbf{q}_s^*)$, which is contrary to the definition of \mathbf{q}_s^* . So, we have $\nabla_{\mathbf{q}_s} h(\mathbf{q}_s^*)^\top \boldsymbol{\delta}_s \geq 0$ or $\nabla_{\mathbf{q}_s} h(\mathbf{q}_s^*)^\top \mathbf{q}'_s \geq \nabla_{\mathbf{q}_s} h(\mathbf{q}_s^*)^\top \mathbf{q}_s^*$. Since we can choose \mathbf{q}'_s arbitrarily in \mathcal{Q}_s , we have $\mathbf{q}_s^* = \operatorname{argmin}_{\mathbf{q}_s} \nabla_{\mathbf{q}_s} h(\mathbf{q}_s^*)^\top \mathbf{q}_s$. Moreover,

$$\nabla_{\mathbf{q}_s} h(\mathbf{q}_s^*)^\top \mathbf{q}_s = \gamma \sum_a v(a, s|\mathbf{q}_s^*) \sum_{s'} V(s') q(s'|a, s). \quad (18)$$

Combine (16) and (18) and the recent claim that $\mathbf{q}_s^* = \operatorname{argmin}_{\mathbf{q}_s} \nabla_{\mathbf{q}_s} h(\mathbf{q}_s^*)^\top \mathbf{q}_s$, we have $\mathbf{q}_s^* = \operatorname{argmin}_{\mathbf{q}_s} g(\boldsymbol{\pi}_s^*, \mathbf{q}_s)$. As such, $(\boldsymbol{\pi}_s^*, \mathbf{q}_s^*)$ is also a saddle point of the minimax problem $\min_{\mathbf{q}_s \in \mathcal{Q}_s} \max_{\boldsymbol{\pi}_s} \{g(\boldsymbol{\pi}_s, \mathbf{q}_s)\}$. We need one more step to prove that the policies determined in the theorem is optimal to the max-min problem $\max_{\boldsymbol{\pi}_s} \min_{\mathbf{q}_s \in \mathcal{Q}_s} \{g(\boldsymbol{\pi}_s, \mathbf{q}_s)\}$. Using the property of the saddle point, for any policies $\boldsymbol{\pi}_s$ we have

$$\begin{aligned} \min_{\mathbf{q}_s \in \mathcal{Q}_s} \{g(\boldsymbol{\pi}_s, \mathbf{q}_s)\} &\leq g(\boldsymbol{\pi}_s, \mathbf{q}_s^*) \leq \max_{\boldsymbol{\pi}_s} \{g(\boldsymbol{\pi}_s, \mathbf{q}_s^*)\} \\ &= g(\boldsymbol{\pi}_s^*, \mathbf{q}_s^*) = \min_{\mathbf{q}_s} \{g(\boldsymbol{\pi}_s^*, \mathbf{q}_s)\}, \end{aligned}$$

which finally implies that $(\boldsymbol{\pi}_s^*, \mathbf{q}_s^*)$ determined in the theorem is also optimal to the original *max-min* problem. we complete the proof.

A.3 Proof of Theorem 3.3

We first prove the following result. Let define a function $f : \mathbb{R}^I \rightarrow \mathbb{R}$ as $f(\mathbf{x}) = \ln \left(\sum_{i=1}^I e^{x_i} \right)$. Given any vector $\mathbf{x}, \mathbf{y} \in \mathbb{R}^I$, the mean value theorem implies that there is $\mathbf{z} = \alpha \mathbf{x} + (1 - \alpha) \mathbf{y}$, $\alpha \in [0, 1]$, such that

$$|f(\mathbf{x}) - f(\mathbf{y})| = |\nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{y})| \leq \sum_{i \in [I]} \|\mathbf{x} - \mathbf{y}\|_\infty \frac{e^{z_i}}{\sum_{i' \in [I]} e^{z_{i'}}} = \|\mathbf{x} - \mathbf{y}\|_\infty. \quad (19)$$

For (i), we first have, for any $s \in \mathcal{S}$, we have

$$\mathcal{T}[V](s) \leq \tilde{\mathcal{T}}[V](s) \leq \eta \ln \left(\sum_{a \in \mathcal{A}} \exp \left(r(a|s)/\eta + \gamma \min_{\mathbf{q}_{sa}} \left\{ \sum_{s' \in \mathcal{S}} q(s'|a, s) V(s') \right\} / \eta + \gamma \xi / \eta \right) \right).$$

So, the inequality in (19) tells us that

$$\begin{aligned} |\mathcal{T}[V](s) - \tilde{\mathcal{T}}[V](s)| &\leq \left| \mathcal{T}[V](s) - \eta \ln \left(\sum_{a \in \mathcal{A}} \exp \left(r(a|s)/\eta + \gamma \min_{\mathbf{q}_{sa}} \left\{ \sum_{s' \in \mathcal{S}} q(s'|a, s) V(s') \right\} / \eta + \gamma \xi / \eta \right) \right) \right| \\ &\leq \gamma \xi, \end{aligned}$$

which means

$$\|\mathcal{T}[V] - \tilde{\mathcal{T}}[V]\|_\infty \leq \gamma \xi. \quad (20)$$

Moreover, using the triangle inequality, We can further write

$$\begin{aligned} \|\mathcal{T}^n[V] - \tilde{\mathcal{T}}^n[V]\|_\infty &\leq \|\tilde{\mathcal{T}}^n[V] - \mathcal{T}[\tilde{\mathcal{T}}^{n-1}[V]]\|_\infty + \|\mathcal{T}[\tilde{\mathcal{T}}^{n-1}[V]] - \mathcal{T}^n[V]\|_\infty \\ &\stackrel{(*)}{\leq} \gamma \xi + \gamma \|\mathcal{T}^{n-1}[V] - \tilde{\mathcal{T}}^{n-1}[V]\|_\infty \\ &\leq \dots \\ &\leq \gamma \xi (1 + \dots + \gamma^{[n-1]}) \\ &= \gamma \xi (1 - \gamma^{[n]}) / (1 - \gamma), \end{aligned}$$

where (*) is due to (20). This is the desired bound.

For (ii), we need the following chain of claims.

- **Claim 1:** For any $V \in \mathbb{R}^{|\mathcal{S}|}$

$$\|\mathcal{T}[V] - V\|_\infty \leq \|\tilde{\mathcal{T}}[V] - \mathcal{T}[V]\|_\infty + \|\tilde{\mathcal{T}}[V] - V\|_\infty \leq \gamma \xi + \|\tilde{\mathcal{T}}[V] - V\|_\infty \quad (21)$$

- **Claim 2:** For any $V \in \mathbb{R}^{|\mathcal{S}|}$

$$\begin{aligned} \|\mathcal{T}^n[V] - V\|_\infty &\leq \|\mathcal{T}^n[V] - \mathcal{T}^{n-1}[V]\|_\infty + \|\mathcal{T}^{n-1}[V] - V\|_\infty \\ &\leq \gamma^{[n-1]} \|\mathcal{T}[V] - V\|_\infty + \|\mathcal{T}^{n-1}[V] - V\|_\infty \\ &\leq \|\mathcal{T}[V] - V\|_\infty (1 + \dots + \gamma^{[n-1]}) \\ &= \|\mathcal{T}[V] - V\|_\infty \frac{1 - \gamma^{[n]}}{1 - \gamma}. \end{aligned}$$

So

$$\|V - V^*\|_\infty \leq \|\mathcal{T}[V] - V\|_\infty \frac{1}{1 - \gamma}.$$

- **Claim 3:** For any $V \in \mathbb{R}^{|\mathcal{S}|}$ and $n \in \mathbb{N}_+$

$$\begin{aligned} \|\tilde{\mathcal{T}}^n[V] - \tilde{\mathcal{T}}^{n-1}[V]\|_\infty &\leq \|\tilde{\mathcal{T}}^n[V] - \mathcal{T}[\tilde{\mathcal{T}}^{n-1}[V]]\|_\infty + \|\tilde{\mathcal{T}}^{n-1}[V] - \mathcal{T}[\tilde{\mathcal{T}}^{n-2}[V]]\|_\infty + \\ &\quad \gamma \|\tilde{\mathcal{T}}^{n-1}[V] - \tilde{\mathcal{T}}^{n-2}[V]\|_\infty \\ &\stackrel{(**)}{\leq} 2\gamma \xi + \gamma \|\tilde{\mathcal{T}}^{n-1}[V] - \tilde{\mathcal{T}}^{n-2}[V]\|_\infty, \end{aligned}$$

where (**) is due to (20). So,

$$\|\tilde{\mathcal{T}}^n[V] - \tilde{\mathcal{T}}^{n-1}[V]\|_\infty \leq 2\xi \gamma \frac{1 - \gamma^{[n-1]}}{1 - \gamma} + \gamma^{[n-1]} \|\tilde{\mathcal{T}}[V] - V\|_\infty.$$

- **Claim 4:** For any $V \in \mathbb{R}^{|\mathcal{S}|}$ and $n \in \mathbb{N}_+$

$$\begin{aligned}
\|\tilde{\mathcal{T}}^n[V] - V^*\|_\infty &\leq \frac{1}{1-\gamma} \|\mathcal{T}[\tilde{\mathcal{T}}^n[V]] - \tilde{\mathcal{T}}^n[V]\|_\infty \\
&\leq \frac{\gamma\xi}{1-\gamma} + \frac{1}{1-\gamma} \|\tilde{\mathcal{T}}^{n+1}[V] - \tilde{\mathcal{T}}^n[V]\|_\infty \\
&\stackrel{(***)}{\leq} \frac{\gamma\xi}{1-\gamma} + \frac{2\gamma\xi}{(1-\gamma)^2} + \frac{\gamma^{[n]}}{1-\gamma} \|\tilde{\mathcal{T}}[V] - V\|_\infty, \tag{22}
\end{aligned}$$

where (***) is due to **Claim 4**.

Thus, to have $\|\tilde{\mathcal{T}}^n[V] - V^*\|_\infty \leq \epsilon$, it is necessary to select $\xi \leq \epsilon(1-\gamma)^2/(4\gamma)$ and $\|\tilde{\mathcal{T}}^{n+1}[V] - \tilde{\mathcal{T}}^n[V]\|_\infty \leq 3\epsilon(1-\gamma)/4$. Note that the latter inequality always occurs if n is large enough, because

$$\begin{aligned}
\|\tilde{\mathcal{T}}^{n+1}[V] - \tilde{\mathcal{T}}^n[V]\|_\infty &\leq 2\xi\gamma \frac{1}{1-\gamma} + \gamma^{[n-1]} \|\tilde{\mathcal{T}}[V] - V\|_\infty \\
&\leq \epsilon(1-\gamma)/2 + \gamma^{[n-1]} \|\tilde{\mathcal{T}}[V] - V\|_\infty
\end{aligned}$$

and the term $\gamma^{[n-1]} \|\tilde{\mathcal{T}}[V] - V\|_\infty$ converges to zero when $n \rightarrow \infty$. Moreover, we see that it would require $n = \mathcal{O}(\ln \epsilon^{-1})$ to have $\gamma^{[n-1]} \|\tilde{\mathcal{T}}[V] - V\|_\infty \leq \epsilon(1-\gamma)/4$.

For the last claim (iii), we write the optimal policy and the approximate policy as

$$\pi^*(a|s) = \frac{\exp(z(a, s|V^*, \mathbf{q}^*))/\eta}{\sum_{a'} \exp(z(a', s|V^*, \mathbf{q}^*)/\eta)}; \text{ and } \tilde{\pi}(a|s) = \frac{\exp(z(a, s|\tilde{V}, \bar{\mathbf{q}}))/\eta}{\sum_{a'} \exp(z(a', s|\tilde{V}, \bar{\mathbf{q}})/\eta)}$$

where $z(a, s|\tilde{V}, \bar{\mathbf{q}}) = r(a|s) + \gamma \mathbb{E}_{\bar{\mathbf{q}}_{sa}}[\tilde{V}(s')]$. We see that, for any $a \in \mathcal{A}, s \in \mathcal{S}$

$$\begin{aligned}
|z(a, s|V^*, \mathbf{q}^*) - z(a, s|\tilde{V}, \bar{\mathbf{q}})| &= \gamma |\mathbb{E}_{\bar{\mathbf{q}}_{sa}}[\tilde{V}(s')] - \min_{\mathbf{q}_{sa}} \mathbb{E}_{\mathbf{q}_{sa}}[V^*(s')]| \\
&\leq \gamma |\mathbb{E}_{\bar{\mathbf{q}}_{sa}}[\tilde{V}(s')] - \min_{\mathbf{q}_{sa}} \mathbb{E}_{\mathbf{q}_{sa}}[\tilde{V}(s')]| + \gamma |\min_{\mathbf{q}_{sa}} \mathbb{E}_{\mathbf{q}_{sa}}[\tilde{V}(s')] \\
&\quad - \min_{\mathbf{q}_{sa}} \mathbb{E}_{\mathbf{q}_{sa}}[V^*(s')]| \\
&\stackrel{(i)}{\leq} \gamma\xi + \gamma |\min_{\mathbf{q}_{sa}} \mathbb{E}_{\mathbf{q}_{sa}}[\tilde{V}(s')] - \min_{\mathbf{q}_{sa}} \mathbb{E}_{\mathbf{q}_{sa}}[V^*(s')]| \tag{23}
\end{aligned}$$

We now consider two cases

- If $\min_{\mathbf{q}_{sa}} \{\mathbb{E}_{\mathbf{q}_{sa}}[\tilde{V}(s')]\} \geq \min_{\mathbf{q}_{sa}} \{\mathbb{E}_{\mathbf{q}_{sa}}[V^*(s')]\}$, then let \mathbf{q}_{sa}^* be a solution attaining the optimal value $\min_{\mathbf{q}_{sa}} \{\mathbb{E}_{\mathbf{q}_{sa}}[V^*(s')]\}$. We have

$$\begin{aligned}
\min_{\mathbf{q}_{sa}} \mathbb{E}_{\mathbf{q}_{sa}}[\tilde{V}(s')] - \min_{\mathbf{q}_{sa}} \mathbb{E}_{\mathbf{q}_{sa}}[V^*(s')] &\leq |\mathbb{E}_{\mathbf{q}_{sa}^*}[\tilde{V}(s') - V^*(s')]| \\
&\leq \|\tilde{V} - V^*\|_\infty \tag{24}
\end{aligned}$$

- If $\min_{\mathbf{q}_{sa}} \{\mathbb{E}_{\mathbf{q}_{sa}}[\tilde{V}(s')]\} < \min_{\mathbf{q}_{sa}} \{\mathbb{E}_{\mathbf{q}_{sa}}[V^*(s')]\}$, then similarly we let \mathbf{q}_{sa}^* be a solution attaining the optimal value $\min_{\mathbf{q}_{sa}} \{\mathbb{E}_{\mathbf{q}_{sa}}[\tilde{V}(s')]\}$ and obtain

$$\min_{\mathbf{q}_{sa}} \mathbb{E}_{\mathbf{q}_{sa}}[\tilde{V}(s')] - \min_{\mathbf{q}_{sa}} \mathbb{E}_{\mathbf{q}_{sa}}[V^*(s')] \leq \|\tilde{V} - V^*\|_\infty \tag{25}$$

Combine (23), (24) and (25) we have

$$|z(a, s|V^*, \mathbf{q}^*) - z(a, s|\tilde{V}, \bar{\mathbf{q}})| \leq \gamma(\xi + \epsilon). \tag{26}$$

We now look at the difference between $\tilde{\pi}(a|s)$ and $\pi^*(a|s)$ as

$$\begin{aligned}
& \left| \ln \frac{\tilde{\pi}(a|s)}{\pi^*(a|s)} \right| \\
& \leq \frac{1}{\eta} |z(a, s|V^*, \mathbf{q}^*) - z(a, s|\tilde{V}, \bar{\mathbf{q}})| + \\
& \quad \left| \ln \left(\sum_{a'} \exp(z(a, s|\tilde{V}, \bar{\mathbf{q}})/\eta) \right) - \ln \left(\sum_{a'} \exp(z(a, s|V^*, \mathbf{q}^*)/\eta) \right) \right| \\
& \stackrel{(i)}{\leq} \frac{1}{\eta} |z(a, s|V^*, \mathbf{q}^*) - z(a, s|\tilde{V}, \bar{\mathbf{q}})| + \frac{1}{\eta} \max_{a'} |z(a', s|V^*, \mathbf{q}^*) - z(a', s|\tilde{V}, \bar{\mathbf{q}})| \\
& \stackrel{(ii)}{\leq} \frac{2}{\eta} (\xi + \epsilon), \tag{27}
\end{aligned}$$

where (i) is due to (19) and (ii) is due to (26). Continue to elaborate (27) we get

$$\frac{|\tilde{\pi}(a|s) - \pi^*(a|s)|}{\pi^*(a|s)} \leq \exp(2(\xi + \epsilon)/\eta) - 1, \tag{28}$$

thus $|\tilde{\pi}(a|s) - \pi^*(a|s)| \leq \exp(2(\xi + \epsilon)/\eta) - 1$, which leads to the desired bound.

B Relevant Algorithms and Discussions

B.1 Hardness of Solving Robust ER-MDP with Non-rectangular Uncertainty Sets

Theorems 3.1 and 3.2 imply that if we can efficiently (i.e., in polynomial time) solve the inner minimization problems $\min_{\mathbf{q}_{s,a}} \mathbb{E}[V(s')]$ in the (s, a) -rectangularity case and $\min_{\mathbf{q}_s \in \mathcal{Q}_s} \{ \sum_{a \in \mathcal{A}} \exp(z(a, s|V^*, \mathbf{q})) \}$ in the (s, a) -rectangularity case, then we can compute the value function as well as the optimal policy in polynomial time as well. We will discuss this in the next question. A relevant question here is that what happens if the uncertainty set is not rectangular. [35] consider the standard robust MDP and show that, if this is the case, then unless $\mathcal{P} = \mathcal{NP}$, it is generally not possible to achieve an ϵ -approximation of the expected accumulated reward in polynomial time. We can show that this result also holds for the robust entropy-regularized MDP model. Our argument is that, for any $\eta > 0$, if there is an algorithm \mathbf{X} that is able to give a ϵ -approximation of $\max_{\boldsymbol{\pi}} \min_{\mathbf{q}} F_{\infty}^{\eta}(\mathbf{r}, \boldsymbol{\pi}, \mathbf{q})$, for any $\epsilon > 0$, where $F_{\infty}^{\eta}(\boldsymbol{\pi}, \mathbf{q})$ is the expected accumulated regularized reward as in (1) but we add η and the reward function \mathbf{r} as parameters to facilitate our arguments. Now, for any $N > 0$ we can solve the regularized problem (in polynomial time) by \mathbf{X} with rewards $\mathbf{r}' = \mathbf{r} \times N$ and approximation error ϵ , then we see that the algorithm will give an $(\epsilon/2)$ -approximation of $(N \max_{\boldsymbol{\pi}} \min_{\mathbf{q}} F_{\infty}^{\eta/N}(\mathbf{r}, \boldsymbol{\pi}, \mathbf{q}))$, or a $(\epsilon/2)$ -approximation of $(\max_{\boldsymbol{\pi}} \min_{\mathbf{q}} F_{\infty}^{\eta/N}(\mathbf{r}, \boldsymbol{\pi}, \mathbf{q}))$. Furthermore, by choosing N large enough, we also have $|\max_{\boldsymbol{\pi}} \min_{\mathbf{q}} F_{\infty}^{\eta/N}(\mathbf{r}, \boldsymbol{\pi}, \mathbf{q}) - \max_{\boldsymbol{\pi}} \min_{\mathbf{q}} F_{\infty}^0(\mathbf{r}, \boldsymbol{\pi}, \mathbf{q})| < \epsilon/2$, noting that $F_{\infty}^0(\mathbf{r}, \boldsymbol{\pi}, \mathbf{q})$ is the objective in the unregularized case. By a triangle inequality, we see that $|\tilde{F} - \max_{\boldsymbol{\pi}} \min_{\mathbf{q}} F_{\infty}^0(\mathbf{r}, \boldsymbol{\pi}, \mathbf{q})| < \epsilon$, which means that Algorithm \mathbf{X} can give a ϵ -approximation of $\max_{\boldsymbol{\pi}} \min_{\mathbf{q}} F_{\infty}^0(\mathbf{r}, \boldsymbol{\pi}, \mathbf{q})$, which contradicts [35]'s claims.

B.2 Approximate Robust Value Iteration

Algorithm 1: Robust value iteration

```

# Compute an  $\epsilon$ -approximation of  $V^*$ 
 $V = V^0 = \mathbf{0}, \bar{V} = \mathbf{1}$ 
repeat
   $\bar{V} = V$ ;
  Solve the inner minimization problem  $\min_{\mathbf{q}_{as}} \mathbb{E}_{\mathbf{q}_{sa}}[V(s)]$  by a  $\xi$ -approximation algorithm,
  where  $\xi = \epsilon(1 - \gamma)^2/(4\gamma)$ . Then, update  $V \leftarrow \tilde{\mathcal{T}}[V]$ .
until  $\|V - \bar{V}\|_\infty \leq 3\epsilon(1 - \gamma)/4$ .
# Compute an  $\epsilon$ -approximation of  $\pi^*$ 
 $V = V^0 = \mathbf{0}, \bar{V} = \mathbf{1}$ 
repeat
   $\bar{V} = V$ ;
  Solve the inner minimization problem  $\min_{\mathbf{q}_{as}} \mathbb{E}_{\mathbf{q}_{sa}}[V(s)]$  by a  $\xi$ -approximation algorithm,
  where  $\xi = \ln(\epsilon + 1)(1 - \gamma)^2/(8\gamma)$ . Then, update  $V \leftarrow \tilde{\mathcal{T}}[V]$ .
until  $\|V - \bar{V}\|_\infty \leq 3 \ln(\epsilon + 1)(1 - \gamma)/8$ .

```

B.3 Robust IRL

In perspective of imitation learning/IRL, we are interested in approximating the log-likelihood function. Assume that the demonstrated data consists of I trajectories and i -th trajectory contains K_i state-action observations. The average log-likelihood function is defined as $\mathcal{L}(\Omega|\theta) = \frac{1}{I} \sum_{i=1}^I \mathcal{L}(\omega_i|\theta)$, where θ is a vector of parameters to be inferred from the data, and $\mathcal{L}(\omega_i|\theta)$ is the log-likelihood value of sequence ω_i , $i = 1, \dots, I$, defined as $\mathcal{L}(\omega_i|\theta) = \sum_{t=0}^{K_i-1} \ln \pi^*(a_t^i|s_t^i)$. The following algorithm describe a robust IRL algorithm that allows to learn from *conservative* behavior.

Algorithm 2: Robust infinite-horizon IRL

```

# Compute an  $\epsilon$ -approximation of  $\mathcal{L}(\Omega|\theta)$ 
 $V = V^0 = \mathbf{0}, \bar{V} = \mathbf{1}$ 
for each sequence  $\omega_i, i \in [I]$  do
  repeat
     $\bar{V} = V$ ;
    Solve  $\min_{\mathbf{q}_{as}} \mathbb{E}_{\mathbf{q}_{sa}}[V(s)]$  by a  $\xi$ -approximation algorithm. where  $\xi = \frac{\epsilon(1-\gamma)^2}{8\gamma^2 \max_i \{K_i\}}$ .
    Update  $V \leftarrow \tilde{\mathcal{T}}[V]$ .
  until  $\|V - \bar{V}\|_\infty \leq 3\epsilon(1 - \gamma)/(8\gamma \max_i \{K_i\})$ 
  Compute  $P(a_k^i|s_k^i, \theta), k = 0, \dots, K_i$ , based on fixed point solution  $V$ , and
   $P(\omega_i|\theta) = \prod_{k=0}^{K_i} P(a_k^i|s_k^i, \theta)$ 
end for
Return  $\tilde{\mathcal{L}}(\Omega|\theta) = 1/I \sum_{i \in [I]} \ln P(\omega_i|\theta)$ 

```

Similarly to the finite case, we can show that if we can compute an ϵ^V -approximation of the fixed point solution V^* , then we can achieve a $(2\gamma\epsilon^V \max_i \{K_i\})$ -approximation of $\mathcal{L}(\omega_i|\theta)$ and $\mathcal{L}(\Omega|\theta)$. Algorithm 2 presents main steps to compute an ϵ -approximation of the log-likelihood function. The computational complexity in the case of single KL divergence bound is $\mathcal{O}(I|\mathcal{S}||\mathcal{A}| \max_s \{N_s\} (\ln \epsilon^{-1})^2)$ and in the case of several bounds with interior-point algorithms, the worst-case complexity is $\mathcal{O}(I|\mathcal{S}||\mathcal{A}| (\max_s \{N_s\})^{7/2} (\ln \epsilon^{-1})^2)$. On the other hand, when the transition probabilities are assumed to be known with certainty, this worst-case complexity becomes $\mathcal{O}(I|\mathcal{S}||\mathcal{A}| (\max_s \{N_s\}) \ln \epsilon^{-1})$.

B.4 Prediction Log-loss Guarantee in Robust ER-MDP

It is also interesting to look at how our robust ER-MDP model is connected to the standard maximum causal entropy principle [38]. Proposition B.1 below shows that, in analogy to [38], the prediction log-loss guarantee holds for the robust ER-MDP model, but with an additional level of robustness

w.r.t uncertain dynamics \mathbf{Q} . This result is also relevant to a claim in [6] saying that the standard ER-MDP is robust for a certain class of reward functions. This implies that our robust ER-MDP model adds another level of robustness to their setting when the dynamics are ambiguous.

Proposition B.1 *Assume that \mathbf{Q} is (a, s) -rectangular and compact, or (a, s) -rectangular, compact and convex, let $(\boldsymbol{\pi}^*, \mathbf{q}^*)$ be the solution determined in Theorems 3.1 or 3.2, then $\boldsymbol{\rho}^t = \boldsymbol{\pi}^*$ and $\mathbf{q}^t = \mathbf{q}^*$, for $t = 0, \dots, \infty$, minimize the prediction log-loss*

$$\min_{\substack{\boldsymbol{\rho}^t \in \Delta^\pi \\ \mathbf{q}^t \in \mathcal{Q} \\ t=0, \dots}} \max_{\substack{\boldsymbol{\pi}^t \in \Delta^\pi, t=0, \dots \\ \mathbb{E}_{\tau \sim (\boldsymbol{\Pi}, \boldsymbol{\rho})} [R(\tau)] = \tilde{\mathbb{E}}^R}} \mathbb{E}_{\boldsymbol{\Pi}, \boldsymbol{\rho}} \left[\sum_{t=0}^{\infty} -\gamma^{[t]} \ln \rho^t(a_t | s_t) \right],$$

where $R(\tau)$ is the accumulated and discounted reward of trajectory $\tau = \{(s_0, a_0), (s_1, a_1), \dots\}$ and $\tilde{\mathbb{E}}^R$ is empirical expectation of the accumulated rewards.

Proof.

Under the assumptions, we see that $(\mathbf{q}^t, \boldsymbol{\pi}^t) = (\mathbf{q}^*, \boldsymbol{\pi}^*)$, $t = 0, \dots$, is a solution to the problem

$$\min_{\mathbf{q}^0, \dots} \max_{\boldsymbol{\pi}^0, \dots} \left\{ \mathbb{E}_{\mathbf{Q}, \boldsymbol{\Pi}} \left[\sum_{t=0}^{\infty} \gamma^{[t]} \left(r(a_t | s_t) - \eta \ln \pi^t(a_t | s_t) \right) \right] \right\}. \quad (29)$$

It is also well-known that the inner maximization optimization problem of (29) can be formulated equivalently as a maximum causal entropy problem [38]

$$\begin{aligned} \sup_{\boldsymbol{\pi}^0, \dots} \quad & \mathbb{E}_{\boldsymbol{\pi}^0, \dots} \left[\sum_{t=0}^{\infty} -\gamma^{[t]} \eta \ln \pi^t(a_t | s_t) \right] \\ \text{subject to} \quad & \mathbb{E}_{\tau \sim (\boldsymbol{\Pi}, \mathbf{Q})} [R(\tau)] = \tilde{\mathbb{E}}^R. \\ & \boldsymbol{\pi}^t \in \Delta^\pi, t = 0, \dots \end{aligned} \quad (30)$$

The prediction log-loss guarantee shown in [38] also implies that if $\boldsymbol{\pi}^*$ is an optimal solution to (30), then it is also a solution to the problem

$$\begin{aligned} \min_{\substack{\boldsymbol{\rho}^t \in \Delta^\pi \\ t=0, \dots}} \max_{\substack{\boldsymbol{\pi}^t \in \Delta^\pi \\ t=0, \dots}} \quad & \mathbb{E}_{\boldsymbol{\Pi}, \mathbf{Q}} \left[\sum_{t=0}^T -\eta \ln \rho(a_t | s_t) \right] \\ \text{subject to} \quad & \mathbb{E}_{\tau \sim (\boldsymbol{\Pi}, \mathbf{Q})} [R(\tau)] = \tilde{\mathbb{E}}^R. \end{aligned} \quad (31)$$

Combine (29), (30) and (31) we obtain the desired result. ■

B.5 Robust General-Regularized MDP

We show how our results can be extended to the general regularized MDP framework introduced in [9]. In a regularized model, a regularized term $\phi_s(\boldsymbol{\pi}_s)$ are added to the reward [9], for any $s \in \mathcal{S}$. The Markov decision problem in the finite-horizon case can be stated as

$$\max_{\substack{\boldsymbol{\pi}^t \in \Delta^\pi \\ t=0, \dots}} \min_{\substack{\mathbf{q}^t \in \mathcal{Q} \\ t=0, \dots}} \left\{ \mathbb{E}_{\tau \sim (\boldsymbol{\Pi}, \mathbf{Q})} \left[\sum_{t=0}^{\infty} \gamma^{[t]} \left(r(a_t | s_t) + \phi_{s_t}(\boldsymbol{\pi}_{s_t}^t) \right) \right] \right\}. \quad (32)$$

It is typically assumed that $\phi_s(\boldsymbol{\pi}_s)$ is concave and bounded. If $\phi_s(\boldsymbol{\pi}_s) = -\sum_{a \in \mathcal{A}} \pi(a|s) \ln \pi(a|s)$ (negative relative entropy), the model becomes the ER-MDP model studied above. In analogy to the ER-MDP, we define the mapping $\mathcal{T}^\phi[V] : \mathbb{R}^{|\mathcal{S}|} \rightarrow \mathbb{R}^{|\mathcal{S}|}$

$$\mathcal{T}^\phi[V](s) = \max_{\boldsymbol{\pi}_s \in \Delta_s^\pi} \min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \mathbb{E}_{\boldsymbol{\pi}_s, \mathbf{q}_s} \left[r(a|s) + \gamma \sum_{s'} q(s'|s, a) V(s') \right] + \phi_s(\boldsymbol{\pi}_s) \right\}.$$

Then the contraction mapping can be verified analogously as in the entropy-regularized models. That is, under both (s, a) and (s) -rectangularity conditions, i.e., for any $V, V' \in \mathbb{R}^{|\mathcal{S}|}$, we have $\mathcal{T}^\phi[V]$ is

a contraction mapping with parameter $\gamma \|\mathcal{T}^\phi[V] - \mathcal{T}^\phi[V']\|_\infty \leq \gamma \|V - V'\|_\infty$, with a note that the contraction property for the non-robust model has been shown in [9]. The other basic properties, except the perfect duality can be proved similarly as well. For example, we also have the Markov optimality saying that, under the two uncertainty assumptions, if $V^{\phi,*}$ is a solution to the equation $\mathcal{T}^\phi[V] = V$, then

$$V^{\phi,*}(s) = \max_{\substack{\boldsymbol{\pi}^t \in \Delta^\pi \\ t=0, \dots}} \min_{\substack{\mathbf{q}^t \in \mathcal{Q} \\ t=0, \dots}} \left\{ \mathbb{E}_{\tau \sim (\boldsymbol{\pi}, \mathbf{Q})} \left[\sum_{t=0}^{\infty} \gamma^t \left(r(a_t | s_t) + \phi_{s_t}(\boldsymbol{\pi}_{s_t}^t) \right) \mid s_0 = s \right] \right\},$$

leading to the result that one can solve the Bellman equation $\mathcal{T}^\phi[V] = V$ to get an optimal policy, as

$$\boldsymbol{\pi}_s^* = \operatorname{argmax}_{\boldsymbol{\pi}_s} \{ \mathbf{w}_s^T \boldsymbol{\pi}_s + \phi_s(\boldsymbol{\pi}_s) \}$$

where $\mathbf{w}_s \in \mathbb{R}^{|\mathcal{A}|}$ with entries

$$w_{sa} = r(a|s) + \gamma \min_{\mathbf{q}_{s'} \in \mathcal{Q}_{s'}} \left\{ \sum_{s'} q(s'|s, a) V^{\phi,*}(s') \right\}.$$

Note that if the convex conjugate function (i.e. Legendre-Fenchel transform) of $-\phi_s(\boldsymbol{\pi}_s)$ can be computed efficiently, then the contraction mapping $\mathcal{T}^\phi[V]$ can be expressed as $\mathcal{T}^\phi[V] = \phi_s^*(\mathbf{w}_s)$, where $\phi_s^*(\mathbf{w}_s)$ is the convex conjugate of $-\phi_s(\boldsymbol{\pi}_s)$ in Δ_s^π , and $\mathbf{w}_s \in \mathbb{R}^{|\mathcal{A}|}$ with entries $w_{sa} = r(a|s) + \gamma \min_{\mathbf{q}_{s'} \in \mathcal{Q}_{s'}} \{ \sum_{s'} q(s'|s, a) V(s') \}$.

If the uncertainty set \mathcal{Q} is only (s) -rectangular, the inner infimum problem in the mapping $\mathcal{T}^\phi[V]$ involves $\boldsymbol{\pi}_s$ as decision variables. Thus, solving the robust Bellman equation is more difficult. However, these *max-min* problems can be solved efficiently by a saddle point algorithm, e.g., the Frank-Wolfe algorithms proposed in [10]. In this context, the computational complexity is more difficult to bound, as compared to what we have in Section 5.

We also can show that the perfect duality also holds in the context, for any uncertainty set if \mathcal{Q} is (s, a) -rectangular and for convex and compact if \mathcal{Q} is (s) -rectangular. The proof can be done by showing that the perfect duality holds for the robust Bellman equation using the *von Neumann's minimax* theorem, analogously to the entropy-regularized case. The perfect duality property would be helpful for solving the robust Bellman equation in the (s) -rectangularity case. More precisely, in case that the convex conjugate of $-\phi_s(\boldsymbol{\pi}_s)$ can be computed conveniently (e.g., by an analytical form), one can solve the min-max counterpart of $\mathcal{T}^\phi[V]$ as

$$\tilde{\mathcal{T}}^\phi[V](s) = \min_{\mathbf{q}_s \in \mathcal{Q}_s} \max_{\boldsymbol{\pi}_s \in \Delta_s^\pi} \{ \boldsymbol{\pi}_s^T \mathbf{w}_s(\mathbf{q}_s) + \phi_s(\boldsymbol{\pi}_s) \} = \min_{\mathbf{q}_s} \phi_s^*(\mathbf{w}_s(\mathbf{q}_s)),$$

where $\mathbf{w}_s(\mathbf{q}_s) \in \mathbb{R}^{|\mathcal{A}|}$ with entries $w_{sa}(\mathbf{q}_s) = r(a|s) + \gamma \sum_{s'} q(s'|s, a) V(s')$. Since $\mathbf{w}_s(\mathbf{q}_s)$ is linear in \mathbf{q}_s , $\phi_s^*(\mathbf{w}_s(\mathbf{q}_s))$ is concave in \mathbf{q}_s , which implies that the problem $\min_{\mathbf{q}_s} \phi_s^*(\mathbf{w}_s(\mathbf{q}_s))$ can be solved efficiently in polynomial time by a convex optimization algorithm (e.g., interior-point). Recall that in the entropy-regularized model, the convex conjugate function of $\phi_s(\cdot)$ has the closed form $\phi_s^*(\mathbf{w}_s) = \ln \left(\sum_{a \in \mathcal{A}} \exp(w_{sa}) \right)$. In the general regularized case, there might be no closed form to compute $\phi_s^*(\mathbf{w}_s)$ and one might need to do it approximately. This would lead to an additional approximation error in the error propagation of the approximate value iteration or (modified) policy iteration.

B.6 Complexity Analyses for the Adversary's Problems

We analyze the computational complexity of solving the adversary's problem, under two rectangularity settings and with uncertainty sets based on several KL bounds.

B.6.1 (s, a) -rectangularity

First, for notational simplicity, we consider a compact version of the inner optimization problem $\min_{\mathbf{x}} \left\{ \sum_{i=1}^{N_s} x_i c_i \mid \mathbf{x} \in \mathcal{X} \subset \Delta(N_s) \right\}$, where $\Delta(|N_s|)$ is the simplex in \mathbb{R}^{N_s} and, N_s is the number of states that can be reached from s . Normally, $N_s \ll |\mathcal{S}|$. In a likelihood model, the uncertainty set has the form $\mathcal{X} = \{ \mathbf{x} \in \Delta(N_s) \mid \sum_i \hat{x}_i \ln x_i \geq \beta \}$, where \hat{x}_i is an empirical estimate and β is

a scalar representing an uncertainty level, such that $\sum_i \hat{x}_i \ln \hat{x}_i \geq \beta$. On the other hand, a relative entropy model takes the form $\mathcal{X} = \{\mathbf{x} \in \Delta(|N_s|) \mid \sum_i x_i \ln(x_i/\hat{x}_i) \leq \beta\}$. It is possible to show that under the above uncertainty sets, one can achieve a ξ -approximation of the inner optimal value by bisection, with complexity $C(\xi) = \mathcal{O}(N_s \ln \xi^{-1})$, for any $s \in \mathcal{S}$. We refer the reader to [27, 17] for detailed discussions.

One might be interested in a mixture of the above models, i.e., uncertainty sets determined by several KL divergence bounds. In the most general form, such an inner minimization problem can be formulated as.

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \mathbf{c}^\top \mathbf{x} = \sum_{i=1}^{N_s} c_i x_i & (33) \\ \text{subject to} \quad & \widehat{\mathbf{X}} \ln \mathbf{x} \geq \boldsymbol{\alpha} \\ & (\mathbf{x}^\top \ln \mathbf{x}) \mathbf{e} - \widehat{\mathbf{Y}} \mathbf{x} \leq \boldsymbol{\beta} \\ & \mathbf{x} \in \Delta(|N_s|), \end{aligned}$$

where $\ln \mathbf{x} = (\ln x_1, \dots, \ln x_{N_s})^\top$, $\widehat{\mathbf{X}} \in \mathbb{R}_+^{K \times N_s}$, $\boldsymbol{\alpha} \in \mathbb{R}^K$ are parameters of the likelihood models, and $\widehat{\mathbf{Y}} \in \mathbb{R}_+^{H \times N_s}$, $\boldsymbol{\beta} \in \mathbb{R}^H$ are parameters of the relative entropy models. In general, it seems not possible to solve the above problem by bisection if $K + H \geq 2$, but we can prove that (33) can be solved by interior-point in polynomial time.

Proposition B.2 *Assume that (33) satisfies the Slater condition, then a ξ -approximation of Problem 33 can be achieved with complexity $C(\xi) = \mathcal{O}((4N_s + H + K + 2)^{1/2} (4N_s + H + K) N_s^2 \ln \xi^{-1})$.*

Proof. By a change of variable, we write an equivalent problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \mathbf{c}^\top \mathbf{x} & (34) \\ \text{subject to} \quad & \sum_i \widehat{X}_{ki} z_i \geq \alpha_k & \forall k \\ & \sum_i y_i - \widehat{Y}_{hi} \leq \beta_h & \forall h \\ & z_i \leq \ln(x_i) & \forall i \\ & y_i \geq x_i \ln x_i & \forall i \\ & \sum_i x_i \geq 1 - \epsilon \\ & -\sum_i x_i \geq -(1 + \epsilon). \end{aligned}$$

With the following notes [26]

- $\Phi(x, z) = -\ln(\ln x - z) - \ln x$ is a 2-self-concordant barrier for the epigraph of $\{(x, z) \mid \ln(x) \geq z, x \geq 0\}$
- $\Gamma(x, y) = -\ln(y - x \ln x) - \ln x$ is a 2-self-concordant barrier for the epigraph of $\{(x, z) \mid x \ln(x) \leq y, x \geq 0\}$

This allows us to construct a barrier function of the feasible set of Problem 34.

$$\begin{aligned} \mathcal{F}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = & -\sum_k \ln \left(\sum_i \widehat{X}_{ki} - \alpha_k \right) - \sum_h \ln \left(-\sum_i (y_i + \widehat{Y}_{hi}) + \beta_h \right) + \sum_i \Phi(x_i, z_i) \\ & + \sum_i \Gamma(x_i, y_i) - \ln \left(\sum_i x_i + \epsilon - 1 \right) - \ln \left(-\sum_i x_i - \epsilon + 1 \right). \end{aligned}$$

We see that $\mathcal{F}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is a self-concordant [26] with variable $4N_s + K + H + 2$. The complexity of the path-following method associated with the aforementioned barrier is

$$\mathcal{O} \left((4N_s + H + K + 2)^{1/2} (4N_s + H + K) N_s^2 \ln \epsilon^{-1} \right).$$

Typically, $H, K \ll N_s$. As a result, the complexity can be bounded by $\mathcal{O}(N_s^{7/2} \ln \xi^{-1})$.

B.6.2 (s)-rectangularity

In this case, we need to solve the inner problem $\min_{\mathbf{q}_s \in \mathcal{Q}_s} \left\{ \sum_{a \in \mathcal{A}} \exp \left(r(a|s) + \gamma \mathbb{E}_{s'} [V^*(s')] \right) \right\}$, for any $s \in \mathcal{S}$, to perform contraction iterations and compute optimal policies. The inner optimization is of the form

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} && \sum_{a \in \mathcal{A}} d_a \exp \left(\sum_{i=1}^{N_s} c_{ai} x_{ai} \right) && (35) \\ & \text{subject to} && \widehat{\mathbf{X}} \ln \mathbf{x} \geq \boldsymbol{\alpha} \\ & && (\mathbf{x}^\top \ln \mathbf{x}) \mathbf{e} - \widehat{\mathbf{Y}} \mathbf{x} \leq \boldsymbol{\beta} \\ & && \mathbf{x} \in \Delta(N_s \times |\mathcal{A}|), \end{aligned}$$

where $\widehat{\mathbf{X}} \in \mathbb{R}_+^{K \times (|\mathcal{A}|N_s)}$ and $\widehat{\mathbf{Y}} \in \mathbb{R}_+^{H \times (|\mathcal{A}|N_s)}$ are parameter matrices of the likelihood models and entropy models, respectively. The proposition below shows that one can solve Problem 35 in polynomial time.

Proposition B.3 *Assume that (35) satisfies the Slater condition, then a ξ -approximation of Problem 35 can be achieved with complexity*

$$C(\xi) = \mathcal{O} \left((4|\mathcal{A}|N_s + 4|\mathcal{A}| + H + K)^{1/2} (4|\mathcal{A}|N_s + H + K) (|\mathcal{A}|N_s)^2 \ln \xi^{-1} \right).$$

Proof. By a change of variable, we write an equivalent problem

$$\begin{aligned} & \underset{\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{t}}{\text{minimize}} && \sum_{a \in \mathcal{A}} t_a && (36) \\ & \text{subject to} && \sum_{i=1}^{N_s} c_{ai} x_{ai} \leq \ln t_a && \forall a \\ & && \sum_{a,i} \widehat{X}_{kai} z_{ai} \geq \alpha_k && \forall k \\ & && \sum_{a,i} y_{ai} - \widehat{Y}_{hai} \leq \beta_h && \forall h \\ & && z_{ai} \leq \ln(x_{ai}) && \forall i \\ & && y_{ai} \geq x_{ai} \ln x_{ai} && \forall i \\ & && \sum_i x_{ai} \geq 1 - \epsilon && \forall a \\ & && - \sum_i x_{ai} \geq -(1 + \epsilon) && \forall a \end{aligned}$$

A self-concordant barrier for the feasible set of Problem 34 can be constructed as

$$\begin{aligned} \mathcal{F}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{t}) = & - \sum_k \ln \left(\sum_{a,i} \widehat{X}_{kai} z_{ai} - \alpha_k \right) - \sum_h \ln \left(- \sum_{a,i} (y_{ai} + \widehat{Y}_{hai}) + \beta_h \right) \\ & + \sum_{a,i} \Phi(x_{ai}, z_{ai}) + \sum_{a,i} \Gamma(x_{ai}, y_{ai}) + \sum_a \Phi \left(t_a, \sum_i c_{ai} x_{ai} \right) \\ & - \sum_a \ln \left(\sum_i x_i + \epsilon - 1 \right) - \sum_a \ln \left(- \sum_i x_i - \epsilon + 1 \right). \end{aligned}$$

$\mathcal{F}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{t})$ is a self-concordant [26] with variable $4|\mathcal{A}|N_s + 4|\mathcal{A}| + K + H$. The complexity of the path-following method associated with the aforementioned barrier is

$$\mathcal{O}\left((4|\mathcal{A}|N_s + 4|\mathcal{A}| + H + K)^{1/2}(4|\mathcal{A}|N_s + H + K)(|\mathcal{A}|N_s)^2 \ln \epsilon^{-1}\right).$$

■

In cases $H, K \ll |\mathcal{A}|N_s$ the complexity is about $\mathcal{O}((|\mathcal{A}|N_s)^{7/2} \ln \xi^{-1})$.