Convergence of a model-free entropy-regularized inverse reinforcement learning algorithm

Titouan Renard* **Andreas Schlaginhaufen*†** **Tingting Ni*‡** **Maryam Kamgarpour**

Abstract

Given a dataset of expert demonstrations, inverse reinforcement learning (IRL) aims to recover a reward for which the expert is optimal. This work proposes a model-free algorithm to solve the entropy-regularized IRL problem. In particular, we employ a stochastic gradient descent update for the reward and a stochastic soft policy iteration update for the policy. Assuming access to a generative model, we prove that our algorithm is guaranteed to recover a reward for which the expert is ε-optimal using an expected number of $\mathcal{O}(1/\varepsilon^2)$ samples of the Markov decision process (MDP). Furthermore, with an expected number of $\mathcal{O}(1/\varepsilon^4)$ samples we prove that the optimal policy corresponding to the recovered reward is ε-close to the expert policy in total variation distance.

1 Introduction

The problem of inverse reinforcement learning (IRL) can be informally characterized as follows. Given observations of an expert acting optimally with respect to an unknown reward in a Markov decision process (MDP), we aim to recover a reward for which the expert is optimal. While early inquiries can be traced back to optimal control theory [8] and to econometrics [16], IRL was first introduced to the machine learning community by Russell [15]. The motivation behind IRL is two-fold: First, IRL is a powerful tool for imitation learning, where the goal is to recover the expert's policy from a dataset of expert demonstrations. This is particularly useful in scenarios such as autonomous driving [13] where it is easy to collect expert demonstrations. Second, compared to other imitation learning methods that only recover a policy, IRL comes with the advantage that recovering a reward provides a more interpretable and transferable description of the task, as the reward can be potentially used to learn optimal policies in a new environment.

A powerful family of imitation learning and IRL algorithms are based on a min-max game between a policy and a reward player [19]. The policy player tries to maximize the expected reward, while the reward player tries to minimize the suboptimality of the expert demonstrations. At a saddle point, we recover a reward for which the expert is nearly optimal along with a policy that approximates the expert's. Hence, this game-theoretic approach is useful for both imitation learning, by recovering the expert's policy, and for IRL, by recovering a reward that rationalizes the expert's behavior. However, when applied to IRL, the non-uniqueness of the optimal policy leads to trivial solutions, like a uniform reward for which all policies are optimal. This degeneracy is usually addressed through an entropy regularization term, which guarantees the uniqueness of the optimal policy, leading to the widely used entropy-regularized IRL framework [23, 7].

*The first three authors contributed equally. Andreas Schlaginhaufen, Tingting Ni, and Maryam Kamgarpour are with the SYCAMORE lab at École Polytechnique Fedérale de Lausanne (EPFL), Switzerland. Correspondence to: {andreas.schlaginhaufen, tingting.ni}@epfl.ch

†Andreas Schlaginhaufen was supported by the Swiss Data Science Center.

‡Tingting Ni was supported by the Swiss National Science Foundation.

Preprint. Under review.

Although approaches based on the aforementioned min-max game have shown tremendous success in imitation learning and IRL applications [7, 4], only little is known about their convergence properties. From an optimization perspective, the min-max problem introduced by [19] is challenging for two reasons. First, since the policy player aims to maximize its expected reward in an MDP, the objective is inherently non-concave in the policy [1]. Second, as in practice we are often only given sample-based access to the MDP, either via a simulator or by collecting samples in the real world, we need to resort to stochastic optimization techniques to build a *model-free* algorithm.

1.1 Related work

Most previous work on the min-max approach to imitation learning and IRL considers the unregularized setting and linear reward classes. For the infinite-horizon unregularized case, the authors of [19] propose a multiplicative weights algorithm to update the reward and show that their algorithm provably recovers a policy that is approximately optimal under the expert's true reward. However, they require access to an approximate solution of the forward MDP problem at each update step of the reward. This results in an inner RL loop, which may lead to high per-iteration costs in practice. For infinite-horizon linear MDPs, the authors of [20] propose a single-loop proximal point algorithm based on the so-called Q-LP formulation of the forward problem [2]. They show that the policy learned by their algorithm efficiently convergences to the expert's, where the convergence is measured in a so-called integral probability metric [12] between the state-action occupancy measures.

For the finite-horizon unregularized case, the authors of [18] analyze a single-loop mirror descent-ascent algorithm and show their algorithm achieves sublinear regret measured in terms of the worst-case difference between the accumulated values of the learner and the expert. Furthermore, the authors of [10] extend this approach to linear MDPs with function approximation. It can be shown that the regret bounds established by [18] and [10] imply convergence of a randomized policy derived from the algorithm iterates in the aforementioned integral probability metric.

Note that all of the above works consider the unregularized setting and their guarantees are on the recovered policy rather than on the reward. To the best of our knowledge, [22] is the only work to provide guarantees for the infinite-horizon entropy-regularized setting. Their method uses an exact soft-policy iteration step to update the policy and stochastic gradient descent to update the reward. They show that the proposed algorithm reaches an ε-stationary solution in $\mathcal{O}(1/\varepsilon^2)$ iterations. However, they do not provide guarantees on the suboptimality of the expert with respect to the recovered reward. Moreover, they require access to the exact state-action value function and an infinitely long trajectory to estimate the reward gradient, which makes the algorithm unimplementable in a model-free setting.

1.2 Contributions

We propose a model-free single-loop entropy-regularized IRL algorithm with a stochastic projected gradient descent update for the reward and a stochastic soft policy iteration [6] update for the policy. Assuming access to a generative model, we show that our algorithm is guaranteed to recover a reward for which the expert is ε-optimal using an expected number of $\mathcal{O}(1/\varepsilon^2)$ samples from the MDP. Furthermore, we prove that with an expected number of $\mathcal{O}(1/\varepsilon^4)$ samples, the optimal policy corresponding to the recovered reward is ε-close to the expert policy in total variation distance. To the best of our knowledge, this is the first work to provide end-to-end guarantees on the rewards for a model-free single-loop entropy-regularized IRL algorithm.

2 Background

2.1 Notation

We use $\mathbb{R}$ and $\mathbb{N}$ to denote the set of real and natural numbers, respectively. For a vector x in $\mathbb{R}^d$, we denote its p-norm by $\|x\|_p$ and its projection onto a closed convex set $\mathcal{X} \subset \mathbb{R}^d$ by $\mathcal{P}_{\mathcal{X}}(x) = \arg\min_{y \in \mathcal{X}} \|x - y\|_2$. For any two vectors $x, y \in \mathbb{R}^d$, we denote the standard inner product by $\langle x, y \rangle$. Furthermore, we let $\mathcal{Y}^{\mathcal{X}}$ be the set of all functions mapping from the set $\mathcal{X}$ to $\mathcal{Y}$. Given a finite set $\mathcal{X}$ the infinity-norm of a vector valued function $f : \mathcal{X} \to \mathbb{R}^d$ over $\mathcal{X}$ is defined as

$\|f\|_\infty := \max_{x\in\mathcal{X}} \|f(x)\|_\infty$. Moreover, we denote the probability simplex over $\mathcal{X}$ by $\Delta_\mathcal{X}$. For $p \in \Delta_\mathcal{X}$, we denote the Shannon entropy of p by $\mathcal{H}(p) := -\sum_{x\in\mathcal{X}} p(x)\log p(x)$. Finally, a random variable X taking values in $\mathbb{N}$ is geometrically distributed with parameter $p \in (0,1)$, denoted as $X \sim \mathrm{Geom}(p)$, if $\Pr(X = k) = (1-p)^k p$.

2.2 Markov decision processes

Throughout this paper, we consider an entropy-regularized MDP defined by the tuple $(\mathcal{S}, \mathcal{A}, P, \nu_0, r, \gamma, \tau)$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote finite state and action spaces, $P \in \Delta_\mathcal{S}^{\mathcal{S}\times\mathcal{A}}$ is a Markovian transition kernel, and $\nu_0 \in \Delta_\mathcal{S}$ is the initial state distribution. Moreover, $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is a reward function, $\gamma \in (0,1)$ a discount factor, and $\tau > 0$ is the regularization parameter. Starting from $s_0 \sim \nu_0$, the agent chooses at each step in time, h, an action $a_h \in \mathcal{A}$, receives reward $r(s_h, a_h)$, and arrives in state $s_{h+1} \sim P(\cdot|s_h, a_h)$. Given a stationary Markov policy $\pi \in \Delta_\mathcal{A}^\mathcal{S}$, the agent selects at each state s_h its next action a_h by sampling from $\pi(\cdot|s_h)$. We use $\mathbb{P}^\pi_{\nu_0}$ to denote the distribution over the sample space $(\mathcal{S}\times\mathcal{A})^\infty = \{(s_0, a_0, s_1, a_1, \ldots) : s_h \in \mathcal{S}, a_h \in \mathcal{A},\ h \in \mathbb{N}\}$ induced by the policy π and the initial distribution ν_0. Moreover, we let $\mathbb{E}_\pi$ be the expectation with respect to $\mathbb{P}^\pi_{\nu_0}$.

The goal of the *forward* MDP problem is to find a policy $\pi \in \Delta_\mathcal{A}^\mathcal{S}$ maximizing the entropy-regularized expected discounted reward

$$J_r^\pi := \mathbb{E}_\pi\left[\sum_{h=0}^\infty \gamma^h r(s_h, a_h)\right] + \tau\mathbb{H}(\pi), \tag{O-RL}$$

where $\mathbb{H}(\pi) := \mathbb{E}_\pi\left[\sum_{h=0}^\infty \gamma^h \mathcal{H}\left(\pi(\cdot|s_h)\right)\right]$ is the discounted causal entropy of π [23]. For a fixed reward r, we denote the optimal policy as

$$\pi_r^* := \arg\max_{\pi\in\Delta_\mathcal{A}^\mathcal{S}} J_r^\pi, \tag{1}$$

and the optimal objective value as $J_r^* := J_r^{\pi_r^*}$. Note that the entropy regularization ensures that π_r^* is unique [5]. To help characterize expectations over trajectories, we introduce the state occupancy measure $\nu^\pi \in \Delta_\mathcal{S}$ defined by

$$\nu^\pi(s) := (1-\gamma)\sum_{h=0}^\infty \gamma^h \mathbb{P}^\pi_{\nu_0}(s_h = s).$$

For a function $f \in \mathbb{R}^\mathcal{S}$, this allows us to rewrite the expectation $\mathbb{E}_\pi\left[\sum_{h=0}^\infty \gamma^h f(s)\right]$ as $\mathbb{E}_{s\sim\nu^\pi}\left[f(s)\right]/(1-\gamma)$ [14].

2.3 Problem statement

The IRL problem is specified as follows: given access to a dataset of expert trajectories,

$$\mathcal{D}^E = \left\{\left(s_0^i, a_0^i, \ldots, s_{H-1}^i, a_{H-1}^i\right)\right\}_{i=1}^N,$$

sampled from an unknown expert policy π^E, we aim to recover a reward $\bar{r}$, in some reward class $\mathcal{R} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, for which the expert is optimal. However, this is problematic for two reasons. First, the expert policy π^E may not be optimal for any reward in $\mathcal{R}$. Second, we only have access to $\mathcal{D}^E$ rather than to π^E itself. To address the first issue, we relax our goal to recovering a reward minimizing the suboptimality of the expert. This leads us to the min-max formulation

$$\min_{r\in\mathcal{R}} \max_{\pi\in\Delta_\mathcal{A}^\mathcal{S}} J_r^\pi - J_r^{\pi^E}. \tag{2}$$

We can interpret (2) as a zero-sum game between a policy player that aims to maximize the MDP objective (O-RL) and a reward player that aims to minimize the suboptimality of the expert policy. If the expert is optimal for some reward in $\mathcal{R}$, then the set of minimizers in (2) coincides with the set of rewards for which the expert is optimal [17].

To address the second issue, we need to estimate $J_r^{\pi^E}$ from the expert data set $\mathcal{D}^E$. To this end, we consider the bounded linear reward class

$$\mathcal{R} = \left\{r_w := \langle w, \phi(\cdot,\cdot)\rangle : \phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^k, w \in W\right\},$$

where w is a weight vector in $W := \left\{w \in \mathbb{R}^k : \|w\|_1 \leq 1\right\}$ and $\phi(s,a)$ is a feature vector in $\mathbb{R}^k$. The features can either be task-specific, such as distance to target and speed in a driving task, or general, allowing for all state-only or state-action rewards, resulting in $k = |\mathcal{S}|$ or $k = |\mathcal{S}||\mathcal{A}|$, respectively. Introducing the feature expectation

$$\sigma^\pi := \mathbb{E}_\pi \left[\sum_{h=0}^{\infty} \gamma^h \phi(s_h, a_h)\right], \tag{3}$$

associated with the policy π, and the empirical expert feature expectation

$$\hat{\sigma}^E := \frac{1}{N} \sum_{i=1}^{N} \sum_{h=0}^{H-1} \gamma^h \phi(s_h^i, a_h^i),$$

we can rewrite $J_{r_w}^\pi = \langle w, \sigma^\pi \rangle + \tau \mathbb{H}(\pi)$ and estimate the expected reward of the expert policy as $\mathbb{E}_{\pi^E}\left[\sum_{h=0}^{\infty} \gamma^t r_w(s_h, a_h)\right] \approx \langle w, \hat{\sigma}^E \rangle$. Plugging this back into (2) leads us to the IRL problem

$$\min_{w \in W} \max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \underbrace{\langle w, \sigma^\pi - \hat{\sigma}^E \rangle + \tau \mathbb{H}(\pi) - \tau \mathbb{H}(\pi^E)}_{=:L(\pi, w)}, \tag{O-IRL}$$

where compared to (2) we replaced $\langle w, \sigma^{\pi^E} \rangle$ with its empirical estimate $\langle w, \hat{\sigma}^E \rangle$. The unknown term $\mathbb{H}(\pi^E)$ is constant in both parameters w and π and thus, is irrelevant for optimization. Note that for a fixed w, the maximizer of $L(\cdot, w)$ is the optimal policy $\pi_{r_w}^*$ defined in (1). Hence, the inner maximization in (O-IRL) is equivalent to the forward MDP problem (O-RL). Our goal is to find $\bar{w}$ minimizing (O-IRL). An approximate expert policy can then be recovered from $\bar{w}$ by solving for the optimal policy corresponding to $r_{\bar{w}}$.

3 Proposed algorithm

To solve the problem described in (O-IRL) with exact gradient information, we employ soft policy iteration for policy updates and projected gradient descent for reward updates. As detailed below, this can be viewed as a simple gradient descent-ascent type algorithm.

Policy update For a known transition law P, we can maximize the objective in (O-RL) via regularized dynamic programming [5]. In particular, for a fixed reward, we update the policy as

$$\pi_r^{t+1}(\cdot|s) \propto \exp\left(Q_r^{\pi^t}(s, \cdot)/\tau\right), \tag{4}$$

with the state-action value function

$$Q_r^{\pi^t}(s,a) := r(s,a) + \mathbb{E}_\pi \left[\sum_{h=1}^{\infty} \gamma^h \left(r(s_h, a_h) + \tau \mathcal{H}(\pi(\cdot|s_h))\right) | s_0 = s, a_0 = a\right]. \tag{5}$$

The above policy update is known as soft policy iteration, which converges linearly to the optimal policy [5]. Additionally, it can be shown that the soft policy iteration update coincides with entropy-regularized natural policy gradient with a specific stepsize [3]. Hence, the policy update (4) can be interpreted as a gradient ascent update. However, note that computing the update (4) requires access to the state-action value function $Q_r^{\pi^t}(s,a)$ for all state-action pairs (s,a).

Reward update To find the optimal reward that minimizes the objective in (O-IRL), we update the reward parameter w using a projected gradient descent update

$$w^{t+1} = \mathcal{P}_W \left(w^t - \eta_w \frac{\partial L(\pi^t, w)}{\partial w}\Big|_{w=w^t}\right) = \mathcal{P}_W \left(w^t - \eta_w \left(\sigma^{\pi^t} - \hat{\sigma}^E\right)\right), \tag{6}$$

where $\mathcal{P}_W$ represents the orthogonal projection onto W, and η_w is the reward learning rate.

3.1 Sampling scheme

Both the policy update (4) and the reward update (6) require access to the transition law P for evaluating the state-action value function (5) and the feature expectation (3), respectively. Hence, to devise a model-free algorithm, we need to estimate the state-action value and the feature expectation from samples of the MDP. To this end, we adopt the geometric sampling scheme, described below, which enables us to get unbiased estimates of the state-action value and the feature expectation.

State-action value estimation We assume access to a generative model of the MDP, allowing us to obtain multiple independent trajectories starting from any arbitrary state-action pair with any policy. Given a policy π, for each state-action pair (s, a), we construct an unbiased estimate of the state-action value $Q_r^\pi(s, a)$ by sampling B independent trajectories $\tau_i := (s, a, \{s_h^i, a_h^i\}_{h=0}^{H_i})_{i=1}^B$ from π, with horizon $H_i \sim \mathrm{Geom}(1-\gamma)$. As shown in [1, Assumption 6.3], the estimator

$$\hat{Q}_r^\pi(s, a) := r(s, a) + \frac{1}{B} \sum_{i=1}^{B} \sum_{h=0}^{H_i} \left(r(s_h^i, a_h^i) + \tau \mathcal{H}(\pi(\cdot|s_h^i)) \right).$$

is an unbiased estimator of $Q_r^\pi(s, a)$. In the following, we will denote the above sampling procedure outputting $\hat{Q}_r^\pi(s, a)$ as $\mathrm{Est}\, Q(s, a, \pi, r, B)$.

Feature expectation estimation Similarly, as for the state-action value estimate, we construct an unbiased estimate of the feature expectation σ^π, by sampling B independent trajectories $\tau_i := (\{s_h^i, a_h^i\}_{h=0}^{H_i})_{i=1}^B$ from π, with horizon $H_i \sim \mathrm{Geom}(1-\gamma)$ and initial state $s_0^i \sim \nu_0$. As shown in Lemma A.1, the estimator

$$\hat{\sigma}^\pi := \frac{1}{B} \sum_{i=1}^{B} \sum_{h=0}^{H_i} \phi(s_h^i, a_h^i),$$

is an unbiased estimator of σ^π. We denote this sampling procedure as $\mathrm{Est}\, \sigma(\pi, B)$.

3.2 Algorithm summary

Combining the above steps, we present Algorithm 1 for learning the reward in (O-IRL), which simultaneously updates the policy with stochastic soft policy iteration and the reward parameter via stochastic projected gradient descent. Note that Algorithm 1 updates the policy and the reward

Algorithm 1 Primal-dual IRL algorithm

Input: Reward learning rate $\eta_w > 0$ and batch size B.
Initialize $\pi^0 \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ and $w^0 = 0$.
Estimate $\hat{\sigma}^E = \frac{1}{N} \sum_{i=0}^{N-1} \sum_{h=0}^{H-1} \gamma^h \phi(s_h^i, a_h^i)$ from $\mathcal{D}^E$.
for $t \leftarrow 0$ **to** $T-1$ **do**
  $r^t = \langle w^t, \phi(\cdot, \cdot) \rangle$.
  `// Estimate values:`
  for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    $\hat{Q}_{r^t}^{\pi^t}(s, a) = \mathrm{Est}\, Q(s, a, \pi^t, r_t, B)$.
  $\sigma^{\pi^t} = \mathrm{Est}\, \sigma(\pi^t, B)$.
  `// Update policy and reward:`
  $\pi^{t+1}(\cdot|s) \propto \exp\left(\hat{Q}_{r^t}^{\pi^t}(s, \cdot)/\tau \right)$.
  $w^{t+1} = \mathcal{P}_W \left(w^t - \eta_w \left(\hat{\sigma}^{\pi^t} - \hat{\sigma}^E \right) \right)$.
Output: Reward $\bar{r} = r_{\bar{w}}$, with $\bar{w} = \frac{1}{T} \sum_{t=0}^{T-1} w^t$.

parameters in a single loop. That is, we do not have to solve an RL problem at every reward step, but we only employ a single approximate policy iteration update. This is in contrast to the algorithm proposed by [19]. Moreover, unlike the algorithm provided by [22], which requires the exact state-action value function (5) for the policy update and an infinitely long trajectory to estimate the feature expectation (3), Algorithm 1 is both model-free and implementable. It requires only a finite expected number of samples from the MDP per iteration.

In the next section, we show that Algorithm 1 enjoys strong guarantees for the recovered reward.

4 Convergence analysis

In Section 4.1, we prove that our algorithm is guaranteed to recover a reward for which the expert is ε-optimal using an expected number of $\mathcal{O}(1/\varepsilon^2)$ samples of the MDP, as detailed in Theorem

4.3. Then, in Section 4.2, we establish that with an expected number of $\mathcal{O}(1/\varepsilon^4)$ samples, the optimal policy corresponding to the recovered reward is ε-close to the expert policy in total variation distance, as stated in Corollary 4.6. Furthermore, we show that the total variation distance is a stronger metric for measuring policy differences compared to the metrics used in [19, 20, 18, 10].

4.1 Reward convergence

To establish guarantees for the recovered reward $\bar{r}$, we require two assumptions. First, Assumption 4.1 below ensures that the policy iterates of Algorithm 1 sufficiently explore the state space.

Assumption 4.1. *The distribution mismatch coefficient, defined by* $\vartheta := \max_{0\leq t\leq T-1} \max_{s\in\mathcal{S}} \nu^{\pi^*_{r^t}}(s)/\nu^{\pi^t}(s)$*, is bounded from above.*

Similar assumptions on the distribution mismatch coefficient are used in the prior literature [1, 11, 21]. Since $\nu^{\pi^t}(s) \geq (1-\gamma)\nu_0(s)$, Assumption 4.1 is satisfied if the initial distribution ν_0 is bounded away from zero.

Second, to show that the expert is approximately optimal for the recovered reward, we need the following approximate realizability assumption that quantifies the best-case optimality of the expert within our reward class $\mathcal{R}$.

Assumption 4.2. *There exists* $\varepsilon_{real} \geq 0$ *such that the expert policy* π^E *is* ε_{real}*-optimal for some reward* $r \in \mathcal{R}$*. That is,*

$$\min_{w\in W} \max_{\pi} L(\pi, w) = \min_{r\in\mathcal{R}} J_r^* - J_r^{\pi^E} \leq \varepsilon_{real}.$$

Assumption 4.2 has been introduced by [19]. The realizability error $\varepsilon_{\text{real}}$ can be reduced by increasing the number of features k and the diameter of W [19]. Next, we are ready to state our main convergence result.

Theorem 4.3. *Suppose Assumptions 4.1 and 4.2 hold, and let* $\eta_w = \frac{(1-\gamma)}{\sqrt{kT}\|\phi\|_\infty}$*. The expert satisfies the following optimality guarantee for the reward* $\bar{r}$ *returned by Algorithm 1:*

$$\mathbb{E}\left[J_{\bar{r}}^* - J_{\bar{r}}^{\pi^E}\right] \leq \varepsilon_{real} + \mathcal{O}\left(\gamma^H\right) + \mathcal{O}\left(1/\sqrt{T}\right).$$

Here, the expectation is taken with respect to all the randomness in Algorithm 1. Moreover, to recover a reward for which the expert is $(\varepsilon + \varepsilon_{real})$*-optimal we require the length of the expert trajectories to be* $H = \mathcal{O}(\log \varepsilon^{-1})$ *and we need an expected number of* $\mathcal{O}(1/\varepsilon^2)$ *samples from the MDP.*

The proof of Theorem 4.3 is based on two key ingredients. First, Lemma 4.4 below shows that the policy iterates π^t converge to the optimal policy for the reward iterates r^t defined in Algorithm 1.

Lemma 4.4. *Suppose Assumption 4.1 holds and let* $\eta_w = \frac{1-\gamma}{\sqrt{kT}\|\phi\|_\infty}$*. We can bound the suboptimality of the policy iterates of Algorithm 1 by*

$$\mathbb{E}\left[\max_{\pi} L(\pi, w^t) - L(\pi^t, w^t)\right] \leq \mathcal{O}\left(1/\sqrt{T}\right).$$

The above lemma shows that if we control the changes of the reward r^t by setting the reward learning rate to $\eta_w = \Theta(1/\sqrt{T})$, the value of the policy π^t converges to the optimal value under r^t at the same speed $\mathcal{O}(1/\sqrt{T})$. We provide the proof in Appendix A.

The second ingredient required for the proof of Theorem 4.3 is the following lemma that shows that our algorithm has sublinear regret with respect to the reward.

Lemma 4.5 (Stochastic online gradient descent regret)**.** *If we set the learning rate to* $\eta_w = \frac{1-\gamma}{\sqrt{kT}\|\phi\|_\infty}$*, we have*

$$\mathbb{E}\left[\sum_{t=0}^{T-1} L(\pi^t, w^t)\right] \leq \mathbb{E}\left[\min_{w\in W} \sum_{t=0}^{T-1} L(\pi^t, w)\right] + \mathcal{O}\left(\sqrt{T}\right).$$

For exact gradient information, Lemma 4.5 is a well-known result in online convex optimization [24]. Since we use an unbiased gradient estimator, this result can be easily extended to our case. We provide a proof adapted to our setting in Appendix A. Equipped with the above two lemmas, we are now ready to prove Theorem 4.3.

Proof of Theorem 4.3. We first upper bound $J_{\bar{r}}^* - J_{\bar{r}}^{\pi^E}$ as follows:

$$\begin{aligned}
&\mathbb{E}\left[J_{\bar{r}}^* - J_{\bar{r}}^{\pi^E}\right] \\
=&\mathbb{E}\left[\max_{\pi} \sum_{t=0}^{T-1} \frac{L(\pi, w^t) + \langle w^t, \sigma^{\pi^E} - \hat{\sigma}^E \rangle}{T}\right] \\
\overset{(i)}{\leq}&\mathbb{E}\left[\max_{\pi} \sum_{t=0}^{T-1} \frac{L(\pi, w^t)}{T}\right] + \mathcal{O}(\gamma^H) \\
\overset{(ii)}{\leq}&\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\max_\pi L(\pi, w^t)}{T}\right] + \mathcal{O}(\gamma^H) \\
\overset{(iii)}{\leq}&\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{L(\pi^t, w^t)}{T}\right] + \mathcal{O}\left(\gamma^H\right) + \mathcal{O}\left(1/\sqrt{T}\right) \\
\overset{(iv)}{\leq}&\mathbb{E}\left[\min_{w \in W} \frac{\sum_{t=0}^{T-1} L(\pi^t, w)}{T}\right] + \mathcal{O}\left(\gamma^H\right) + \mathcal{O}\left(1/\sqrt{T}\right).
\end{aligned} \tag{7}$$

Here, (i) follows from the truncation error of the empirical expert feature expectation, as detailed in Lemma A.1, and (ii) holds since $\max_\pi L(\pi, \cdot)$ is a pointwise maximum of affine functions and therefore convex. Moreover, in (iii) and (iv) we used Lemma 4.4 and 4.5, respectively. Next, notice that we can express $L(\pi^t, w)$ as a function of the state-action occupancy measure, $\mu^\pi(s,a) := \nu^\pi(s)\pi(a|s)$. In particular, we can rewrite $L(\pi, w) = \bar{L}(\mu^\pi, w)$, with

$$\bar{L}(\mu, w) := \langle r_w - \tau \log \pi^\mu, \mu \rangle - \langle w, \hat{\sigma}^E \rangle - \tau \mathbb{H}(\pi^E),$$

where π^μ denotes the policy induced by μ. It can be shown that $\bar{L}(\cdot, w)$ is concave [17]. Therefore, if $\bar{\mu} := \frac{1}{T}\sum_{t=0}^{T-1} \mu^{\pi^t}$ and $\bar{\pi} := \pi^{\bar{\mu}}$, we have by Jensen's inequality that

$$\frac{1}{T}\sum_{t=0}^{T-1} L(\pi^t, w) = \frac{1}{T}\sum_{t=0}^{T-1} \bar{L}(\mu^{\pi^t}, w) \leq \bar{L}\left(\bar{\mu}, w\right) = L(\bar{\pi}, w).$$

Plugging this back into (7), we have

$$\begin{aligned}
\mathbb{E}\left[J_{\bar{r}}^* - J_{\bar{r}}^{\pi^E}\right] \leq&\mathbb{E}\left[\min_{w \in W} L(\bar{\pi}, w)\right] + \mathcal{O}\left(\gamma^H\right) + \mathcal{O}\left(1/\sqrt{T}\right) \\
\leq&\varepsilon_{\text{real}} + \mathcal{O}\left(\gamma^H\right) + \mathcal{O}\left(1/\sqrt{T}\right),
\end{aligned}$$

where the last step follows by Assumption 4.2. To find the reward for which the expert is $(\varepsilon + \varepsilon_{\text{real}})$-optimal, we need to set $H = \mathcal{O}\left(\log \varepsilon^{-1}\right)$ and $T = \mathcal{O}(1/\varepsilon^2)$. Since each trajectory in $\operatorname{Est} Q(s, a, \pi, r, B)$ and $\operatorname{Est} \sigma(\pi^t, B)$ has an expected length of $1/(1-\gamma)$, the expected total number of samples used by Algorithm 1 is $2BT/\left(1-\gamma\right) = \mathcal{O}(1/\varepsilon^2)$. □

4.2 Policy convergence

In Theorem 4.3, we quantified the optimality of the expert for the recovered reward $\bar{r}$. In this section, we show that the optimal policy corresponding to the recovered reward is also close to the expert policy. We first introduce a total variation metric for measuring the distance to the expert policy

$$\max_s \|\pi^E(\cdot|s) - \pi(\cdot|s)\|_{\text{TV}}. \tag{8}$$

We define π to be ε-close to the expert policy if the total variation metric described above is bounded by ε. To ensure convergence in the policy, we need the following additional assumption. The state

occupancy measure $\nu^{\pi^E}(s)$ generated by the expert satisfies $\vartheta^E := \min_{s\in\mathcal{S}} \nu^{\pi^E}(s) > 0$. The above assumption ensures that the expert sufficiently explores the state space of the MDP. Similar to Assumption 4.1, it is satisfied when the initial distribution ν_0 is bounded away from zero. Under Assumption 4.2, we have the following convergence guarantee for the optimal policy corresponding to the recovered reward $\bar{r}$.

Corollary 4.6. *Suppose Assumptions 4.1, 4.2, and 4.2 hold, and let $\eta_w = \frac{1-\gamma}{\sqrt{kT}\|\phi\|_\infty}$. Algorithm 1 requires an expected number of $\mathcal{O}(1/\varepsilon^4)$ samples to ensure that the optimal policy corresponding to the recovered reward $\bar{r}$ is $(\varepsilon + \sqrt{\varepsilon_{real}})$-close to the expert policy.*

Proof. The result follows from

$$\begin{aligned} J^*_{\bar{r}} - J^{\pi^E}_{\bar{r}} &\overset{(i)}{=} \frac{\tau}{1-\gamma}\mathbb{E}_{s\sim\nu^E}\left[\mathrm{KL}(\pi^E(\cdot|s)||\pi^*_{r^{\bar{w}}}(\cdot|s))\right] \\ &= \frac{\tau}{1-\gamma}\sum_{s\in\mathcal{S}}\nu^E(s)\mathrm{KL}(\pi^E(\cdot|s)||\pi^*_{r^{\bar{w}}}(\cdot|s)) \\ &\overset{(ii)}{\geq} \frac{2\tau}{1-\gamma}\sum_{s\in\mathcal{S}}\nu^E(s)\|\pi^E(\cdot|s) - \pi^*_{r^{\bar{w}}}(\cdot|s)\|^2_{\mathrm{TV}} \\ &\overset{(iii)}{\geq} \frac{2\tau\vartheta^E}{1-\gamma}\max_s \|\pi^E(\cdot|s) - \pi^*_{r^{\bar{w}}}(\cdot|s)\|^2_{\mathrm{TV}}, \end{aligned}$$

where the equality (i) follows from the soft suboptimality Lemma [11, Lemma 26], (ii) holds by Pinsker's inequality, and (iii) by Assumption 4.2. Therefore, to ensure $\max_s \|\pi^E(\cdot|s) - \pi^*_{r^{\bar{w}}}(\cdot|s)\|_{\mathrm{TV}} \leq \varepsilon + \sqrt{\varepsilon_{\mathrm{real}}}$, we need $J^*_{\bar{r}} - J^{\pi^E}_{\bar{r}}$ to be upper bounded by $\Omega\left(\varepsilon^2 + \varepsilon_{\mathrm{real}}\right)$. By Theorem 4.3, we need an expected number of $\mathcal{O}(1/\varepsilon^4)$ samples in total. □

In Corollary 4.6, we show that the optimal policy corresponding to the reward recovered by Algorithm 1 converges to the expert's policy in the total variation metric (8). In the following, we demonstrate that (8) is a stronger metric for measuring policy convergence compared to the metrics used by [19, 20, 18, 10]. In particular, the authors of [19] prove convergence in the metric

$$\langle w_{\mathrm{true}}, \sigma^\pi - \sigma^{\pi^E}\rangle, \tag{9}$$

where $w_{\mathrm{true}} \in W$ is an unknown true reward parameter. Moreover, [20, 18, 10] provide their convergence guarantees in terms of the *integral probability metric* [12]

$$\delta_{\mathcal{R}}(\mu^\pi, \mu^{\pi^E}) := \max_{r\in\mathcal{R}}\langle r, \mu^\pi - \mu^{\pi^E}\rangle = \max_{w\in W}\langle w, \sigma^\pi - \sigma^{\pi^E}\rangle, \tag{10}$$

between the state-action occupancy measures μ^π and μ^{π^E}. The metric (10) measures the worst-case difference in the unregularized expected value between the recovered policy and the expert policy. It is easy to see that the integral probability metric (10) is stronger compared to the metric (9). In the following proposition, we demonstrate that the total variation metric (8) is a stronger metric for measuring policy convergence compared to the integral probability metric (10).

Proposition 4.7. *1) If the policy π is ε-close to the expert policy in the total variation metric, then it is also ε-close to the expert policy in the integral probability metric.*

2) Convergence in the integral probability metric does not imply convergence in the total variation metric.

Proof. To prove 1), we bound (10) as follows:

$$\begin{aligned} \max_{w\in W}\langle w, \sigma^\pi - \sigma^{\pi^E}\rangle &\overset{(i)}{\leq} \max_{w\in W}\|w\|_1\|\sigma^\pi - \sigma^{\pi^E}\|_\infty \\ &\overset{(ii)}{\leq} \|\sigma^\pi - \sigma^{\pi^E}\|_\infty \\ &\overset{(iii)}{\leq} \frac{2\|\phi\|_\infty}{(1-\gamma)^2}\max_s \|\pi^E(\cdot|s) - \pi(\cdot|s)\|_{\mathrm{TV}}, \end{aligned}$$

where (i) follows from Hölder's inequality, (ii) holds because $\|w\|_1 \leq 1$, and (iii) uses the Lipschitz continuity of σ^π with respect to π, as shown in Lemma B.7.

To prove 2), we consider a one-state MDP, as illustrated in Figure 1, where the policy is optimal in the integral probability metric (10) but is far from the expert policy in the total variation metric.

Consider an MDP with a single state $\mathcal{S} = \{s_1\}$ and two actions $\mathcal{A} = \{a_1, a_2\}$. We let the feature vector be a scalar constant $\phi(s, a) = 1$, and we consider the expert policy $\pi^E(a_1|s_1) = 1/2$ and the policy $\pi(a_2|s_1) = 1$. Then,

$$\max_{w \in W} \langle w, \sigma^\pi - \sigma^{\pi^E} \rangle = 0, \quad \max_s \|\pi^E(\cdot|s) - \pi(\cdot|s)\|_{\text{TV}} = \frac{1}{2}.$$

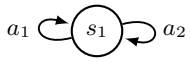

Figure 1: One-state MDP

□

As shown by Proposition 4.7, convergence in the total variation metric (8) is stronger than convergence in the integral probability metric (10). This is because the total variation metric directly measures the difference between policies, while the integral probability metric measures the difference in their unregularized expected value, ignoring the entropy regularization values. In the above example, the expert policy π^E has maximum entropy and is therefore realizable with $\varepsilon_{\text{real}} = 0$. Therefore, Algorithm 1 is guaranteed to recover a reward with a corresponding optimal policy that is ε-close to the expert policy in the total variation metric. However, in the case of a large realizability error $\varepsilon_{\text{real}}$, Corollary 4.6 fails to provide such strong guarantees. Hence, compared with [19, 20, 18, 10], we can get stronger convergence guarantees for the optimal policy corresponding to the recovered reward if the expert is realizable with a small realizability error $\varepsilon_{\text{real}}$.

5 Discussion and conclusion

We proposed a model-free single-loop algorithm to tackle the entropy-regularized IRL problem. Our algorithm simultaneously updates the policy using stochastic soft policy iteration and the reward parameters via stochastic projected gradient descent. We provided theoretical guarantees for the recovered reward and characterized the algorithm's sample complexity. Moreover, we demonstrated that the optimal policy under the recovered reward is close to the expert policy, measured using the total variation metric. Furthermore, we showed that this metric is stronger than the metrics used by [19, 20, 18, 10].

Since our proposed algorithm uses a stochastic soft policy iteration update for the policy, it requires re-estimating the state-action values for all state-action pairs at each time step. This may be infeasible for large state and action spaces and may require a large batch size to control the variance of the policy iterates. This highlights a limitation in our theoretical results, as we have only demonstrated convergence in expectation without providing a high probability bound. Therefore, a potential avenue for future research involves redesigning the policy update steps to establish convergence with a high probability bound and validating the algorithm on benchmarks or real-world scenarios.

References

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

[2] J. Bas-Serrano, S. Curi, A. Krause, and G. Neu. Logistic q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3610–3618. PMLR, 2021.

[3] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2022.

[4] J. Fu, K. Luo, and S. Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.

[5] M. Geist, B. Scherrer, and O. Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR, 2019.

[6] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*. PMLR, 2018.

[7] J. Ho and S. Ermon. Generative adversarial imitation learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2016.

[8] R. E. Kálmán. When is a linear control system optimal? *Journal of Basic Engineering*, 1964.

[9] G. Lan. Policy mirror descent for reinforcement learning: linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, 2021.

[10] Z. Liu, Y. Zhang, Z. Fu, Z. Yang, and Z. Wang. Learning from demonstration: Provably efficient adversarial policy imitation with linear function approximation. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[11] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[12] A. Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.

[13] D. A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.

[14] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

[15] S. Russell. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 101–103, 1998.

[16] J. Rust. Do people behave according to bellman's principle of optimality? *Journal of Economic Perspectives*, 1992.

[17] A. Schlaginhaufen and M. Kamgarpour. Identifiability and generalizability in constrained inverse reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023.

[18] L. Shani, T. Zahavy, and S. Mannor. Online apprenticeship learning. In *AAAI Conference on Artificial Intelligence*, 2021.

[19] U. Syed and R. E. Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems*, 2007.

[20] L. Viano, A. Kamoutsi, G. Neu, I. Krawczuk, and V. Cevher. Proximal point imitation learning. *Advances in Neural Information Processing Systems*, 35:24309–24326, 2022.

[21] D. Ying, Y. Ding, and J. Lavaei. A dual approach to constrained markov decision processes with entropy regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 1887–1909. PMLR, 2022.

[22] S. Zeng, C. Li, A. Garcia, and M. Hong. Maximum-likelihood inverse reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 35:10122–10135, 2022.

[23] B. D. Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.

[24] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, 2003.

A Proof of main lemmas

Lemma A.1 (Property of the estimators)**.** *The estimators $\hat{Q}_r^\pi(s,a)$ and $\hat{\sigma}^\pi$ returned by* $\operatorname{Est} Q(s,a,\pi,r,B)$ *and* $\operatorname{Est}\sigma(\pi,B)$ *respectively, are unbiased, meaning*

$$\mathbb{E}\left[\hat{Q}_r^\pi(s,a)\right] = Q_r^\pi(s,a),\ \mathbb{E}\left[\hat{\sigma}^\pi\right] = \sigma^\pi.$$

For the estimator $\hat{\sigma}^E$, we have, for all $w \in W$,

$$\langle \sigma^{\pi^E}, w\rangle - \frac{\gamma^H\|\phi\|_\infty}{1-\gamma} \le \langle \mathbb{E}\left[\hat{\sigma}^E\right], w\rangle.$$

For the reward gradient estimator $\hat{\nabla}_w L(\pi,w)$, we have

$$\mathbb{E}\|\hat{\nabla}_w L(\pi,w)\|_\infty \le \frac{2\|\phi\|_\infty}{1-\gamma},\ \mathbb{E}\|\hat{\nabla}_w L(\pi,w)\|_2^2 \le \frac{6k\|\phi\|_\infty^2}{(1-\gamma)^2}.$$

Here, the expectations are with respect to the randomness of the corresponding sampling strategies.

Proof of Lemma A.1. From [1, Assumption 6.3], we have $\mathbb{E}\left[\hat{Q}_r^\pi(s,a)\right] = Q_r^\pi(s,a)$ and (s_H, a_H) with $H \sim \operatorname{Geom}(1-\gamma)$ is sampled from the state-action occupancy measure μ^π. Therefore, we have

$$\mathbb{E}\left[\hat{\sigma}^\pi\right] = \mathbb{E}\left[\frac{1}{B(1-\gamma)}\sum_{i=1}^{B}\sum_{h=0}^{H_i}\phi(s_h^i,a_h^i)\right] = \mathbb{E}_{(s_h^i,a_h^i)\sim\mu^\pi(s,a)}\left[\frac{1}{B}\sum_{i=1}^{B}\sum_{h=0}^{H_i}\phi(s_h^i,a_h^i)\right] = \sigma^\pi.$$

For estimator $\hat{\sigma}^E$, we have

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^E\right] &= \mathbb{E}_{\pi^E}\left[\sum_{t=0}^{H-1}\gamma^t\phi(s,a)\right] \\
&= \mathbb{E}_{\pi^E}\left[\sum_{t=0}^{\infty}\gamma^t\phi(s,a)\right] - \mathbb{E}_{\pi^E}\left[\sum_{t=H}^{\infty}\gamma^t\phi(s,a)\right] \\
&= \sigma^{\pi^E} - \mathbb{E}_{\pi^E}\left[\sum_{t=H}^{\infty}\gamma^t\phi(s,a)\right].
\end{aligned}$$

Taking the inner product with $w \in W$ on both sides of the above equality, we have

$$\begin{aligned}
\langle \mathbb{E}\left[\hat{\sigma}^E\right], w\rangle &= \langle \sigma^{\pi^E}, w\rangle - \mathbb{E}_{\pi^E}\left[\sum_{t=H}^{\infty}\gamma^t\langle\phi(s,a),w\rangle\right] \\
&\ge \langle \sigma^{\pi^E}, w\rangle - \frac{\gamma^H\|\phi\|_\infty}{1-\gamma},
\end{aligned}$$

where the last inequality follows from $\langle\phi(s,a),w\rangle \le \|\phi\|_\infty\|w\|_1 \le \|\phi\|_\infty$.

For reward gradient estimator $\hat{\nabla}_w L(\pi,w)$, we have

$$\begin{aligned}
\mathbb{E}\left\|\hat{\nabla}_w L(\pi,w)\right\|_\infty = \mathbb{E}\left\|\hat{\sigma}^\pi - \hat{\sigma}^E\right\|_\infty &= \mathbb{E}\left\|\frac{1}{B}\sum_{i=1}^{B}\sum_{h=0}^{H_i}\phi(s_h^i,a_h^i) - \frac{1}{N}\sum_{j=1}^{N}\sum_{h=0}^{H-1}\gamma^h\phi(s_h^j,a_h^j)\right\|_\infty \\
&\le \mathbb{E}\left\|\frac{1}{B}\sum_{i=1}^{B}\sum_{h=0}^{H_i}\phi(s_h^i,a_h^i)\right\|_\infty + \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{h=0}^{H-1}\gamma^h\phi(s_h^i,a_h^i)\right\|_\infty \\
&\overset{(i)}{\le} \|\phi\|_\infty\mathbb{E}_{H\sim\operatorname{Geom}(1-\gamma)}H + \frac{\|\phi\|_\infty}{1-\gamma} \\
&\le \frac{(1+\gamma)\|\phi\|_\infty}{1-\gamma} \le \frac{2\|\phi\|_\infty}{1-\gamma},
\end{aligned}$$

where (i) we apply the expectation of $\mathrm{Geom}(1-\gamma)$ is $\frac{\gamma}{1-\gamma}$. Similarly, we can bound $\mathbb{E}\left\|\hat{\nabla}_w L(\pi,w)\right\|_2^2$ as follows:

$$\begin{aligned}
\mathbb{E}\left\|\hat{\nabla}_w L(\pi,w)\right\|_2^2 &= \mathbb{E}\left\|\frac{1}{B}\sum_{i=1}^{B}\sum_{h=0}^{H_i}\phi(s_h^i,a_h^i) - \frac{1}{N}\sum_{j=1}^{N}\sum_{h=0}^{H-1}\gamma^h\phi(s_h^j,a_h^j)\right\|_2^2 \\
&\le 2\mathbb{E}\left\|\frac{1}{B}\sum_{i=1}^{B}\sum_{h=0}^{H_i}\phi(s_h^i,a_h^i)\right\|_2^2 + 2\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{h=0}^{H-1}\gamma^h\phi(s_h^i,a_h^i)\right\|_2^2 \\
&\overset{(i)}{\le} 2k\|\phi\|_\infty^2\mathbb{E}_{H\sim\mathrm{Geo}(1-\gamma)}H^2 + \frac{2k\|\phi\|_\infty^2}{(1-\gamma)^2} \\
&\le \frac{2k(1+\gamma+\gamma^2)\|\phi\|_\infty^2}{(1-\gamma)^2} \le \frac{6k\|\phi\|_\infty^2}{(1-\gamma)^2},
\end{aligned}$$

where (i) we apply the fist moment of $\mathrm{Geom}(1-\gamma)$ is $\frac{\gamma+\gamma^2}{(1-\gamma)^2}$. □

Lemma 4.4. *Suppose Assumption 4.1 holds and let $\eta_w = \frac{1-\gamma}{\sqrt{kT}\|\phi\|_\infty}$. We can bound the suboptimality of the policy iterates of Algorithm 1 by*

$$\mathbb{E}\left[\max_\pi L(\pi,w^t) - L(\pi^t,w^t)\right] \le \mathcal{O}\left(1/\sqrt{T}\right).$$

Proof of Lemma 4.4. We first upper bound the improvement of the suboptimality gap as follows:

$$\begin{aligned}
&\mathbb{E}\left[\left(\max_\pi L(\pi,w^{t+1}) - L(\pi^{t+1},w^{t+1})\right) - \left(\max_\pi L(\pi,w^t) - L(\pi^t,w^t)\right)\right] \\
=&\mathbb{E}\left[\left(J^*_{r^{t+1}} - J^{\pi^{t+1}}_{r^{t+1}}\right) - \left(J^*_{r^t} - J^{\pi^t}_{r^t}\right)\right] \\
\overset{(i)}{=}&\mathbb{E}\left[J^*_{r^{t+1}} - J^{\pi^{t+1}}_{r^{t+1}} - J^{\pi^{t+1}}_{r^t} + J^{\pi^{t+1}}_{r^t} - J^*_{r^t} + J^{\pi^t}_{r^t}\right] \\
\le&\mathbb{E}\left[\left|J^*_{r^{t+1}} - J^*_{r^t}\right| + \left|J^{\pi^{t+1}}_{r^{t+1}} - J^{\pi^{t+1}}_{r^t}\right| - \left(J^{\pi^{t+1}}_{r^t} - J^{\pi^t}_{r^t}\right)\right] \\
\overset{(ii)}{\le}&\frac{2\eta_w\|\phi\|_\infty}{(1-\gamma)^2} - \mathbb{E}\left[J^{\pi^{t+1}}_{r^t} - J^{\pi^t}_{r^t}\right] \\
\overset{(iii)}{\le}&\frac{2\eta_w\|\phi\|_\infty}{(1-\gamma)^2} - \frac{\tau}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{s\sim\nu^{\pi^{t+1}}}\mathrm{KL}(\pi^t(\cdot|s)||\pi^{t+1}(\cdot|s))\right] \\
\le&\frac{2\eta_w\|\phi\|_\infty}{(1-\gamma)^2} - \frac{\tau}{1-\gamma}\mathbb{E}\left[\left(\min_s\frac{\nu^{\pi^{t+1}}(s)}{\nu^{\pi^*_{r^t}}(s)}\right)\mathbb{E}_{s\sim\nu^{\pi^*_{r^t}}}\mathrm{KL}(\pi^t(\cdot|s)||\pi^{t+1}(\cdot|s))\right] \\
\overset{(iv)}{\le}&\frac{2\eta_w\|\phi\|_\infty}{(1-\gamma)^2} - \frac{\tau}{\vartheta(1-\gamma)}\mathbb{E}\left[\mathbb{E}_{s\sim\nu^{\pi^*_{r^t}}}\mathrm{KL}(\pi^t(\cdot|s)||\pi^{t+1}(\cdot|s))\right] \\
\overset{(v)}{\le}&\frac{2\eta_w\|\phi\|_\infty}{(1-\gamma)^2} - \frac{1}{\vartheta}\mathbb{E}\left[J^*_{r^t} - J^{\pi^t}_{r^t}\right] \\
=&\frac{2\eta_w\|\phi\|_\infty}{(1-\gamma)^2} - \frac{1}{\vartheta}\mathbb{E}\left[\max_\pi L(\pi,w^t) - L(\pi^t,w^t)\right]
\end{aligned}$$

where (i) holds by adding and subtracting $J^{\pi^{t+1}}_{r^t}$, (ii) holds by Lemma B.5, (iii) holds by Lemma B.3, (iv) holds by Assumption 4.1 and (v) holds by Lemma B.4. Rearrange the above inequality, we have

$$\mathbb{E}\left[\max_\pi L(\pi,w^{t+1}) - L(\pi^{t+1},w^{t+1})\right] \le \frac{2\eta_w\|\phi\|_\infty}{(1-\gamma)^2} + \left(1-\frac{1}{\vartheta}\right)\mathbb{E}\left[\max_\pi L(\pi,w^t) - L(\pi^t,w^t)\right].$$

Recursively applying the above inequality, we obtain

$$\begin{aligned}\mathbb{E}\left[\max_{\pi} L(\pi, w^t) - L(\pi^t, w^t)\right] \leq & \frac{2\eta_w \vartheta \|\phi\|_\infty}{(1-\gamma)^2} + \left(1 - \frac{1}{\vartheta}\right)^t \left(\max_{\pi} L(\pi, w^0) - L(\pi^0, w^0)\right) \\ = & \frac{2\vartheta}{\sqrt{kT}(1-\gamma)},\end{aligned}$$

where we apply $\eta_w = \frac{1-\gamma}{\sqrt{kT}\|\phi\|_\infty}$ and $\max_\pi L(\pi, w^0) - L(\pi^0, w^0) = 0$ since $r^{(0)} = 0$ in the last step. □

Lemma 4.5 (Stochastic online gradient descent regret)**.** *If we set the learning rate to* $\eta_w = \frac{1-\gamma}{\sqrt{kT}\|\phi\|_\infty}$*, we have*

$$\mathbb{E}\left[\sum_{t=0}^{T-1} L(\pi^t, w^t)\right] \leq \mathbb{E}\left[\min_{w \in W} \sum_{t=0}^{T-1} L(\pi^t, w)\right] + \mathcal{O}\left(\sqrt{T}\right).$$

Proof. We start from the projected descent step

$$\begin{aligned}\|w^{(t+1)} - w^*\|_2^2 = & \left\|\mathcal{P}_W\left(w^t - \eta_w \hat{\nabla}_w L(\pi^t, w^t)\right) - w^*\right\|_2^2 \\ \overset{(i)}{\leq} & \left\|w^t - \eta_w \hat{\nabla}_w L(\pi^t, w^t) - w^*\right\|_2^2 \\ = & \|w^t - w^*\|_2^2 - 2\eta_w \langle \hat{\nabla}_w L(\pi^t, w^t), w^t - w^* \rangle + \eta_w^2 \|\hat{\nabla}_w L(\pi^t, w^t)\|_2^2,\end{aligned}$$

where (i) holds by the non expansiveness of projection. Therefore

$$\langle \hat{\nabla}_w L(\pi^t, w^t), w^t - w^* \rangle \leq \frac{\|w^t - w^*\|_2^2 - \|w^{t+1} - w^*\|_2^2}{2\eta_w} + \frac{\eta_w}{2} \|\hat{\nabla}_w L(\pi^t, w^t)\|_2^2.$$

Taking the expectation on both sides, we have

$$\mathbb{E}\left[\left\langle \nabla_w L(\pi^t, w^t), w^t - w^* \right\rangle\right] \leq \mathbb{E}\left[\frac{\|w^t - w^*\|_2^2 - \|w^{t+1} - w^*\|_2^2}{2\eta_w} + \frac{3k\eta_w \|\phi\|_\infty^2}{(1-\gamma)^2}\right], \tag{11}$$

where we use Lemma A.1. Let w^* be the following optimizer:

$$w^* = \arg\min_{w \in W} \sum_{t=0}^{T-1} L(\pi^t, w).$$

We have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1} L(\pi^t, w^t) - \min_{w \in W} \sum_{t=0}^{T-1} L(\pi^t, w)\right] = & \mathbb{E}\left[\sum_{t=0}^{T-1} L(\pi^t, w^t) - \sum_{t=0}^{T-1} L(\pi^t, w^*)\right] \\ \overset{(i)}{\leq} & \mathbb{E}\left[\sum_{t=0}^{T-1} \left\langle \nabla_w L(\pi^t, w^t), w^t - w^* \right\rangle\right] \\ \overset{(ii)}{\leq} & \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\|w^t - w^*\|_2^2 - \|w^{t+1} - w^*\|_2^2}{2\eta_w} + \frac{3kT\eta_w \|\phi\|_\infty^2}{(1-\gamma)^2}\right] \\ \leq & \mathbb{E}\left[\frac{\|w^{T-1} - w^*\|_2^2}{2\eta_w}\right] + \frac{3kT\eta_w \|\phi\|_\infty^2}{(1-\gamma)^2} \\ \leq & \frac{1}{\eta_w} + \frac{3kT\eta_w \|\phi\|_\infty^2}{(1-\gamma)^2} \\ \overset{(iii)}{\leq} & \frac{3\sqrt{2kT}\|\phi\|_\infty}{1-\gamma},\end{aligned}$$

where (i) holds by convexity, (ii) by (11), and (iii) by plugging in $\eta_w = \frac{1-\gamma}{\sqrt{2kT}\|\phi\|_\infty}$. □

B Supporting lemmas

Lemma B.1 (Soft suboptimality [11, Lemma 26]). *For any policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ and reward $r \in \mathcal{R}$, we have*

$$J_r^* - J_r^\pi = \frac{\tau}{1-\gamma}\mathbb{E}_{s\sim\nu^\pi} KL(\pi(\cdot|s)||\pi^*(\cdot|s)).$$

Lemma B.2 (Soft performance difference [9]). *For any two policies $\pi, \pi' \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, we have*

$$J_r^\pi - J_r^{\pi'} = \frac{1}{1-\gamma}\left(\mathbb{E}_{(s,a)\sim\mu^\pi}\left[A_r^{\pi'}(s,a)\right] - \tau\mathbb{E}_{s\sim\nu^\pi}\left[\mathrm{KL}\left(\pi(\cdot|s)||\pi'(\cdot|s)\right)\right]\right),$$

where $A_r^{\pi'}(s,a) := Q_r^{\pi'}(s,a) - V_r^{\pi'}(s) - \tau\log\pi'(a|s)$ is the advantage function.

Lemma B.3 (Performance improvement for the policy).

$$\mathbb{E}\left[J_{r^t}^{\pi^{t+1}} - J_{r^t}^{\pi^t}\right] = \frac{\tau}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{s\sim\nu^{\pi^{t+1}}}\left[\mathrm{KL}\left(\pi^t(\cdot|s)||\pi^{t+1}(\cdot|s)\right)\right]\right]$$

This result extends [3, Lemma 1] to the stochastic setting.

Proof. From the soft value iteration update,

$$\pi^{t+1}(a|s) = \frac{1}{Z^t(s)}\exp\left(\hat{Q}_{r^t}^{\pi^t}(s,a)/\tau\right) = \frac{1}{Z^t(s)}\exp\left(\frac{Q_{r^t}^{\pi^t}(s,a) + \Delta^t(s,a)}{\tau}\right),$$

where $\Delta^t(s,a) := \hat{Q}_{r^t}^{\pi^t}(s,a) - Q_{r^t}^{\pi^t}(s,a)$. It follows that

$$\tau\log Z^t(s) = Q_{r^t}^{\pi^t}(s,a) + \Delta^t(s,a) - \tau\log\pi^{t+1}(a|s). \tag{12}$$

Let Using Lemma B.2, we have that

$$\mathbb{E}\left[J_{r^t}^{\pi^{t+1}} - J_{r^t}^{\pi^t}\right] = \frac{1}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{(s,a)\sim\mu^{\pi^{t+1}}}\left[A_{r^t}^{\pi^t}(s,a)\right] - \tau\mathbb{E}_{s\sim\nu^{\pi^{t+1}}}\left[\mathrm{KL}\left(\pi^{t+1}(\cdot|s)||\pi^t(\cdot|s)\right)\right]\right]$$

$$\begin{aligned}
&\overset{(i)}{=} \frac{1}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{(s,a)\sim\mu^{\pi^{t+1}}}\left[Q_{r^t}^{\pi^t}(s,a) - V_{r^t}^{\pi^t}(s) - \tau\log\pi^{t+1}(a|s)\right]\right]\\
&\overset{(ii)}{=} \frac{1}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{s\sim\nu^{\pi^{t+1}}}\mathbb{E}_{a\sim\pi^{t+1}(\cdot|s)}\left[\tau\log Z^t(s) - V_{r^t}^{\pi^t}(s) - \Delta^t(s,a)\right]\right]\\
&\overset{(iii)}{=} \frac{1}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{s\sim\nu^{\pi^{t+1}}}\mathbb{E}_{a\sim\pi^t(\cdot|s)}\left[\tau\log Z^t(s) - V_{r^t}^{\pi^t}(s)\right]\right]\\
&\overset{(iv)}{=} \frac{1}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{s\sim\nu^{\pi^{t+1}}}\mathbb{E}_{a\sim\pi^t(\cdot|s)}\left[Q_{r^t}^{\pi^t}(s,a) + \Delta^t(s,a) - \tau\log\pi^{t+1}(a|s) - V_{r^t}^{\pi^t}(s)\right]\right]\\
&\overset{(v)}{=} \frac{1}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{s\sim\nu^{\pi^{t+1}}}\mathbb{E}_{a\sim\pi^t(\cdot|s)}\left[A_{r^t}^{\pi^t}(s,a) + \Delta^t(s,a) + \tau\log\frac{\pi^t(a|s)}{\pi^{t+1}(a|s)}\right]\right]\\
&\overset{(vi)}{=} \frac{\tau}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{s\sim\nu^{\pi^{t+1}}}\left[\mathrm{KL}\left(\pi^t(\cdot|s)||\pi^{t+1}(\cdot|s)\right)\right]\right].
\end{aligned}$$

Here, we use the definition of the advantage in (i) and (12) in (ii). In (iii) we use

$$\mathbb{E}\left[\mathbb{E}_{s\sim\nu^{\pi^{t+1}}}\mathbb{E}_{a\sim\pi^t(\cdot|s)}\left[\Delta^t(s,a)\right]\right] = \mathbb{E}\left[\mathbb{E}_{s\sim\nu^{\pi^{t+1}}}\mathbb{E}_{a\sim\pi^t(\cdot|s)}\left[\mathbb{E}\left[\Delta^t(s,a)|\pi^t, r^t\right]\right]\right] = 0, \tag{13}$$

where the last equality follows from Lemma A.1. In (iv) we again plug in (12). Finally, (v) follows from rearranging and (vi) from Equation (13) and $\mathbb{E}_{a\sim\pi^t(\cdot|s)}\left[A_{r^t}^{\pi^t}(s,a)\right] = 0$. □

Lemma B.4 (suboptimality gap for policy). *For any iterates r^t and π^t generated by Algorithm 1, we have*

$$\mathbb{E}\left[J_{r^t}^* - J_{r^t}^{\pi^t}\right] \le \frac{\tau}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{s\sim\nu_{r^t}^*}\left[\mathrm{KL}\left(\pi^t(\cdot|s)||\pi^{t+1}(\cdot|s)\right)\right]\right].$$

This result extends [3, Lemma 5] to the stochastic setting.

Proof. From Lemma B.2, it follows that

$$\begin{aligned}
\mathbb{E}\left[J^*_{r^t} - J^{\pi^t}_{r^t}\right] &= \frac{1}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{(s,a)\sim\mu^*_{r^t}}\left[A^{\pi^t}_{r^t}(s,a)\right] - \tau\mathbb{E}_{s\sim\nu^*_{r^t}}\left[\mathrm{KL}\left(\pi^*_{r^t}(\cdot|s)||\pi^t(\cdot|s)\right)\right]\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{(s,a)\sim\mu^*_{r^t}}\left[Q^{\pi^t}_{r^t}(s,a) - V^{\pi^t}_{r^t}(s) - \tau\log\pi^*_{r^t}(a|s)\right]\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{s\sim\nu^*_{r^t}}\left[\underbrace{\mathbb{E}_{a\sim\pi^*_{r^t}(\cdot|s)}\left[Q^{\pi^t}_{r^t}(s,a) + \Delta^t(s,a) - \tau\log\pi^*_{r^t}(a|s)\right]}_{(A)} - \underbrace{V^{\pi^t}_{r^t}(s)}_{(B)}\right]\right],
\end{aligned} \tag{14}$$

where we used $\Delta^t(s,a) := \hat{Q}^{\pi^t}_{r^t}(s,a) - Q^{\pi^t}_{r^t}(s,a)$ and Equation (13) in the last step. Next, we bound (A) and (B) separately. For (A) we have by Jensen's inequality

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi^*_{r^t}(\cdot|s)}\left[Q^{\pi^t}_{r^t}(s,a) + \Delta^t(s) - \tau\log\pi^*_{r^t}(a|s)\right] \\
=&\tau\sum_{a\in\mathcal{A}}\pi^*_{r^t}(a|s)\log\left(\frac{\exp\left(\left(Q^{\pi^t}_{r^t}(s,a) + \Delta^t(s,a)\right)/\tau\right)}{\pi^*_{r^t}(a|s)}\right) \\
\leq&\tau\log\left(\sum_{a\in\mathcal{A}}\exp\left(\left(Q^{\pi^t}_{r^t}(s,a) + \Delta^t(s,a)\right)/\tau\right)\right) = \tau\log Z^t(s).
\end{aligned}$$

For (B) the definition of the value function and the soft policy iteration update yield

$$\begin{aligned}
V^{\pi^t}_{r^t}(s) &= \mathbb{E}_{a\sim\pi^t(\cdot|s)}\left[Q^{\pi^t}_{r^t}(s,a) - \tau\log\pi^t(a|s)\right] \\
&= \mathbb{E}_{a\sim\pi^t(\cdot|s)}\left[Q^{\pi^t}_{r^t}(s,a) - \tau\log\pi^{t+1}(a|s)\right] - \tau\,\mathrm{KL}\left(\pi^t(\cdot|s)||\pi^{t+1}(\cdot|s)\right) \\
&= \mathbb{E}_{a\sim\pi^t(\cdot|s)}\left[\tau\log Z^t(s) - \Delta^t(s,a)\right] - \tau\,\mathrm{KL}\left(\pi^t(\cdot|s)||\pi^{t+1}(\cdot|s)\right).
\end{aligned}$$

Plugging these bounds for (A) and (B) back into (14), using again that Equation (13), we arrive at the desired inequality

$$\mathbb{E}\left[J^*_{r^t} - J^{\pi^t}_{r^t}\right] \leq \frac{1}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{s\sim\nu^*_{r^t}}\left[\tau\,\mathrm{KL}\left(\pi^t(\cdot|s)||\pi^{t+1}(\cdot|s)\right)\right]\right].$$

□

Lemma B.5. *For any reward iterates r_t, r_{t+1} generated by Algorithm 1 and any policy $\pi \in \Delta^{\mathcal{S}}_{\mathcal{A}}$, we have*

$$|J^\pi_{r^t} - J^\pi_{r^{t+1}}| \leq \frac{2\eta_w\|\phi\|_\infty}{(1-\gamma)^2}, \tag{15}$$

$$|J^*_{r^t} - J^*_{r^{t+1}}| \leq \frac{2\eta_w\|\phi\|_\infty}{(1-\gamma)^2}. \tag{16}$$

Proof. Inequality (15) holds since

$$\begin{aligned}
|J^\pi_{r^t} - J^\pi_{r^{t+1}}| &\overset{(i)}{=} \left|\mathbb{E}_\pi\left[\sum_{t=0}^{+\infty}\gamma^t\big(r_1(s_t,a_t) - \cancel{\tau\log\pi(a_t|s_t)} - r^{t+1}(s_t,a_t) + \cancel{\tau\log\pi(a_t|s_t)}\big)\right]\right| \\
&= \left|\mathbb{E}_\pi\left[\sum_{t=0}^{+\infty}\gamma^t\big(r^t(s_t,a_t) - r^{t+1}(s_t,a_t)\big)\right]\right| \\
&\leq \left[\sum_{t=0}^{+\infty}\gamma^t\left\|r^t - r^{t+1}\right\|_\infty\right]
\end{aligned}$$

$$\overset{(ii)}{=} \eta_w \sum_{t=0}^{+\infty} \gamma^t \left\| \hat{\nabla}_w L(\pi, w^t) \right\|_\infty$$
$$\overset{(iii)}{\leq} \frac{2\eta_w \|\phi\|_\infty}{(1-\gamma)^2},$$

where (i) holds by the definition of J_r^π, (ii) holds by plugging reward updating form and non-expansiveness of projection and (iii) holds by Lemma A.1.

Inequality (16) holds since

$$|J^*_{r^t} - J^*_{r^{t+1}}| \leq \max_{\pi \in \Pi} |J^\pi_{r^t} - J^\pi_{r^{t+1}}| \overset{(i)}{\leq} \frac{2\eta_w \|\phi\|_\infty}{(1-\gamma)^2},$$

where (i) holds by inequality (15). □

Lemma B.6 (Lipschitz continuity of occupancy measure in policy)**.** *Let μ^π denote the occupancy measure corresponding to the policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$. Then, for any $\pi_1, \pi_2 \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ we have*

$$\|\mu^{\pi_1} - \mu^{\pi_2}\|_1 \leq \frac{1}{1-\gamma} \max_s \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_1 .$$

Proof. We can upper bound $\|\mu^{\pi_1} - \mu^{\pi_2}\|_1$ as follows

$$\begin{aligned}\|\mu^{\pi_1} - \mu^{\pi_2}\|_1 &\leq \sum_{s,a} |\nu^{\pi_1}(s)(\pi_1(a|s) - \pi_2(a|s))| + \sum_{s,a} |(\nu^{\pi_1}(s) - \nu^{\pi_2}(s))\pi_2(a|s)| \\ &\leq \max_s \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_1 + \|\nu^{\pi_1} - \nu^{\pi_2}\|_1 ,\end{aligned}$$

where we used the triangle and Hölder's inequality. From the Bellman flow constraints [14]

$$\nu^\pi(s) = \gamma \sum_{s',a'} P(s|s',a')\mu^\pi(s',a') + (1-\gamma)\nu_0(s),$$

it follows that

$$\begin{aligned}\|\nu^{\pi_1} - \nu^{\pi_2}\|_1 &= \gamma \sum_s \left| \sum_{s',a'} P(s|s',a')(\mu^{\pi_1}(s',a') - \mu^{\pi_2}(s',a')) \right| \\ &\leq \gamma \sum_{s',a'} \underbrace{\sum_s P(s|s',a')}_{=1} |\mu^{\pi_1}(s',a') - \mu^{\pi_2}(s',a')| \\ &= \gamma \|\mu^{\pi_1} - \mu^{\pi_2}\|_1 ,\end{aligned}$$

where we again used the triangle inequality. Hence, it follows that

$$\max_s \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_1 \geq \|\mu^{\pi_1} - \mu^{\pi_2}\|_1 - \|\nu^{\pi_1} - \nu^{\pi_2}\|_1 \geq (1-\gamma) \|\mu^{\pi_1} - \mu^{\pi_2}\|_1 .$$

□

Lemma B.7. *For any two policies $\pi_1, \pi_2 \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, we have*

$$\|\sigma^{\pi_1} - \sigma^{\pi_2}\|_\infty \leq \frac{2\|\phi\|_\infty}{(1-\gamma)^2} \max_s \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_{TV}.$$

Proof. The proof follows from Hölder's inequality and Lemma B.6 above:

$$\begin{aligned}\|\sigma^{\pi_1} - \sigma^{\pi_2}\|_\infty &= \max_{1 \leq i \leq k} \max_{s,a} \left| \frac{1}{1-\gamma} \langle \phi_i, \mu^{\pi_1} - \mu^{\pi_2} \rangle \right| \\ &\leq \frac{1}{1-\gamma} \max_{1 \leq i \leq k} \|\phi_i\|_\infty \|\mu^{\pi_1} - \mu^{\pi_2}\|_1 \\ &\leq \frac{\|\phi\|_\infty}{(1-\gamma)^2} \max_s \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_1 \\ &\leq \frac{2\|\phi\|_\infty}{(1-\gamma)^2} \max_s \|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_{TV} .\end{aligned}$$

□