Stochastic convex optimization for provably efficient apprenticeship learning

Angeliki Kamoutsi
ETH Zurich
kamoutsa@ethz.ch

Goran Banjac
ETH Zurich
gbanjac@ethz.ch

John Lygeros
ETH Zurich
jlygeros@ethz.ch

Abstract

We consider large-scale Markov decision processes (MDPs) with an unknown cost function and employ stochastic convex optimization tools to address the problem of imitation learning, which consists of learning a policy from a finite set of expert demonstrations. We adopt the apprenticeship learning formalism, which carries the assumption that the true cost function can be represented as a linear combination of some known features. Existing inverse reinforcement learning algorithms come with strong theoretical guarantees, but are computationally expensive because they use reinforcement learning or planning algorithms as a subroutine. On the other hand state-of-the-art policy gradient based algorithms (like IM-REINFORCE, IM-TRPO and GAIL), achieve significant empirical success in challenging benchmark tasks, but are less well understood in terms of theory. With an emphasis on non-asymptotic guarantees of performance, we propose a method that directly learns a policy from expert demonstrations, bypassing the intermediate step of learning the cost function, by formulating the problem as a single convex optimization problem over occupancy measures. We develop a computationally efficient algorithm and derive high confidence excess-loss bounds on the quality of the extracted policy, utilizing results from uncertain convex optimization and recent works in approximate linear programming for solving forward MDPs.

1 Introduction

The goal of apprenticeship learning (AL) in a Markov decision process (MDP) environment without cost function is to learn a policy that achieves or even surpasses the performance of a policy demonstrated by an expert. A usual assumption is that the unknown true cost function can be represented as a weighted combination of some known basis functions, where the true unknown weights specify how different desiderata should be traded off. An argument for this assumption is that in practice the unknown cost function depends on just a few key properties, but the desirable weighting is unknown.

A lot of methods have been proposed to solve the apprenticeship learning problem. The most naive approach is *behavior cloning*, which casts the problem as a supervised learning problem, in which the goal is to learn a map from states to optimal actions. Although behavior cloning is simple and easy to implement, the crucial i.i.d. assumption made in supervised learning is violated. As a result, the approach suffers from the problem of *cascading errors* which is related to *covariate shift* [13]. Later works like DAgger[19] eliminate distribution mismatch by formulating the problem as a no regret algorithm in an online learning setting. However, these kind of algorithms require interaction with the expert, which is a different learning scenario from the one considered in this paper. Most importantly their sample and computational complexity scales polynomially with the horizon of the problem, which in our case is infinite.

Inverse Reinforcement learning (IRL) [4] is a prevalent approach to AL. In this paradigm, the learner first infers the unknown cost function that the expert tries to optimize and then uses it to

Optimization Foundations for Reinforcement Learning Workshop at NeurIPS 2019, Vancouver, Canada.

reproduce the optimal behavior. IRL algorithms do not suffer from the problem of cascading errors because the training takes place over entire expert trajectories, rather than individual actions. In addition, since the recovered cost function "explains" the expert behavior, they can easily generalize to unseen states or even new MDP environments. Note however, that most existing IRL algorithms [4, 22, 17, 3, 25, 18, 16, 14, 15] are computationally expensive because they use reinforcement learning as a subroutine.

On the other hand, one can frame the problem as a single convex program [23], bypassing the intermediate step of learning the cost function. Although the associated program can be solved exactly for small-sized MDPs, the approach suffers from the curse of dimensionality, making it intractable for large-scale problems, arising in, e.g., autonomous driving with increasing number of sensors and decision aspects. Provably efficient convex approximation schemes for the convex formulation of AL [23] in the context of large-scale MDPs remain unexplored. However, it is worth noting, that the formulations and reasoning in [23] formed the ground and inspired later state-of-the-art algorithms [13, 12]. In particular, the authors in [13] developed a gradient-based optimization formulation over parameterized policies for apprenticeship learning, and then presented algorithms which are parallel to the policy gradient RL counterparts [24, 20]. The sequel paper [12], draws a connection between the policy optimization formulation and generative adversarial networks [7], from which an analogous imitation learning algorithm is derived. These approaches are model-free and scale to large and continuous environments. However, in general the policy optimization problem is highly non-convex and as a result remains hampered by limited theoretical understanding. In particular, these methods provide no guarantees into the quality of the points to which they converge.

With an emphasis on non-asymptotic guarantees of stability and performance, in this work we propose an approximation scheme for the convex formulation of AL [23]. In particular, the objective is to minimize the ℓ_1-distance between the feature expectation vector of the expert and the learner, subject to linear constraints ensuring that the optimization variable is an occupation measure induced by a policy. Our AL algorithm and its theoretical analysis builds upon recent innovations in approximate LP for large-scale discounted MDPs [1, 2] and can be seen like the AL analogue of their algorithhms. Similarly to [1], we control the complexity by limiting our search to the linear subspace defined by a small number of features. We then convert the initial program to an unconstrained convex optimization problem. To this end, we use a surrogate loss function by adding a multiple of the total constraint violations to the initial objective. We then construct unbiased subgradient estimators and apply the stochastic subgradient descent algorithm. In this way, by combining bounds in the stochastic convex optimization literature and concentration inequalities, we are able to give high confidence regret bounds showing that the performance of our algorithm approaches the best achievable by any policy in the comparison class. A salient feature of the algorithm is that the iteration and sample complexity do not depend on the size of the state space but instead on the number of approximation features.

Moreover, it is worth mentioning that since our methodology is based on the LP formulation of MDPs [9, 10, 11, 5], is naturally extensible to unconventional problems involving additional safety constraints or secondary costs, where traditional dynamic programming techniques are not applicable [8, 6, 21].

To the best of our knowledge this is the first time that a performance bound is derived for a policy-optimization-based algorithm for AL. We hope that the techniques proposed in this work provide a starting point for developing provably efficient AL algorithms

Notation and conventions. We denote by $A_{i,:}$ and $A_{:,j}$ the i-th row and j-th column of a matrix A, respectively. For $p \in [1, \infty]$, we denote by $\|\cdot\|_p$ the p-norm in $\mathbf{R}^n$. The corresponding induced matrix norm is defined by $\| A \|_p = \sup_{\| x \|_p \leq 1} \| Ax \|_p$. For vectors x and y, we denote by $\langle x, y \rangle$ the usual inner product. Moreover, $x \leq y$ denotes elementwise inequality. We define $[x]_+ = \max\{0, x\}$ and $[x]_- = -\min\{0, x\}$. The set of probability measures on a set X is denoted by $\mathcal{P}(X)$.

2 Preliminaries

Consider a finite MDP described by a tuple $\mathcal{M}_c \triangleq \big(\mathcal{X}, \mathcal{A}, P, \gamma, \nu_0, c\big)$, where $\mathcal{X} = \{x_1, \ldots, x_{|\mathcal{X}|}\}$ is the state space, $\mathcal{A} = \{a_1, \ldots, a_{|\mathcal{A}|}\}$ is the action space, $P : \mathcal{X} \times \mathcal{A} \to \mathcal{P}(\mathcal{X})$ is the transition law, $\gamma \in (0, 1)$ is the discount factor, $\nu_0 \in \mathcal{P}(\mathcal{X})$ is the initial probability distribution of the system state, and $c : \mathcal{X} \times \mathcal{A} \to \mathbf{R}$ is the one-stage cost function.

The model $\mathcal{M}_c$ represents an infinite horizon controlled discrete-time stochastic system whose evolution is described as follows. At time step t, if the system is in state $x_t = x \in \mathcal{X}$, and the action $a_t = a \in \mathcal{A}$ is taken, then (i) the cost $c(x, a)$ is incurred, and (ii) the system moves to the next state x_{t+1}, which is an $\mathcal{X}$-valued random variable with probability distribution $P(\cdot|x, a)$. Once transition into the new state has occurred, a new action is chosen and the process is repeated.

A stationary Markov policy is a map $\pi : \mathcal{X} \mapsto \mathcal{P}(\mathcal{A})$, and $\pi(a|x)$ denotes the probability of choosing action a, while being in state x. We denote the space of stationary Markov policies by Π_0.

Given a policy $\pi \in \Pi_0$, we denote by $\mathbf{P}^{\pi}_{\nu_0}$ the induced probability measure on the canonical sample space $\Omega \triangleq (\mathcal{X} \times \mathcal{A})^{\infty}$, i.e., $\mathbf{P}^{\pi}_{\nu_0}[\cdot] = \text{Prob}[\cdot \mid \pi, x_0 \sim \nu_0]$. The expectation operator with respect to $\mathbf{P}^{\pi}_{\nu_0}$ is denoted by $\mathbf{E}^{\pi}_{\nu_0}$.

The optimal control problem is given by $\min_{\pi \in \Pi_0} \eta_c(\pi)$, where $\eta_c(\pi) \triangleq \mathbf{E}^{\pi}_{\nu_0}\left[\sum_{t=0}^{\infty} \gamma^t c(x_t, a_t)\right]$, is the total expected discounted cost of a policy π.

For every policy $\pi \in \Pi_0$, we define the *γ-discounted occupancy measure* $\mu^{\pi} : \mathcal{X} \times \mathcal{A} \to \mathbf{R}_+$, by $\mu^{\pi}(x, a) \triangleq \sum_{t=0}^{\infty} \gamma^t \mathbf{P}^{\pi}_{\nu_0}[x_t = x, a_t = a]$. The occupancy measure can be interpreted as the (unormalized) discounted visitation frequency of state-action pairs when acting according to policy π. Moreover, it holds that $\eta_c(\pi) = \sum_{x,a} \mu^{\pi}(x, a) c(s, a) = \mathbf{E}_{\mu^{\pi}}[c(x, a)]$.

3 Apprenticeship Learning Framework

3.1 Problem statement

Consider now the Markov decision model without a cost function, $\mathcal{M} \triangleq (\mathcal{X}, \mathcal{A}, P, \gamma, \nu_0)$. Assume that instead, we have access to a finite number m of i.i.d sample trajectories $\{(x_0^k, a_0^k, x_1^k, a_1^k, \ldots, x_t^k, a_t^k, \ldots)\}_{k=1}^{m}$ coming from an expert policy π_E. Note that the expert policy could also be history dependent. We impose the following assumptions:

Assumption 1 (Apprenticeship learning)**.**

(A1) *π_E is a nearly optimal policy for the discounted MDP corresponding to the model $\mathcal{M}^{c_{\text{true}}} = (\mathcal{X}, \mathcal{A}, P, \gamma, \nu_0, c_{\text{true}})$;*

(A2) *$c_{\text{true}} \in \mathcal{C}_{\text{lin}} = \{\sum_{i=1}^{n_c} w_i \psi_i \mid \|w\|_{\infty} \leq 1\}$, where $\{\psi_i\}_{i=1}^{n_c} \subset \mathbf{R}^{|\mathcal{X}||\mathcal{A}|}$ are fixed basis vectors, such that $\|\psi_i\|_{\infty} \leq 1$ for all $i = 1, \ldots, n_c$.*

The goal of apprenticeship learning is to find a policy π, such that $\langle \mu^{\pi}, c_{\text{true}} \rangle \leq \langle \mu^{\pi_E}, c_{\text{true}} \rangle$. Since the cost function c_{true} is unknown, AL algorithms search for a policy π that satisfies $\langle \mu^{\pi}, c \rangle \leq \langle \mu^{\pi_E}, c \rangle$, for all $c \in \mathcal{C}_{\text{lin}}$.

Therefore, an AL algorithm seeks a policy that performs better than the expert across $\mathcal{C}_{\text{lin}}$, by optimizing the objective

$$\min_{\pi \in \Pi} \sup_{c \in \mathcal{C}_{\text{lin}}} \left(\langle \mu^{\pi}, c \rangle - \langle \mu^{\pi_E}, c \rangle \right). \tag{1}$$

We highlight that one can consider other linearly parameterized cost classes, e.g., $\mathcal{C}_{\text{lin},2} = \{\sum_{i=1}^{n_c} w_i \psi_i \mid \|w\|_2 \leq 1\}$ [4], or $\mathcal{C}_{\text{convex}} = \{\sum_{i=1}^{n_c} w_i \psi_i \mid w_i \geq 0, \ \sum_{i=1}^{n_c} w_i = 1\}$ [22, 23]. The resoning and the analysis are similar.

3.2 The convex optimization view

In the remainder of the paper we will use the following vector notation borrowed from [1]. The transition law is a matrix $P \in \mathbf{R}_+^{|\mathcal{X}||\mathcal{A}| \times |\mathcal{X}|}$ so that $\sum_{x' \in \mathcal{X}} P_{(x,a),x'} = 1$, the initial probability distribution is a vector $\nu_0 \in \mathbf{R}_+^{|\mathcal{X}|}$ so that $\|\nu_0\|_1 = 1$, and the cost function is a vector $c \in \mathbf{R}^{|\mathcal{X}||\mathcal{A}|}$. Finally, for a stationary Markov policy $\pi \in \Pi_0$ we define the matrix $M^{\pi} \in \mathbf{R}_+^{|\mathcal{X}| \times |\mathcal{X}||\mathcal{A}|}$ that encodes π as $M^{\pi}_{x_i,(x_j,a_k)} = \pi(a_k \mid x_i)$, if $i = j$, and $M^{\pi}_{x_i,(x_j,a_k)} = 0$ otherwise.

Next, we will characterize the set of occupancy measures in terms of linear constraint satisfaction. To this aim let $\mathcal{F} \triangleq \left\{\mu \in \mathbf{R}^{|\mathcal{X}||\mathcal{A}|} \mid (B - \gamma P)^T \mu = \nu_0, \ \mu \geq 0\right\}$, where $B \in \{0, 1\}^{|\mathcal{X}||\mathcal{A}| \times |\mathcal{X}|}$ is a

binary matrix defined by $B_{(x_i,a_k),x_j} = 1$, if $i = j$, and $B_{(x_i,a_k),x_j} = 0$ otherwise. The constraints that define the set $\mathcal{F}$ are also known as *Bellman flow constraints*.

Proposition 1 ([23, Theorem 2])**.** *It holds that, $\mathcal{F} = \{\mu^\pi \mid \pi \in \Pi_0\}$. Indeed, for every $\pi \in \Pi_0$, we have that $\mu^\pi \in \mathcal{F}$. Moreover, for every feasible solution $\mu \in \mathcal{F}$, we can obtain a stationary Markov policy $\pi_\mu \in \Pi_0$ by $\pi_\mu(a|x) \triangleq \frac{\mu(x,a)}{\sum_{a' \in \mathcal{A}} \mu(x,a')}$. Then, the corresponding induced occupancy measure μ^{π_μ} satisfies $\mu^{\pi_\mu} = \mu$.*

Let $\mathbf{\Psi} \triangleq [\psi_1 | \dots | \psi_{n_c}] \in \mathbf{R}^{|\mathcal{X}||\mathcal{A}| \times n_c}$ be the cost basis matrix. For a policy $\pi \in \Pi$, we define its *feature expectation vector* as

$$\langle \mu^\pi, \mathbf{\Psi} \rangle \triangleq \mathbf{\Psi}^T \mu^\pi \in \mathbf{R}^{n_c}.$$

In other words, for every $i = 1, \dots, n_c$,

$$\langle \mu^\pi, \mathbf{\Psi} \rangle_i = \langle \mu^\pi, \psi_i \rangle = \eta_{\psi_i}(\pi).$$

Lemma 1. *For every π, the following holds*

$$\sup_{c \in \mathcal{C}_{\text{lin}}} \left(\langle \mu^\pi, c \rangle - \langle \mu^{\pi_E}, c \rangle \right) = \| \langle \mu^\pi, \mathbf{\Psi} \rangle - \langle \mu^{\pi_E}, \mathbf{\Psi} \rangle \|_1.$$

By Lemma 1, we get that (1) is equivalent to

$$\min_{\pi \in \Pi} \| \langle \mu^\pi, \mathbf{\Psi} \rangle - \langle \mu^{\pi_E}, \mathbf{\Psi} \rangle \|_1. \tag{2}$$

Note that although the objective function in (2) is convex in μ^π, the whole program is nonconvex in π. However, combining Proposition 1 with the fact that $\langle \mu^\pi, c \rangle = \eta_c(\pi)$ for every policy π, and every cost c, we conclude that the apprenticeship learning objective (2) can be stated equivalently as a convex optimization program:

$$\min_{\mu \in \mathcal{F}} \| \langle \mu, \mathbf{\Psi} \rangle - \langle \mu^{\pi_E}, \mathbf{\Psi} \rangle \|_1. \tag{3}$$

Note that the $|\mathcal{X}||\mathcal{A}|$ linear constraints given by $\mu \geq 0$ ensure that μ is a nonnegative measure, while the $|\mathcal{X}|$ linear constraints given by $(B - \gamma P)^T \mu = \nu_0$ ensure that μ is an occupancy measure generated by a stationary Markov policy.

4 Algorithm and main result

We have to take into account that in practice we do not have access to the whole policy π_E, but instead can observe i.i.d. trajectory samples distributed according to $\mathbf{P}_{\nu_0}^{\pi_E}$. For a multi-sample $\{(x_0^k, a_0^k, x_1^k, a_1^k, \dots, x_t^k, a_t^k, \dots)\}_{k=1}^m \sim (\mathbf{P}_{\nu_0}^{\pi_E})^m$ consider the Monte Carlo approximation $\langle \widehat{\mu^{\pi_E}, \mathbf{\Psi}} \rangle \in \mathbf{R}^{n_c}$ of the expert feature expectation vector, *i.e.* for each $i = 1, \dots, n_c$,

$$\langle \widehat{\mu^{\pi_E}, \mathbf{\Psi}} \rangle_i = \langle \widehat{\mu^{\pi_E}, \psi_i} \rangle \triangleq \frac{1}{m} \sum_{t=0}^{\infty} \sum_{j=1}^{m} \gamma^t \psi_i(x_t^j, a_t^j).$$

Moreover, under Assumption (A2), the following is a pointwise bound on Ω^m:

$$\| \langle \widehat{\mu^{\pi_E}, \mathbf{\Psi}} \rangle \|_\infty \leq 1/(1-\gamma). \tag{4}$$

We are interested in optimizing the empirical convex objective (5) for large-scale MDPs.

$$\min_{\mu \in \mathcal{F}} \| \langle \mu, \mathbf{\Psi} \rangle - \langle \widehat{\mu^{\pi_E}, \mathbf{\Psi}} \rangle \|_1, \tag{5}$$

which is a random convex program on $(\Omega^m, (\mathbf{P}_{\nu_0}^{\pi_E})^m)$.

Our main aim is (i) to provide a computationally efficient algorithm whose complexity does not grow with the size of the state and action spaces, and (ii) to obtain explicit probabilistic performance bounds on the quality of the extracted solution. To this end, we will design and analyze the apprenticeship learning analogue of the algorithm proposed in [1] for the forward average-cost MDP problem. Most of the tools from the forward MDP setting [1, 2] can be used for the AL formulation with the

appropriate modifications. We will present the main reasoning and results in this section with proofs presented in the appendix.

As the first step, instead of optimizing over the whole space $\mathbf{R}^{|\mathcal{X}||\mathcal{A}|}$, we optimize over the linear hull of a small number of selected feature vectors $\{\phi_i\}_{i=1}^d \subset \mathbf{R}^{|\mathcal{X}||\mathcal{A}|}$. In this way, we reduce significantly the number of optimization variables. Let $\mathbf{\Phi} \triangleq [\phi_1 | \dots | \phi_d] \in \mathbf{R}^{|\mathcal{X}||\mathcal{A}| \times d}$ be the feature matrix. The corresponding reduced program is

$$\min_{\theta \in \mathbf{R}^d : \mathbf{\Phi}\theta \in \mathcal{F}} \|\mathbf{\Psi}^T \mathbf{\Phi}\theta - \langle \widehat{\mu^{\pi_E}, \mathbf{\Psi}} \rangle\|_1. \tag{6}$$

Note however that for an arbitrary vector $u \in \mathbf{R}^{|\mathcal{X}||\mathcal{A}|}$ which is not necessarily in $\mathcal{F}$, we can still define a policy $\pi_u \in \mathbf{R}^{|\mathcal{X}||\mathcal{A}|}$ by $\pi_u(a|x) = \frac{[u(x,a)]_+}{\sum_{a' \in \mathcal{A}} [u(x,a')]_+}$. If $u(x,a) \leq 0$, for all $a \in \mathcal{A}$, we let $\pi_u(\cdot \mid x)$ be the uniform distribution [**?**].

Then, one has that $\mu^{\pi_u} = u$ if and only if $u \in \mathcal{F}$. In the general case, the following lemma, which is the discounted cost analogue of [1, Lem. 2], quantifies how close the generated occupancy measure μ^{π_u} is to u according to the degree of constraint violation.

Lemma 2. *For any $u \in \mathbf{R}^{|\mathcal{X}||\mathcal{A}|}$, it holds that $\|\mu^{\pi_u} - u\|_1 \leq \frac{2\|[u]_-\|_1 + \|(B - \gamma P)^T u - \nu_0\|_1}{1-\gamma}$.*

For any $\theta \in \mathbf{R}^d$, we define $\pi_\theta \triangleq \pi_{\mathbf{\Phi}\theta}$ and $\mu_\theta \triangleq \mu^{\pi_{\mathbf{\Phi}\theta}}$. As already discussed, $\mu_\theta = \mathbf{\Phi}\theta$ if and only if θ is feasible for (6). In the general case, one can bound the distance between the occupancy measure μ_θ and the vector $\mathbf{\Phi}\theta$ by applying Lemma 2.

Let $(\rho, \lambda) > 0$ be positive constants, $\Theta \triangleq \{\theta \in \mathbf{R}^d \mid \|\theta\|_2 \leq \rho\}$, and $\Pi_\Theta : \mathbf{R}^d \mapsto \Theta$ the Euclidean projection onto Θ. Consider the following surrogate loss function which is obtained by adding a positive multiple of the constraint violations to the initial objective function:

$$\begin{aligned}
\mathcal{L}(\theta) &\triangleq \|\mathbf{\Psi}^T \mathbf{\Phi}\theta - \langle \widehat{\mu^{\pi_E}, \mathbf{\Psi}} \rangle\|_1 + \lambda \underbrace{\|[\mathbf{\Phi}\theta]_-\|_1}_{:= V_1(\theta)} + \lambda \underbrace{\|(B - \gamma P)^T (\mathbf{\Phi}\theta) - \nu_0\|_1}_{:= V_2(\theta)} \\
&= \sum_{i=1}^{n_c} |\mathbf{\Psi}_{:,i}^T \mathbf{\Phi}\theta - \langle \widehat{\mu^{\pi_E}, \psi_i} \rangle| + \lambda \sum_{(x,a) \in \mathcal{X} \times \mathcal{A}} [\mathbf{\Phi}_{(x,a),:}\theta]_- + \lambda \sum_{x \in \mathcal{X}} |(B - \gamma P)_{:,x}^T \mathbf{\Phi}\theta - \nu_0(x)|.
\end{aligned}$$

We are interested in the reduced unconstrained convex optimization program of the form $\min_{\theta \in \Theta} \mathcal{L}(\theta)$.

A subgradient of $\mathcal{L}$ at θ is given by

$$\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) = &\sum_{i=1}^{n_c} (\mathbf{\Phi}^T \mathbf{\Psi}_{:,i}) \operatorname{sign}\left(\mathbf{\Psi}_{:,i}^T \mathbf{\Phi}\theta - \langle \widehat{\mu^{\pi_E}, \psi_i} \rangle\right) \\
&+ \lambda \sum_{x \in \mathcal{X}} (\mathbf{\Phi}^T (B - \gamma P)_{:,x}) \operatorname{sign}\left((B - \gamma P)_{:,x}^T \mathbf{\Phi}\theta - \nu_0(x)\right) \\
&- \lambda \sum_{(x,a) \in \mathcal{X} \times \mathcal{A}} (\mathbf{\Phi}_{(x,a),:}^T) \mathbf{1}_{\{\mathbf{\Phi}_{(x,a),:}\theta < 0\}}.
\end{aligned}$$

Suppose that $q_1 \in \mathcal{P}(\mathcal{X} \times \mathcal{A})$ and $q_2 \in \mathcal{P}(\mathcal{X})$ assign to each element a strictly positive probability. We propose a method for apprenticeship learning shown in Algorithm 1. It uses an unbiased estimate of $\nabla_\theta \mathcal{L}(\theta)$ for fixed expert trajectory samples, *i.e.*

Algorithm 1 Stochastic subgradient descent for apprenticeship learning (SGD-AL).

given cost matrix $\mathbf{\Psi}$, feature matrix $\mathbf{\Phi}$, number of expert samples m, number of iterations T, learning rate $\eta > 0$, radius $\rho > 0$, regularization parameter $\lambda > 0$
Set $\theta_0 = 0$
Sample $\{(x_0^k, a_0^k, x_1^k, a_1^k, \ldots, x_t^k, a_t^k, \ldots)\}_{k=1}^m \sim (\mathbf{P}_{\nu_0}^{\pi_E})^m$
for $t = 1, \ldots, T$ **do**
  Sample $(x^{(t)}, a^{(t)}) \sim q_1$ and $y^{(t)} \sim q_2$
  Compute $g_t(\theta_t)$ via (7)
  Update $\theta_{t+1} = \Pi_\Theta\big(\theta_t - \eta g_t(\theta_t)\big)$
$\hat{\theta}_T = \frac{1}{T}\sum_{t=1}^T \theta_t$
Return $\pi_{\hat{\theta}_T}$

$$
\begin{aligned}
g_t(\theta) = &\sum_{i=1}^{n_c} (\mathbf{\Phi}^T \mathbf{\Psi}_{:,i}) \operatorname{sign}\left(\mathbf{\Psi}_{:,i}^T \mathbf{\Phi}\theta - \widehat{\langle \mu^{\pi_E}, \psi_i \rangle}\right) \\
&+ \lambda \frac{\mathbf{\Phi}^T (B - \gamma P)_{:,y^{(t)}}}{q_2(y^{(t)})} \operatorname{sign}\left((B - \gamma P)_{:,y^{(t)}}^T \mathbf{\Phi}\theta - \nu_0(y^{(t)})\right) \\
&- \lambda \frac{\mathbf{\Phi}_{(x^{(t)},a^{(t)}),:}^T}{q_1(x^{(t)}, a^{(t)})} \mathbf{1}_{\{\mathbf{\Phi}_{(x^{(t)},a^{(t)}),:}\theta < 0\}},
\end{aligned} \tag{7}
$$

where $(x^{(t)}, a^{(t)}) \sim q_1$ and $y^{(t)} \sim q_2$.

Regret bounds for Algorithm 1 can be obtained by using results from the stochastic convex optimization literature and statistical learning theory.

Assumption 2. *All entries of the feature matrix $\mathbf{\Phi}$ are positive,* i.e. *every feature vector ϕ_i is a measure which assigns a non-zero measure to each pair (x, a). Moreover, $\| \mathbf{\Phi} \|_1 = \frac{1}{1-\gamma}$.*

We define the following constants:

$$
\begin{aligned}
C_1 &\triangleq \max_{(x,a) \in \mathcal{X} \times \mathcal{A}} \frac{\|\mathbf{\Phi}_{(x,a),:}^T\|_2}{q_1(x, a)}, \\
C_2 &\triangleq \max_{x \in \mathcal{X}} \frac{\|\mathbf{\Phi}^T (B - \gamma P)_{:,x}\|_2}{q_2(x)}.
\end{aligned}
$$

These constants appear in our performance bounds. We would like to choose appropriate distributions so that C_1 and C_2 are small, since they appear in the error bound. We refer the reader to [1] for a thorough discussion on the choice of the distributions.

Observe from (7) that for all $\theta \in \Theta$ we have the following bound:

$$
\|g_t(\theta)\|_2 \leq \|\mathbf{\Phi}\|_2 \sum_{i=1}^{n_c} \|\psi_i\|_2 + \lambda(C_1 + C_2) =: K. \tag{8}
$$

Theorem 1. *Let $\varepsilon \in (0, 1)$, $\delta \in (0, 1)$, $\rho > 0$, $\lambda = 1/\varepsilon$, $m \geq \frac{32 n_c^2 \log(\frac{4 n_c}{\delta})}{(1-\gamma)\varepsilon^2}$, $T \geq \frac{4\rho^2}{\varepsilon^2}\left(\frac{2\|\mathbf{\Psi}\|_\infty}{\lambda(1-\gamma)} + 1\right)^2 \Delta^2$ with $\Delta \triangleq K + \sqrt{10 \log \frac{2}{\delta}} + \sqrt{5d \log(1 + \frac{\rho^2 T}{d})}$, and $\eta = \rho/(K\sqrt{T})$. Then, with probability at least $1 - \delta$, Algorithm 1 generates $\pi_{\hat{\theta}_T}$ so that for all $\theta \in \Theta$,*

$$
\begin{aligned}
\|\mathbf{\Psi}^T \mu_{\hat{\theta}_T} - \langle \mu^{\pi_E}, \mathbf{\Psi} \rangle\|_1 \leq \|\mathbf{\Psi}^T \mu_\theta - \langle \mu^{\pi_E}, \mathbf{\Psi} \rangle\|_1 + \left(\frac{4\|\mathbf{\Psi}\|_\infty}{1-\gamma} + \frac{1}{\varepsilon}\right)(V_1(\theta) + V_2(\theta)) \\
+ \left(\frac{2\|\mathbf{\Psi}\|_\infty}{1-\gamma}\right)\left(\|\mathbf{\Psi}\|_\infty \|\mathbf{\Phi}\|_1 \rho\sqrt{d} + \frac{n_c}{1-\gamma}\right)\varepsilon + \varepsilon.
\end{aligned}
$$

References

[1] Y. Abbasi-Yadkori, P. L. Bartlett, and A. Malek. Linear programming for large-scale Markov decision problems. In *International Conference on Machine Learning (ICML)*, pages 496–504, 2014.

[2] Yasin Abbasi-Yadkori, Peter L. Bartlett, Xiao Dong Chen, and Alan Malek. Large-scale markov decision problems via the linear programming dual. *ArXiv*, abs/1901.01992, 2019.

[3] P. Abbeel, D. Dolgov, A. Y. Ng, and S. Thrun. Apprenticeship learning for motion planning with application to parking lot navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1083–1090, 2008.

[4] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2004.

[5] V. S. Borkar. A convex analytic approach to Markov decision processes. *Probability Theory and Related Fields*, 78(4):583–602, 1988.

[6] F. Dufour and T. Prieto-Rumeau. Finite linear programming approximations of constrained discounted Markov decision processes. *SIAM Journal on Control and Optimization*, 51(2):1298–1324, 2013.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[8] O. Hernández-Lerma, J. González-Hernández, and R. López-Martínez. Constrained average cost Markov control processes in Borel spaces. *SIAM Journal on Control and Optimization*, 42(2):442–468, 2003.

[9] O. Hernández-Lerma and J. B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer-Verlag New York, 1996.

[10] O. Hernández-Lerma and J. B. Lasserre. *Further Topics on Discrete-Time Markov Control Processes*. Springer-Verlag New York, 1999.

[11] O. Hernández-Lerma and J. B. Lasserre. The linear programming approach. In *Handbook of Markov Decision Processes: Methods and Applications*, pages 377–407. Springer US, 2002.

[12] J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4565–4573, 2016.

[13] J. Ho, J. K. Gupta, and S. Ermon. Model-free imitation learning with policy optimization. In *International Conference on Machine Learning (ICML)*, pages 2760–2769, 2016.

[14] S. Levine, Z. Popović, and V. Koltun. Feature construction for inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1342–1350, 2010.

[15] S. Levine, Z. Popović, and V. Koltun. Nonlinear inverse reinforcement learning with Gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*, pages 19–27, 2011.

[16] G. Neu and C. Szepesvári. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 295–302, 2007.

[17] A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 663–670, 2000.

[18] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *International Conference on Machine Learning (ICML)*, 2006.

[19] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 627–635, 2011.

[20] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and M. Moritz. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, pages 1889–1897, 2015.

[21] E. Shafieepoorfard, M. Raginsky, and S. P. Meyn. Rational inattention in controlled Markov processes. In *American Control Conference (ACC)*, pages 6790–6797, 2013.

[22] U. Syed and R. E. Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1449–1456, 2007.

[23] Umar Syed, Michael Bowling, and Robert E. Schapire. Apprenticeship learning using linear programming. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1032–1039, New York, NY, USA, 2008. ACM.

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

[25] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *National Conference on Artificial Intelligence (AAAI)*, pages 1433–1438, 2008.

A Proofs

Proof of Lemma 1. Fix a $\pi \in \Pi$. Then, for every $c = \sum_{i=1}^{n_c} w_i \psi_i$ with $\|w\|_\infty \leq 1$, it holds that

$$\begin{aligned}\langle \mu^\pi, c\rangle - \langle \mu^{\pi_E}, c\rangle &= \sum_{i=1}^{n_c} w_i \left(\langle \mu^\pi, \psi_i\rangle - \langle \mu^{\pi_E}, \psi_i\rangle\right) \\ &\leq \|w\|_\infty \|\langle \mu^\pi, \mathbf{\Psi}\rangle - \langle \mu^{\pi_E}, \mathbf{\Psi}\rangle\|_1 \\ &\leq \|\langle \mu^\pi, \mathbf{\Psi}\rangle - \langle \mu^{\pi_E}, \mathbf{\Psi}\rangle\|_1,\end{aligned}$$

and thus,

$$\sup_{c \in \mathcal{C}_{\text{lin}}} \left(\langle \mu^\pi, c\rangle - \langle \mu^{\pi_E}, c\rangle\right) \leq \|\langle \mu^\pi, \mathbf{\Psi}\rangle - \langle \mu^{\pi_E}, \mathbf{\Psi}\rangle\|_1.$$

Next, let $\tilde{w} \in \mathbf{R}^{n_c}$ be defined by

$$\tilde{w}_i = \text{sign}\left(\langle \mu^\pi, \psi_i\rangle - \langle \mu^{\pi_E}, \psi_i\rangle\right).$$

Then for $\tilde{c} \triangleq \sum_{i=1}^{n_c} \tilde{w}_i \psi_i$, we have $\tilde{c} \in \mathcal{C}_{\text{lin}}$ and

$$\langle \mu^\pi, \tilde{c}\rangle - \langle \mu^{\pi_E}, \tilde{c}\rangle = \|\langle \mu^\pi, \mathbf{\Psi}\rangle - \langle \mu^{\pi_E}, \mathbf{\Psi}\rangle\|_1,$$

which proves that

$$\sup_{c \in \mathcal{C}_{\text{lin}}} \left(\langle \mu^\pi, c\rangle - \langle \mu^{\pi_E}, c\rangle\right) \geq \|\langle \mu^\pi, \mathbf{\Psi}\rangle - \langle \mu^{\pi_E}, \mathbf{\Psi}\rangle\|_1.$$

This concludes the proof. □

In case of a stationary Markov policy, the induced discounted occupancy measure has the following form.

Lemma 3. *Let $\pi \in \Pi_0$ be a stationary Markov policy. Then for all $x \in \mathcal{X}$, $a \in \mathcal{A}$ and $t \in \mathbf{N}_0$, it holds that $\mathbf{P}^\pi_{\nu_0}[x_t = x, a_t = a] = \left[\nu_0^T M^\pi (PM^\pi)^t\right]_{(x,a)}$. In particular, $(\mu^\pi)^T = \sum_{t=0}^\infty \gamma^t \nu_0^T M^\pi (PM^\pi)^t$.*

Proof. For $t = 0$ we have

$$\mathbf{P}^\pi_{\nu_0}[x_0 = x, a_0 = a] = \nu_0(x)\pi(a|x) = [\nu_0^\top M^\pi]_{(x,a)}.$$

Next, assume that the result holds for $t - 1$. Then,

$$\begin{aligned}\mathbf{P}^\pi_{\nu_0}[x_t = x, a_t = a] &= \mathbf{P}^\pi_{\nu_0}[a_t = a \mid x_t = x]\, \mathbf{P}^\pi_{\nu_0}[x_t = x] \\ &= \pi(a|x) \sum_{x' \in \mathcal{X}} \sum_{a' \in \mathcal{A}} P_{(x',a'),x}\, \mathbf{P}^\pi_{\nu_0}[x_{t-1} = x', a_{t-1} = a'].\end{aligned}$$

By the induction assumption, we conclude that

$$\begin{aligned}\mathbf{P}^\pi_{\nu_0}[x_t = x, a_t = a] &= \pi(a|x) \sum_{x' \in \mathcal{X}} \sum_{a' \in \mathcal{A}} \left[\nu_0^\top M^\pi (PM^\pi)^{t-1}\right]_{(x',a')} P_{(x',a'),x} \\ &= \pi(a|x) \left[\nu_0^\top M^\pi (PM^\pi)^{t-1} P\right]_x \\ &= \left[\nu_0^\top M^\pi (PM^\pi)^t\right]_{(x,a)}.\end{aligned}$$

□

Proof of Lemma 2. We provide a refined proof and bound similar to [2, Lemma 13]. We have

$$\begin{aligned}\|\underbrace{(B - \gamma P)^T [u]_+ - \nu_0}_{=:-w}\|_1 &\leq \|(B - \gamma P)^T [u]_-\|_1 + \|(B - \gamma P)^T u - \nu_0\|_1 \\ &\leq (1 + \gamma)\|[u]_-\|_1 + \|(B - \gamma P)^T u - \nu_0\|_1,\end{aligned}$$

where we have used the fact that $\|B^T\|_1 = \|B\|_\infty = 1$ and $\|P^T\|_1 = \|P\|_\infty = 1$. By virtue of Lemma 3, we have

$$\begin{aligned}
(\mu^{\pi_u})^T &= \sum_{t=0}^{\infty} \gamma^t \nu_0^T M^{\pi_u} (P M^{\pi_u})^t \\
&= \sum_{t=0}^{\infty} \gamma^t (w + (B - \gamma P)^T [u]_+)^T M^{\pi_u} (P M^{\pi_u})^t \\
&= \sum_{t=0}^{\infty} \gamma^t w^T M^{\pi_u} (P M^{\pi_u})^t + \sum_{t=0}^{\infty} \gamma^t [u]_+^T (P M^{\pi_u})^t - \sum_{t=0}^{\infty} \gamma^{t+1} [u]_+^T (P M^{\pi_u})^{t+1} \\
&= \sum_{t=0}^{\infty} \gamma^t w^T M^{\pi_u} (P M^{\pi_u})^t + [u]_+^T,
\end{aligned}$$

where in the third equality we used $[u]_+^T B M^{\pi_u} = [u]_+^T$. Therefore,

$$\begin{aligned}
\|\mu^{\pi_u} - [u]_+\|_1 &= \|\sum_{t=0}^{\infty} \gamma^t ((M^{\pi_u})^T P^T)^t (M^{\pi_u})^T w\|_1 \\
&\le \sum_{t=0}^{\infty} \gamma^t \|M^{\pi_u}\|_\infty^t \|P\|_\infty^t \|M^{\pi_u}\|_\infty \|w\|_1 \\
&\le \frac{1}{1-\gamma} \left((1+\gamma) \|[u]_-\|_1 + \|(B - \gamma P)^T u - \nu_0\|_1 \right),
\end{aligned}$$

where in the last inequality we used $\|M^{\pi_u}\|_\infty = \|B\|_\infty = \|P\|_\infty = 1$. Finally, the triangle inequality gives

$$\|\mu^{\pi_u} - u\|_1 \le \|\mu^{\pi_u} - [u]_+\|_1 + \|[u]_-\|_1 \le \frac{2\|[u]_-\|_1 + \|(B - \gamma P)^T u - \nu_0\|_1}{1-\gamma}. \qquad \square$$

Proof of Theorem 1. The proof combines techniques presented in the proofs of [1, Theorem 2] and [2, Lemma 14] and the Hoeffding's bound.

We first fix an expert trajectory multi-sample $\{(x_0^k, a_0^k, x_1^k, a_1^k, \ldots, x_t^k, a_t^k, \ldots)\}_{k=1}^m \sim (\mathbf{P}_{\nu_0}^{\pi_E})^m$. Then, by virtue of [1, Thm. 3] and by the uniform bound of the unbiased subgradient estimates (8), we get that if the learning rate is $\eta = \rho/(K\sqrt{T})$, then with probability at least $1 - \delta/2$ (the corresponding probability space is $((\mathcal{X} \times \mathcal{A})^T \times \mathcal{X}^T), q_1^T \otimes q_2^T$),

$$\mathcal{L}(\hat{\theta}_T) - \min_{\theta \in \Theta} \mathcal{L}(\theta) \le \frac{\rho K}{\sqrt{T}} + \sqrt{\frac{1 + 4\rho^2 T}{T^2} \left(2 \log \frac{2}{\delta} + d \log \left(1 + \frac{\rho^2 T}{d} \right) \right)}. \tag{9}$$

Integrating over the whole probability space $(\Omega^m, (\mathbf{P}_{\nu_0}^{\pi_E})^m)$, we conclude that (9) holds with probability at least $1 - \delta/2$, where the corresponding probability space is $(\Omega^m \times (\mathcal{X} \times \mathcal{A})^T \times \mathcal{X}^T), (\mathbf{P}_{\nu_0}^{\pi_E})^m \otimes q_1^T \otimes q_2^T)$.

Substituting $\mathcal{L}(\hat{\theta}_T)$ and $\mathcal{L}(\theta)$ by their definitions, and using the inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we obtain that with probability at least $1 - \delta/2$, for all $\theta \in \Theta$,

$$\|\mathbf{\Psi}^T \mathbf{\Phi} \hat{\theta}_T - \widehat{\langle \mu^{\pi_E}, \mathbf{\Psi} \rangle}\|_1 + \lambda V_1(\hat{\theta}_T) + \lambda V_2(\hat{\theta}_T) \le \|\mathbf{\Psi}^T \mathbf{\Phi} \theta - \widehat{\langle \mu^{\pi_E}, \mathbf{\Psi} \rangle}\|_1 + \lambda V_1(\theta) + \lambda V_2(\theta) + \frac{\rho}{\sqrt{T}} \Delta. \tag{10}$$

For all multisamples that (10) holds, and for all $\theta \in \Theta$,

$$\begin{aligned}\|\mathbf{\Psi}^T\mu_{\hat{\theta}_T} - \widehat{\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle}\|_1 &\leq \|\mathbf{\Psi}^T\mu_{\hat{\theta}_T} - \mathbf{\Psi}^T\mathbf{\Phi}\hat{\theta}_T\|_1 + \|\mathbf{\Psi}^T\mathbf{\Phi}\hat{\theta}_T - \widehat{\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle}\|_1 \\ &\leq \|\mathbf{\Psi}\|_\infty \frac{2V_1(\hat{\theta}_T) + V_2(\hat{\theta}_T)}{1-\gamma} + \|\mathbf{\Psi}^T\mathbf{\Phi}\theta - \widehat{\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle}\|_1 \\ &\quad + \lambda V_1(\theta) + \lambda V_2(\theta) + \frac{\rho}{\sqrt{T}}\Delta, \end{aligned} \tag{11}$$

where we used the triangle inequality in the first step, and Lemma 2 together with the bound (10) in the second one. Moreover by (10),

$$V_1(\hat{\theta}_T) + V_2(\hat{\theta}_T) \leq \frac{1}{\lambda}\left(\|\mathbf{\Psi}\|_\infty\|\mathbf{\Phi}\|_1\rho\sqrt{d} + \frac{n_c}{1-\gamma}\right) + V_1(\theta) + V_2(\theta) + \frac{\rho}{\lambda\sqrt{T}}\Delta, \tag{12}$$

where we used that $\|\,\theta\,\|_1 \leq \rho\sqrt{d}$, and the pointwise bound $\|\widehat{\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle}\|_1 \leq n_c\,\|\widehat{\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle}\|_\infty \leq n_c/(1-\gamma)$.

Once more, by the triangle inequality and Lemma 2, we get

$$\begin{aligned}\|\mathbf{\Psi}^T\mathbf{\Phi}\theta - \widehat{\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle}\|_1 &\leq \|\mathbf{\Psi}^T\mathbf{\Phi}\theta - \mathbf{\Psi}^T\mu_\theta\|_1 + \|\mathbf{\Psi}^T\mu_\theta - \widehat{\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle}\|_1 \\ &\leq \|\mathbf{\Psi}\|_\infty \frac{2V_1(\theta) + V_2(\theta)}{1-\gamma} + \|\mathbf{\Psi}^T\mu_\theta - \widehat{\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle}\|_1. \end{aligned} \tag{13}$$

Therefore, by combining (11),(12) and (13), we get that we probability at least $1-\delta/2$, , for all $\theta \in \Theta$,

$$\begin{aligned}\|\mathbf{\Psi}^T\mu_{\hat{\theta}_T} - \widehat{\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle}\|_1 \leq \|\mathbf{\Psi}^T\mu_\theta - \widehat{\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle}\|_1 + \frac{2\|\mathbf{\Psi}\|_\infty}{\lambda(1-\gamma)}\left(\|\mathbf{\Psi}\|_\infty\|\mathbf{\Phi}\|_1\rho\sqrt{d} + \frac{n_c}{1-\gamma}\right) \\ + \left(\frac{4\|\mathbf{\Psi}\|_\infty}{1-\gamma} + \lambda\right)(V_1(\theta) + V_2(\theta)) + \left(\frac{2\|\mathbf{\Psi}\|_\infty}{\lambda(1-\gamma)} + 1\right)\frac{\rho}{\sqrt{T}}\Delta.\end{aligned}$$

For $T \geq \frac{4\rho^2}{\varepsilon^2}\left(\frac{2\|\mathbf{\Psi}\|_\infty}{\lambda(1-\gamma)} + 1\right)^2\Delta^2$ and $\lambda = 1/\varepsilon$, it follows that with probability at least $1-\delta/2$, for all $\theta \in \Theta$,

$$\begin{aligned}\|\mathbf{\Psi}^T\mu_{\hat{\theta}_T} - \widehat{\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle}\|_1 \leq \|\mathbf{\Psi}^T\mu_\theta - \widehat{\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle}\|_1 + \left(\frac{4\|\mathbf{\Psi}\|_\infty}{1-\gamma} + \frac{1}{\varepsilon}\right)(V_1(\theta) + V_2(\theta)) \\ + \left(\frac{2\|\mathbf{\Psi}\|_\infty}{1-\gamma}\right)\left(\|\mathbf{\Psi}\|_\infty\|\mathbf{\Phi}\|_1\rho\sqrt{d} + \frac{n_c}{1-\gamma}\right)\varepsilon + \varepsilon/2.\end{aligned}$$

We conclude the proof by using Hoeffding's inequality with confidence $\delta/(2n_c)$ and approximation accuracy $\varepsilon/(4n_c)$. In particular, we have that for $m \geq \frac{32n_c\log(4n_c/\delta)}{2(1-\gamma)\varepsilon^2}$ and for all $i = 1, \ldots, n_c$,

$$\left|\langle\mu^{\pi_E}, \psi_i\rangle - \widehat{\langle\mu^{\pi_E}, \psi_i\rangle}\right| \leq \varepsilon/(4n_c),$$

with probability $(\mathbf{P}_{\nu_0}^{\pi_E})^m$ at least $(1-\delta/(2n_c))$. Note that under Assumption (A2), it holds that $\sum_{t=0}^{\infty}\gamma^t\psi_i(x_t, a_t) \leq 1/(1-\gamma)$ for all $(x_t, a_t) \in \mathcal{X}\times\mathcal{A}$ and for all $i = 1, \ldots, n_c$.

A union bound gives that for $m \geq \frac{32n_c\log(\frac{4n_c}{\delta})}{2(1-\gamma)\varepsilon^2}$,

$$\|\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle - \widehat{\langle\mu^{\pi_E}, \mathbf{\Psi}\rangle}\|_1 \leq \varepsilon/4,$$

with probability $(\mathbf{P}_{\nu_0}^{\pi_E})^m$ at least $(1-\delta/2)$. Integrating over the whole space $\left((\mathcal{X}\times\mathcal{A})^T \times \mathcal{X}^T, q_1^T \otimes q_2^T\right)$ we have the same statement with probability $(\mathbf{P}_{\nu_0}^{\pi_E})^m \otimes q_1^T \otimes q_2^T$ at least $(1-\delta/2)$.

Finally, a simple union bound concludes the proof. □