Inverse Reinforcement Learning in a Continuous State Space with Formal Guarantees

Gregory Dexter
Department of Computer Science
Purdue University
West Lafayette, IN, USA
gdexter@purdue.edu

Kevin Bello
Department of Computer Science
Purdue University
West Lafayette, IN, USA
kbellome@purdue.edu

Jean Honorio
Department of Computer Science
Purdue University
West Lafayette, IN, USA
jhonorio@purdue.edu

Abstract

Inverse Reinforcement Learning (IRL) is the problem of finding a reward function which describes observed/known expert behavior. The IRL setting is remarkably useful for automated control, in situations where the reward function is difficult to specify manually or as a means to extract agent preference. In this work, we provide a new IRL algorithm for the *continuous* state space setting with *unknown* transition dynamics by modeling the system using a basis of *orthonormal functions*. Moreover, we provide a proof of correctness and formal guarantees on the sample and time complexity of our algorithm. Finally, we present synthetic experiments to corroborate our theoretical guarantees.

1 Introduction

Reinforcement learning in the context of a Markov Decision Process (MDP) aims to produce a policy function such that the discounted accumulated reward of an agent following the given policy is optimal. One difficulty in applying this framework to real world problems is that for many tasks of interest, there is no obvious reward function to maximize. For example, a car merging onto a highway must account for many different criteria, e.g., avoiding crashing and maintaining adequate speed, and there is not an obvious formula of these variables producing a reward which reflects expert behavior [3]. The Inverse Reinforcement Learning (IRL) problem consists of having as input an MDP with no reward function along with a policy, and the goal is to find a reward function so that the MDP plus the learned reward function is optimally solved by the given policy. Describing expert behavior in terms of a reward function has been shown to provide a succinct, robust, and interpretable representation of expert behavior for purposes of reinforcement learning [3]. Furthermore, this representation of agent preference can give insight on the preferences or goals of an agent.

Despite much work over the years, there still exist major limitations in our understanding of IRL. One major gap is the lack of algorithms for the setting of continuous state space. This setting is of primary importance since many promising applications, such as autonomous vehicles and robotics, occur in a continuous setting. Existing algorithms for IRL in the continuous setting generally do not directly solve the IRL problem, but instead take a policy matching approach which leads to a less robust description of agent preference [3]. Finally, there is a lack of formal guarantees for IRL methods which make principled understanding of the problem difficult.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

For a comprehensive list of results on the IRL problem, we refer the reader to [3] as we next describe works that are closest to ours. Throughout the years, various approaches have been proposed to solve the IRL problem in the continuous state setting. Aghasadeghi and Bretl [1] used a finite set of parametric basis functions to model the policy and reward function for a policy matching approach. Boularias et al. [6] matched the observed policy while maximizing Relative Entropy and applied this method to continuous tasks through discretization. Levine and Koltun [12] maximized the likelihood of the observed policy while representing the reward function as a Gaussian Process. Finn et al. [8] modeled the reward function through a neural network to allow for more expressive rewards. Ramponi et al. [18] and Pirotta and Restelli [17] search for an optimal reward based on minimizing the gradient of the continuous policy. Metelli et al. [14] selects a reward function over a set of basis functions such that the policy gradient is zero. These previous methods rely on heuristic arguments and empirical results to support each approach, but do not provide any theoretical guarantees or the insight that comes with them. Finally, Komanduru and Honorio [11] and Metelli et al. [15] provided theoretical analysis similar to our work, but only for the finite state space setting.

Contributions. In contrast to the work above, this paper provides a *formally guaranteed solution* to the IRL problem over a continuous state space when transition dynamics of the MDP are *unknown*. We accomplish this by representing the estimated transition dynamics as an infinite sum of basis functions from $\mathbb{R}^2 \to \mathbb{R}$. Then, under natural conditions on the representations, we prove that a linear program can recover a correct reward function. We develop a model of IRL in the continuous setting using the idea of infinite matrices, which in contrast to the more typical non-parametric functional analysis approach, allow us to derive *"explicit, computable, and meaningful error bounds"* [20]. We exhibit our main theorem by proving that if the transition functions of the MDP have appropriately bounded partial derivatives, then by representing the transition functions over a trigonometric basis, our algorithm returns a correct reward function with high probability. However, we emphasize that our results apply under a range of appropriate technical conditions and orthonormal bases. To the best of our knowledge, we present the first algorithm for IRL in the continuous setting with formal guarantees on correctness, sample complexity, and computational complexity. The total sample and computational complexity of our algorithm depends on the series representations of the transition dynamics over a given basis. However, if a given set of transition functions can be represented with sufficiently small error over a given basis by a $k \times k$ matrix of coefficients, then our algorithm (see Algorithm 2) returns a solution to the IRL problem with probability at least $1 - \delta$ in $\mathcal{O}(k^2 \log \frac{k}{\delta})$ samples by solving a linear program with $\mathcal{O}(k)$ variables and constraints. We show how the magnitude of k needed, i.e., the size of the representation, depends on characteristics of a given IRL problem. This sample complexity matches that of the discrete case algorithm of [11]. For more detailed complexity results, see Section 4. We experimentally validate our results by testing our IRL algorithm on a problem with random polynomial transition functions. Finally, we show how our algorithm can be efficiently extended to the d-dimensional setting.

2 Preliminaries

In this section, we first define the Markov Decision Process (MDP) formalizing the IRL problem. We briefly introduce the Bellman Optimality Equation, which gives necessary and sufficient conditions for a solution to the IRL problem. We then describe the problem of estimating the transition dynamics of the MDP. Finally, we show how to represent the Bellman Optimality Criteria for the given MDP in terms of infinite-matrices and infinite-vectors. For now, we restrict our attention to a one dimensional continuous state space.

Definition 1 *A Markov Decision Process (MDP) is a 5-tuple* $(S, A, \{P_a(s'|s)\}_{a \in A}, \gamma, R)$ *where*

- $S = [-1, 1]$ *is an uncountable set of states*[1]*,*
- A *is a finite set of actions ,*
- $\{P_a(s'|s)\}_{a \in A}$ *is a set of probability density functions such that* $P_a(s'|s) : S \times S \to \mathbb{R}^+$*.* $P_a(s'|s)$ *represents the transition probability density function from state* s *to state* s' *when taking action* a*,*

[1]While this paper considers $S = [-1, 1]$, S can easily be generalize to any interval.

- $\gamma \in [0, 1)$ *is the discount factor,*
- $R : S \to \mathbb{R}$ *is the reward function.*

We consider the inverse reinforcement problem. That is, given an MDP without the reward function (MDP $\setminus R$) along with a policy $\pi : S \to A$, our goal is to determine a reward function R such that the policy π is optimal for the MDP under the learned reward function R.

Bellman Optimality Criteria. First we define the value function and Q-function

$$V^\pi(s_0) = \mathbb{E}[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + ...|\pi], \tag{1}$$

$$Q^\pi(s, a) = R(s) + \gamma \mathbb{E}[V^\pi(s')], \quad s' \sim P_a(s'|s). \tag{2}$$

Using these terms, we present the *Bellman Optimality Equation* which gives necessary and sufficient conditions for policy π to be optimal for a given MDP [21].

$$\pi(s) \in \arg\max_{a \in A} Q^\pi(s, a), \;\; \forall\, s \in S. \tag{3}$$

Estimation Problem. We impose the additional task of estimating the transition dynamics of the MDP. While the policy is assumed to be exactly known, the transition probability density functions $\{P_a(s'|s)\}_{a \in A}$ are unknown. Instead, we assume that for an arbitrary state s, we can sample the one-step transition $s' \sim P_a(\cdot|s)$, as previously done by Komanduru and Honorio [11].

Infinite Matrix Representation. We now introduce a natural assumption on the transition probability density functions, which will make further analysis possible.

Let $\{\phi_n(s)\}_{n \in \mathbb{N}}$ be a countably infinite set of orthonormal basis functions over S. That is

$$\int_{-1}^{1} \phi_n^2(s) ds = 1, \text{ and } \int_{-1}^{1} \phi_i(s)\phi_j(s) ds = 0, \; \forall\, i \neq j.$$

Examples of bases that fulfill the above conditions are the Legendre polynomials, the Chebyshev polynomials, and an appropriately scaled trigonometric basis.

We assume that the transition probability density functions $\{P_a(s'|s)\}_{a \in A}$ can be represented over the given basis by,

$$P_a(s'|s) = \sum_{i,j=1}^{\infty} \phi_i(s') \mathcal{Z}_{ij}^{(a)} \phi_j(s), \;\; \mathcal{Z}_{ij}^{(a)} \in \mathbb{R}. \tag{4}$$

In Section 5, we demonstrate sufficient conditions on the transition functions to guarantee that our algorithm is correct when using a trigonometric basis as an example. The following proposition justifies considering only reward functions which are a linear combination of $\{\phi_n(s)\}_{n \in \mathbb{N}}$.

Proposition 1 *If the transition functions can be represented over $\{\phi_n(s)\}_{n \in \mathbb{N}}$ as in Equation 4 and there is a at least one reward function such that the policy is optimal, then there is at least one reward function R equal to a linear combination of $\{\phi_n(s)\}_{n \in \mathbb{N}}$ such that the Bellman Optimally criterion is fulfilled for the given policy.*

(Missing proofs can be found in the Appendix A.)

This justifies considering only reward functions represented by a series such as,

$$R(s) = \sum_{k=1}^{\infty} \alpha_k \phi_k(s), \quad \alpha_k \in \mathbb{R}. \tag{5}$$

The above conditions are equivalent to requiring that the transition function is a linear operator on the function space spanned by the Schauder basis $\{\phi_n(s)\}$. Proposition 1 guarantees that if there exists a reward which optimally solves the IRL problem, then there is an optimal reward in the closure of the function space spanned by $\{\phi_n(s)\}_{n \in \mathbb{N}}$.

It will be useful to identify the coefficients $\mathcal{Z}_{ij}^{(a)}$ and α_k in Equations 4 and 5 as entries of an infinite-matrix and infinite-vector respectively. This identification allows us to keep all the usual definitions of matrix-matrix multiplication, matrix-vector multiplication, matrix transposition, matrix powers, as well as addition and subtraction.

We define additional simplifying notation. Let the infinite matrix $\mathcal{Z}^{(a)}$ be defined as the matrix of coefficients associated with $P_a(s'|s)$. Let the infinite vector α correspond to coefficients of an arbitrary reward function R. We drop the superscript a and write $\mathcal{Z}$ where possible. Let $[^k\mathcal{Z}]$ represent the k-truncation of $\mathcal{Z}$. That is, $[^k\mathcal{Z}]_{ij} = \mathcal{Z}_{ij}$ for all $i, j \leq k$, and $[^k\mathcal{Z}]_{ij} = 0$ otherwise. Define the complement of the k-truncation of $\mathcal{Z}$ as $[^\infty_k\mathcal{Z}] = \mathcal{Z} - [^k\mathcal{Z}]$. Let $\phi(s)$ be an infinite-vector where the n-th entry equals $\phi_n(s)$. Note that in contrast to finite matrices, typical objects, such as matrix products, do not always exist in the infinite matrix case. We add additional conditions on the infinite matrices where necessary to ensure that all needed operations are well-defined. This typically amounts to requiring absolute summability of the reward coefficients α and bounded row absolute-summability of the matrix $\mathcal{Z}^{(a)}$.

We define operations on infinite matrices and vectors to follow typical conventions. Let $\mathcal{A}$ and $\mathcal{B}$ be infinite matrices and ν be an infinite vector. Then let $\|\mathcal{A}\|_\infty$ denote the induced infinity norm defined as $\|\mathcal{A}\|_\infty = \max_{i\in\mathbb{N}} \sum_{j=1}^\infty |\mathcal{A}_{ij}|$. Let the vector ℓ_1-norm be defined as $\|\nu\|_1 = \sum_{i=1}^\infty |\nu_i|$ and the ℓ_∞-norm defined as $\|\nu\|_\infty = \max_{i\in\mathbb{N}} |\nu_i|$. Let $[\mathcal{A}+\mathcal{B}]_{ij} = \mathcal{A}_{ij} + \mathcal{B}_{ij}$. Let the infinite matrix product be defined as $[\mathcal{A}\mathcal{B}]_{ik} = \sum_{j=1}^\infty \mathcal{A}_{ij}\mathcal{B}_{jk}$. Matrix-vector multiplication is defined as $[\mathcal{A}\nu]_i = \sum_{j=1}^\infty \mathcal{A}_{ij}\nu_j$ and $[\nu^\top\mathcal{A}]_j = \sum_{i=1}^\infty \nu_i\mathcal{A}_{ij}$. Matrix transposition is defined as $[\mathcal{A}^\top]_{ij} = \mathcal{A}_{ji}$. Define matrix power as the repeated product of a matrix, i.e., $\mathcal{A}^r = \mathcal{A}\mathcal{A}...\mathcal{A}$, r times. The outer product of infinite vectors ν and μ is denoted as $\nu\mu^\top$ and equals an infinite matrix such that $[\nu\mu^\top]_{ij} = \nu_i\mu_j$. These definitions match those of finite matrices and vectors except that the index set is countably infinite. For more information on infinite matrices, we refer the reader to [7].

Finally, we can give an infinite matrix analogue to the equations of Ng and Russell [16] to characterize the optimal solutions. We restrict our definition to only strictly optimal policies as motivated by Komanduru and Honorio [11]. For simplicity, we also assign the given optimal policy π as $\pi(s) \equiv a_1$, $\forall s \in S$, as done in [16] and [11]. Following the same intuition as in the discrete case, we can represent the expectation of a reward function after a sequence of transitions through matrix-matrix and matrix-vector multiplication. Therefore, we can use the following definition to write the difference in Q-functions of the optimal action and any other action under the optimal policy:

$$\mathcal{F}^{(a)} = \sum_{r=0}^{\infty} \gamma^r (\mathcal{Z}^{(a_1)})^r (\mathcal{Z}^{(a_1)} - \mathcal{Z}^{(a)}). \tag{6}$$

Necessary and sufficient conditions for policy optimality. The following Lemma gives necessary and sufficient conditions for policy π to be strictly optimal for a given MDP of the form described above. The condition is analogous to the condition in [16]. However, a key difference is that entries of the infinite transition matrices represent coefficients of basis functions and not transition probabilities directly. Therefore, our method of proof is necessarily different.

Lemma 2 *The given policy $\pi(s)$ is strictly optimal for the* $\mathrm{MDP}(S, A, \{P_a(s'|s)\}_{a\in A}, \gamma, R = \sum_{n=1}^\infty \alpha_n\phi_n(s))$ *if and only if,*

$$\alpha^\top \mathcal{F}^{(a)}\phi(s) > 0, \quad \forall s \in S, \ \forall\, a \in A \setminus \{a_1\}.$$

Lemma 2 will be used in Section 4 to prove correctness of our main result, Algorithm 2, as it provides a relatively easy to verify sufficient condition on the reward function representation.

3 Estimating MDP transition dynamics

As stated in the previous section, we consider the problem where the true transition dynamics of the MDP are unknown. Thus, in this section, we propose an algorithm to estimate the coefficient matrix $\mathcal{Z}^{(a)}$ and derive high-probability error bounds on the estimated matrices $\widehat{\mathcal{F}}^{(a)}$. We denote any estimated matrices with a hat, i.e., $\widehat{\mathcal{Z}}^{(a)}$ is the estimation of $\mathcal{Z}^{(a)}$ and $\widehat{\mathcal{F}}^{(a)}$ is derived from $\widehat{\mathcal{Z}}^{(a)}$.

Since Algorithm 1 is based on random samples, in our next theorem we show that, with probability at least $1-\delta$, Algorithm 1 returns an estimate of the true matrix $\mathcal{Z}^{(a)}$ with bounded error measured by the induced infinity norm.

Algorithm 1 `EstimateZ`

Input: Action $a \in A$, iteration count $n \in \mathbb{N}$, truncation parameter $k \in \mathbb{N}$.

1. Sample $\bar{s} \in \mathbb{R}^n$ such that each $\bar{s}_r$ is independent and $\bar{s}_r \sim \mathrm{Uniform}(-1, 1)$.
3. Sample $s' \in \mathbb{R}^n$ such that each s'_r is independent and $s'_r \sim P_a(\cdot|\bar{s}_r)$.
4. Compute $\widehat{\mathcal{Z}}^{(a)} = \frac{2}{n}\sum_{r=1}^{n} \phi(s'_r)\phi(\bar{s}_r)^\top$.

Output: return $[^k \widehat{\mathcal{Z}}^{(a)}]$.

Theorem 3 *Let $\widehat{\mathcal{Z}}^{(a)}$ be the output of Algorithm 1 for inputs (a, n, k), $\varepsilon > 0$, and $\delta \in (0, 1)$. If*

$$\|[^\infty_k \mathcal{Z}]^{(a)}\|_\infty \leq \varepsilon \quad \textit{and} \quad n \geq \frac{8k^2}{\varepsilon^2} \log \frac{2k^2}{\delta},$$

then $\|\widehat{\mathcal{Z}}^{(a)} - \mathcal{Z}^{(a)}\|_\infty \leq 2\varepsilon$ with probability at least $1 - \delta$.

Theorem 3 guarantees an ε-approximation of $\mathcal{Z}^{(a)}$ with probability at least $1 - \delta$ in $n \in \mathcal{O}(\frac{k^2}{\varepsilon^2} \log \frac{k}{\delta})$ samples. The condition that the truncation error is less than ε can be guaranteed by properties of the *true* MDP, as we will demonstrate in Section 5. Next, we show how the error in estimating $\mathcal{Z}^{(a)}$ propagates to the error in the estimate of infinite matrix $\mathcal{F}^{(a)}$.

Lemma 4 *If for all $a \in A$, $\|\widehat{\mathcal{Z}}^{(a)} - \mathcal{Z}^{(a)}\|_\infty \leq \varepsilon$ and $\|\mathcal{Z}^{(a)}\|_\infty, \|\widehat{\mathcal{Z}}^{(a)}\|_\infty \leq \Delta < \frac{1}{\gamma}$, then for all $a \in A \setminus \{a_1\}$,*

$$\|\widehat{\mathcal{F}}^{(a)} - \mathcal{F}^{(a)}\|_\infty \leq \frac{2\varepsilon}{(1 - \gamma\Delta)^2}.$$

Lemma 4 matches the intuition that the error of the $\{\mathcal{F}^{(a)}\}$ matrices is larger than that of the $\{\mathcal{Z}^{(a)}\}$ matrices due to the infinite summation (see Equation 6). The propagation of error will be smaller if there is a lower discount factor or the entries of the matrices $\{\mathcal{Z}^{(a)}\}$ are smaller.

4 The IRL algorithm

In this section, we present our main algorithm that recovers a strictly optimal reward function with high probability under general conditions on the infinite-matrix representations. Particularly, in Theorem 6 we give the precise assumptions on the series representation of the true MDP to guarantee the linear program in Algorithm 2 recovers an optimal reward function with high probability when using the estimated $\widehat{\mathcal{F}}^{(a)}$ matrices described in Section 3.

Algorithm 2 `ContinuousIRL`

Input: Discount factor $\gamma \in (0, 1)$, interval cover parameter $c > 0$, number of samples $n \in \mathbb{N}$, truncation parameter $k \in \mathbb{N}$.

1. Compute $\{\widehat{\mathcal{Z}}^{(a)}\}$ by calling Algorithm 1 with parameters (a, n, k) for each $a \in A$.
2. Compute $\{\widehat{\mathcal{F}}^{(a)}\}$, for each $a \neq a_1$ by using the $\{\widehat{\mathcal{Z}}^{(a)}\}$ in step 1 and Equation 6.
3. Compute a set $\bar{S} \subset [-1, 1]$ of $\lceil 2/c \rceil$ elements such that for all $s \in [-1, 1]$, there exists $\bar{s} \in \bar{S}$ such that $|s - \bar{s}| \leq c$.[2]
4. Solve the following linear program. Denote this solution by $\widehat{\alpha}$.

$$\min_\alpha \|\alpha\|_1 \ \text{ s.t. } \alpha^\top \widehat{\mathcal{F}}^{(a)} \phi(\bar{s}) \geq 1, \forall \bar{s} \in \bar{S}, a \in A \backslash \{a_1\}.$$

Output: return reward function $\widehat{R}(s) = \sum_{i=1}^{k} \widehat{\alpha}_i \phi_i(s)$.

[2] Let $\lceil \cdot \rceil$ denote the ceiling function.

We now provide an intuitive explanation of Algorithm 2. For a fixed reward vector α, the term $\alpha^\top \mathcal{F}\phi(s)$ can be regarded as a function from $S \to \mathbb{R}$. This function has a series representation over the basis functions $\{\phi_n(s)\}_{n\in\mathbb{N}}$ with coefficients given by the infinite-vector $\mathcal{F}^\top \alpha$. Therefore, the Bellman Optimality criteria can be fulfilled if α is chosen so that the coefficients $\mathcal{F}^\top \alpha$ make $\alpha^\top \mathcal{F}\phi(s)$ positive over S. However, bounds for the minimum of a function given its series representation are generally not tight enough and tighter approximations require expensive calculations such as solving a semi-definite program [13]. Instead, we exactly compute $\widehat{\mathcal{F}}\phi(\bar{s})$ for all points in a finite covering set $\bar{S}$. Using Lipschitz continuity of $\alpha^\top \widehat{\mathcal{F}}\phi(s)$, we can lower-bound $\alpha^\top \widehat{\mathcal{F}}\phi(s)$ over S. Since $\alpha^\top \widehat{\mathcal{F}}\phi(\bar{s})$ is linear in α, linear optimization can be used to ensure the lower bound is positive over S.

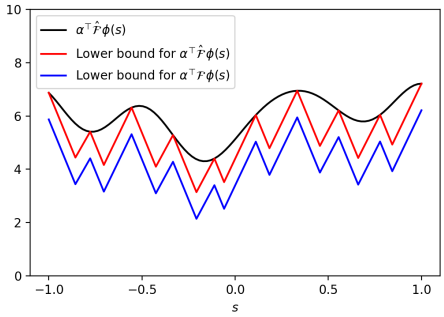

Figure 1: Visualization of Algorithm 2. Here the black line represents $\alpha^\top \widehat{\mathcal{F}}\phi(s)$ and the red line is a lower bound for this function by using the Lipschitz assumption. The blue line is then a lower bound for the true Bellman criteria $\alpha^\top \mathcal{F}\phi(s)$ where the distance between the red and blue scales with $\|\mathcal{F} - \widehat{\mathcal{F}}\|_\infty$. Algorithm 2 finds a sparse α such that $\alpha^\top \mathcal{F}^{(a)}\phi(s)$ is positive over S.

We extend the definition of β-separability given by Komanduru and Honorio [11] to the continuous case in order to provide finite sample guarantees for Algorithm 2. Our experiments indicate that β is effective in measuring the intrinsic difficulty of a given IRL problem.

Definition 2 *(β-separability). We say that the IRL problem is β-separable if there exists a reward vector $\bar{\alpha}$ such that,*

$$\|\bar{\alpha}\|_1 = 1, \text{ and}$$
$$\bar{\alpha}^\top \mathcal{F}^{(a)}\phi(s) \geq \beta > 0, \ \forall\, s \in S, \ a \in A \setminus \{a_1\}.$$

The condition of β-separability does not reduce the applicability of the algorithm since every IRL problem with a strictly optimal solution $\bar{\alpha}$ is β-separable for some $\beta > 0$. Furthermore, $\|\bar{\alpha}\|_1$ is guaranteed to be well-defined, since if there exists α such that $\alpha^\top \mathcal{F}^{(a)}\phi(s) \geq \beta > 0$, then there exists a finite truncation of α called $\bar{\alpha}$ such that $\bar{\alpha}^\top \mathcal{F}^{(a)}\phi(s) \geq \beta' > 0$.

Next, we provide conditions on the smoothness of the expected optimal reward, estimation error of $\mathcal{F}^{(a)}$, and the size of the covering set $\bar{S}$ so that the minimization problem in Algorithm 2 outputs a correct solution.

Lemma 5 *If the IRL problem is β-separable, $\|\mathcal{F}^{(a)}\phi'(s)\|_\infty \leq \rho$,[3] $\|\widehat{\mathcal{F}}^{(a)} - \mathcal{F}^{(a)}\|_\infty \leq \frac{\beta - c\rho}{2}$ for all $s \in S$, $a \in A \setminus \{a_1\}$, and $\bar{S} \subset S$ such that there exists $\bar{s} \in \bar{S}$ so $|\bar{s} - s| < c$ for all $s \in S$, then the solution (called $\widehat{\alpha}$) to the optimization problem*

$$\text{minimize}_\alpha \ \|\alpha\|_1$$
$$\text{s.t.} \ \ \alpha^\top \widehat{\mathcal{F}}^{(a)}\phi(\bar{s}) \geq 1$$

satisfies,

$$\widehat{\alpha}^\top \mathcal{F}\phi(s) > 0, \quad \forall\, s \in S.$$

In Figure 1, we provide an intuitive illustration of the correctness of Algorithm 2, which relates to the blue line denoting $\alpha^\top \mathcal{F}^{(a)}\phi(s)$ being positive over S for a given $a \in A$. First, the difference

[3] Let $\phi'_n(s) = \frac{d}{ds}\phi_n(s)$.

between the red and blue line must be small, which corresponds to $\|\widehat{\mathcal{F}}^{(a)} - \mathcal{F}^{(a)}\|_\infty$. Secondly, the piece-wise linear bounds composing the red line must be sufficiently "high". The output $\widehat{\alpha}$ will be chosen so that all points on the black curve corresponding to $\bar{s} \in \bar{S}$ will equal at least 1.

We now have the necessary components to prove general conditions on the infinite matrix representation under which Algorithm 2 outputs a solution to the IRL problem with high probability.

Theorem 6 *If the given MDP without reward function,* $(\mathrm{MDP} \setminus R)$*, is* β*-separable,* $\beta > c\rho$*, and for all* $a \in A,\ s \in [-1, 1]$*, the following conditions hold:*

$$(i)\ \|\mathcal{F}^{(a)}\phi'(s)\|_\infty < \rho, \quad (ii)\ \|\mathcal{Z}^{(a)}\|_\infty \leq \Delta < \frac{1}{2\gamma}, \quad (iii)\ \|[^\infty_k \mathcal{Z}]^{(a)}\|_\infty \leq \frac{\beta - c\rho}{8}\left(\frac{1}{2} - \gamma\Delta\right)^2,$$

$$n \geq \max\left\{32, \frac{8^3}{(\beta - c\rho)^2(\frac{1}{2} - \gamma\Delta)^4}\right\} k^2 \log \frac{2k^2|A|}{\delta},$$

then Algorithm 2 called with parameters (γ, c, n, k) *returns a correct reward function with probability at least* $1 - \delta$ *by solving a linear program with* k *variables and* $\mathcal{O}(\frac{k|A|}{c})$ *constraints.*

Remark 6.1 *The conditions (i)-(iii) on the infinite-matrix representations of the transition functions given in Theorem 6 are natural for many orthonormal bases. We explain the purpose of each condition and how they relate to assumptions on the MDP without reward function* $(\mathrm{MDP} \setminus R)$*.*

The first condition (i) enforces that the Bellman Optimality condition $\alpha^\top \mathcal{F}\phi(s)$ is ρ-Lipschitz continuous when $\|\alpha\|_1 = 1$. A finite value for ρ exists over the basis $\{\phi_n(s)\}_{n\in\mathbb{N}}$ whenever there is a convergent representation of the derivative of expected optimal reward. Intuitively, this means that expected reward of the optimal policy does not change too quickly with slight changes in the starting state.

The second condition (ii) is needed to enforce that the infinite summation which equals $\mathcal{F}$ (see Equation 6) converges in the induced infinity norm. Conditions on the MDP which enforce this assumption depend on the underlying basis of functions.

Finally, the third condition (iii) is always satisfied by some $k \in \mathbb{N}$ since it is assumed that each transition function has a convergent series representation. Different orthonormal bases provide different guarantees on bounding k given assumptions on the transition function. We will show an example of such bounds for a trigonometric basis in Section 5.

We note that our sample complexity,

$$n \in \mathcal{O}\left(\frac{1}{(\beta - c\rho)^2} k^2 \log \frac{k}{\delta}\right),$$

matches the complexity of the discrete setting algorithm given by Komanduru and Honorio [11] when $\frac{\beta}{c\rho} = \mathcal{O}(1)$.

5 Fourier instantiation

The conditions (i) - (iii) of Theorem 6 were chosen to maintain generality while being easily verified by leveraging classic results on orthonormal bases. To illustrate this claim, we prove that Theorem 6 applies to a class of transition functions over a trigonometic basis by using standard methods of proof for Fourier series. Due to space constraints, we leave the technical derivations to Appendix B and focus on the high-level idea in this section. First, we define the basis of trigonometric functions $\{\phi_n(s)\}$ as

$$\phi_n(s) = \cos\left(\lfloor n/2 \rfloor \pi s\right) \quad \text{for all odd } n, \quad \phi_n(s) = \sin\left((n/2)\pi s\right) \quad \text{for all even } n.$$

Lemma 7 *If* $P_a(\tilde{s}|s) = \sum_{i,j=1}^{\infty} \phi_i(\tilde{s}) \mathcal{Z}_{ij}^{(a)} \phi_j(s)$ *where* $\{\phi_n(s)\}$ *are the trigonometric basis functions,* $0 < \Delta < \frac{1}{2\gamma}$, $\varepsilon > 0$, *and*

$$\left| \frac{\partial^6}{\partial \tilde{s}^3 \partial s^3} P_a(\tilde{s}|s) \right| < \frac{\pi^6 \Delta}{\zeta(3)},$$

then

$$(i)\ \|\mathcal{F}\phi'(s)\|_\infty < \frac{4\pi\Delta\zeta(2)}{\zeta(3)}, \quad (ii)\ \|\mathcal{Z}\|_\infty < \Delta, \quad (iii)\ \|[^{\infty}_{k}\mathcal{Z}]\|_\infty < \varepsilon,$$

with truncation parameter k *on the order of* $k \in \mathcal{O}\left(\sqrt{\frac{\Delta}{\varepsilon}}\right)$, *where* $\zeta(r) = \sum_{n=1}^{\infty} \frac{1}{n^r}$ *is the Riemann zeta function.*

A classical result on Fourier series is that if a periodic function f is of class $\mathbb{C}^p$, then the magnitude of the n-th coefficient of its Fourier series is on the order of $\mathcal{O}(1/n^p)$. This can be used to bound the entries of $\mathcal{Z}^{(a)}$ and $\mathcal{F}^{(a)}$. Then, known inequalities of the real Riemann Zeta function and its variants can be used to guarantee the absolute infinite summations that make up conditions (i)-(iii) are appropriately bounded. An additional condition of $\Delta < \frac{1}{4\gamma}$ is then sufficient to guarantee Algorithm 2 has a simplified sample complexity of $\mathcal{O}(\beta^{-3} \log 1/\beta\delta)$ (see Appendix B).

6 Experiments

We validate our theoretical results by testing Algorithm 1 and 2 on randomly generated IRL problems using polynomial transition function. We give high level results in this section and leave the details of the experiment in Appendix C.

The following graphs indicate that the theoretical sample complexity of Algorithm 1 given by Theorem 3 is correct. We observe that the number of samples n to achieve estimation error ε is less than $\frac{8k^2}{\varepsilon^2} \log \frac{2k^2}{\delta}$, the theoretically guaranteed sample complexity.

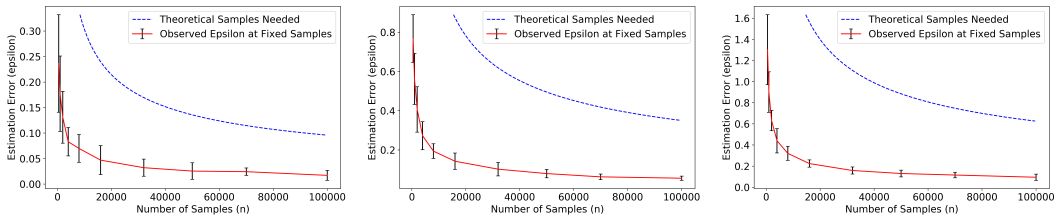

Figure 2: The red line shows the error in estimating the coefficient matrix as $\|[^{k}\mathcal{Z}] - \widehat{\mathcal{Z}}\|_\infty$ versus the number of samples n using Algorithm 1 for truncation parameter $k = 5$, $k = 15$, $k = 25$. Error bars represent two standard deviations of the observed error. The line 'Theoretical Samples Needed' shows the corresponding sample complexity given by Theorem 3.

Next, we verify the sample complexity of Algorithm 2, while varying the representation size controlled by truncation parameter k. We note that the separability parameter β has a substantial impact on the sample complexity, as indicated by the complexity result of Theorem 6. We explore this more in Appendix C.

We test the sample complexity result of Algorithm 2 by varying the truncation parameter k and number of samples n for a fixed IRL problem; that is, the actions, discount factor, and transition functions do not change. By keeping the IRL problem fixed, we can derive the following simplified relation from the sample complexity of Theorem 6 where $C > 0$ is a constant.

$$
\begin{aligned}
n &= \frac{8^3}{(\beta - c\rho)^2(\frac{1}{2} - \gamma\Delta)^4} k^2 \log \frac{2k^2|A|}{\delta} \\
\iff \frac{n}{k^2} &= \frac{8^3}{(\beta - c\rho)^2(\frac{1}{2} - \gamma\Delta)^4} (\log 2k^2|A| + \log \frac{1}{\delta}) \\
\iff \log \frac{1}{\delta} &= C\frac{n}{k^2} - \log 2k^2|A|
\end{aligned}
$$

The relation $\log \frac{1}{\delta} = C\frac{n}{k^2} - \log 2k^2|A|$ indicates that we expect to see an linear relationship between $\log {}^1\!/\delta$ and ${}^n\!/k^2$. Indeed, we observe this relationship for various values of k.

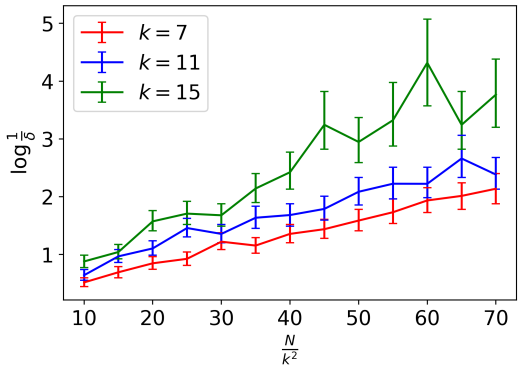

Figure 3: This plot shows the linear relationship between $\log {}^1\!/\delta$ and ${}^n\!/k^2$ while varying the truncation parameter k. The error bars show 95% confidence intervals computed via bootstrapping from 240 trials. Other constants: $\beta^{-1} = 26.5$, $\gamma = 0.7$, $c = 0.05$, $\Delta = 1.45$, and $|A| = 3$.

7 IRL in d-dimensions

In this section, we discuss IRL in the d-dimensional setting, i.e., $S = [-1, 1]^d$. One obvious way of extending Algorithm 2 to the d-dimensional setting would be to modify Lemma 5 to use ℓ_2-Lipschitz continuity of $\alpha^\top \mathcal{F}^{(a)}\phi(s)$ with an orthonormal basis $\{\phi_n(s)\}_{n\in\mathbb{N}}$ over $[-1, 1]^d$. Such a d-dimensional basis can be constructed by taking the product of 1-dimensional basis functions. This approach has the advantage that it allows for weak assumptions on the d-dimensional transition functions, which only need to satisfy the conditions of Theorem 6 when represented over some basis. However, this would require an exponential increase in the size of the covering set $\bar{S}$ used in Algorithm 2 and number of samples needed if the representativeness of the transition functions in each dimension are not restricted further.

It is unlikely that the general d-dimensional case be handled without exponential scaling in d since this also occurs in the finite state space setting. To see this, notice that if $|S|$ is finite then an IRL problem with state space S^d can be equivalent to any IRL problem on a space with $|S|^d$ states when the transition function on S^d is arbitrary. Therefore, it is necessary to add additional assumptions to reduce the worst-case computational and statistical complexity.

To simplify the d-dimensional setting, we make the assumption that the total d-dimensional state space of the MDP can be decomposed as the direct sum of T subspaces of dimension at most q such that the total transition function can be written as a product of transition functions on each of these component subspaces. In other words, we can decompose the state space as follows, $S = S^{(1)} \oplus S^{(2)} \oplus ... \oplus S^{(T)}$ such that $\dim(S^{(j)}) \leq q$. Then, given this decomposition, we require that for all $a \in A$, $P_a(s_1|s_0) = \prod_{j=1}^{T} P_a(s_1^{(j)}|s_0^{(j)})$, where $s^{(j)}$ denotes the projection of s to $S^{(j)}$. If the transition functions of a given MDP satisfy such assumptions, then we can solve the IRL problem in this setting in $\mathcal{O}(T\exp(q))$ times the computational time of solving the 1-dimensional IRL problem due to the following theorem.

Theorem 8 *Let* $(S, A, \{P_a(s_1|s_0)\}, \gamma)$ *represent an* $(\text{MDP} \setminus R)$ *such that* $S = [-1, 1]^d$ *with some decomposition* $S = S^{(1)} \oplus S^{(2)} \oplus ... \oplus S^{(T)}$ *such that* $S^{(i)} \subset S$*, and* $\dim(S^{(i)}) \leq q$*. Let the transition function operates independently on these components, i.e.,* $P_a(s_1|s_0) = \prod_{j=1}^{T} P_a(s_1^{(j)}|s_0^{(j)})$ *for all* $a \in A$ *with* $s^{(j)} = \text{Proj}_{S^{(j)}}(s)$*. If each component IRL problem with transition functions* $\{P_a(s_1^{(j)}|s_0^{(j)})\}_{a \in A}$ *is solved by reward* $R^{(j)}$*, then the total* d*-dimensional IRL problem is solved by the reward function* $R(s) = \sum_{j=1}^{T} R^{(j)}(s^{(j)})$*.*

This assumption is motivated by neurological evidence that human cognition handles the curse-of-dimensionality by decomposing a high-dimensional task into multiple subtasks [9]. This assumption was successfully used by Rothkopf and Ballard [19] in their Modular IRL algorithm to represent human navigation preference, implying that reward functions generated under this assumption could be particularly effective at representing human behavior. Further research on the type of assumptions which are empirically justified and allow strong theoretical guarantees in d-dimensions would be an interesting future topic of study.

8 Discussion

We have presented a formulation of IRL in the continuous setting by representing the transition functions with infinite-matrices over an orthonormal function basis. We show in Theorem 6 that under natural assumptions on the transition function representations, Algorithm 2 returns a reward function solving the IRL problem with high probability.

Our method of estimating the transition matrices has a sample complexity of $\mathcal{O}\left(\frac{k^2}{(\beta - c\rho)^2} \log \frac{k}{\delta}\right)$, which is same as the discrete IRL method in [11] when $\frac{\beta}{c\rho} = \mathcal{O}(1)$. The computational complexity of Algorithm 2 is dominated by solving a linear program with k variables and $\mathcal{O}(\frac{k|A|}{c})$ constraints. Furthermore, we show how our main theorem, Theorem 6, can be used to guarantee Algorithm 2 for a class of transition functions when represented over a trigonometric basis. We verify our theoretical results by empirically testing our method on randomly generated IRL problems with polynomial transition functions. Finally, we discuss how our method can be applied in the d-dimensional setting.

In future work, it would useful to better characterize transition functions that appear in applications of IRL. Much like the wavelet basis allows for sparse representations of audio and video, it is possible that particular orthonormal bases would be better suited for certain tasks where IRL is used. Better understanding of transition functions which appear in practice, such as whether bounds on its partial derviatives are justified, paired with more sophisticated results on orthonormal bases, such as those given by Izumi and Izumi [10], could allow for stronger theoretical guarantees in particular domains. Another potential line of future work could be determining which assumptions allow IRL to scale efficiently with dimension of the state space, while maintaining effectiveness on practical tasks. While our assumption on the decomposability of d-dimensional transition functions is supported by prior literature on human motor control, it is unclear how limiting this assumption would be for other applications of IRL. Different methods of simplifying IRL in the continuous d-dimensional setting may provide greater flexibility.

References

[1] Navid Aghasadeghi and Timothy Bretl. Maximum entropy inverse reinforcement learning in continuous state spaces with path integrals. In *International Conference on Intelligent Robots and Systems*, 2011.

[2] Horst Alzer. On some inequalities for the gamma and psi functions. *Mathematics of computation*, 66(217):373–389, 1997.

[3] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *ArXiv 1806.06877*, 2020.

[4] Necdet Batir. New inequalities for the hurwitz zeta function. *Proceedings Mathematical Sciences*, 2008.

[5] John J Benedetto and Wojciech Czaja. *Integration and modern analysis*. Springer, 2009.

[6] Abdeslam Boularias, Jens Kober, and Jan Peters. Relative entropy inverse reinforcement learning. In *AISTATS*, 2011.

[7] R.G. Cooke. *Infinite Matrices and Sequence Spaces*. Macmillan, 1950. URL `https://books.google.com/books?id=T9Y0AAAAMAAJ`.

[8] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *ICML*, 2016.

[9] Samuel J Gershman, Bijan Pesaran, and Nathaniel D Daw. Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *Journal of Neuroscience*, 29(43):13524–13531, 2009.

[10] Masako Izumi and Shin-ichi Izumi. Absolute convergence of fourier series of convolution functions. *Journal of Approximation Theory*, 1968.

[11] Abi Komanduru and Jean Honorio. On the correctness and sample complexity of inverse reinforcement learning. In *NeurIPS*, 2019.

[12] Sergey Levine and Vladlen Koltun. Continuous inverse optimal control with locally optimal examples. In *ICML*, 2012.

[13] Jeremy W McLean and Hugo J Woerdeman. Spectral factorizations and sums of squares representations via semidefinite programming. *SIAM journal on matrix analysis and applications*, 2002.

[14] Alberto Metelli, Matteo Pirotta, and Marcello Restelli. Compatible reward inverse reinforcement learning. In *The Thirty-first Annual Conference on Neural Information Processing Systems-NIPS 2017*, 2017.

[15] Alberto Maria Metelli, Giorgia Ramponi, Alessandro Concetti, and Marcello Restelli. Provably efficient learning of transferable rewards. In *International Conference on Machine Learning*, pages 7665–7676. PMLR, 2021.

[16] Andrew Y Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *ICML*, 2000.

[17] Matteo Pirotta and Marcello Restelli. Inverse reinforcement learning through policy gradient minimization. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[18] Giorgia Ramponi, Amarildo Likmeta, Alberto Maria Metelli, Andrea Tirinzoni, and Marcello Restelli. Truly batch model-free inverse reinforcement learning about multiple intentions. In *International Conference on Artificial Intelligence and Statistics*, pages 2359–2369. PMLR, 2020.

[19] Constantin A Rothkopf and Dana H Ballard. Modular inverse reinforcement learning for visuomotor behavior. *Biological cybernetics*, 107(4):477–490, 2013.

[20] PN Shivakumar and KC Sivakumar. A review of infinite matrices and their applications. *Linear Algebra and its Applications*, 430(4):976–998, 2009.

[21] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Supplementary material: Inverse Reinforcement Learning in a Continuous State Space with Formal Guarantees

A Proofs of lemmas and theorems

A.1 Additional lemma

Lemma 9 *Let s_0 be the starting state, let $(\boldsymbol{a})_n$ represent a sequence of actions and let $\mathcal{M} = \mathcal{Z}^{(\boldsymbol{a}_r)}\mathcal{Z}^{(\boldsymbol{a}_{r-1})}...\mathcal{Z}^{(\boldsymbol{a}_1)}$ i.e., the product of matrices in $\{\mathcal{Z}^{(a)}\}$ left multiplied in order of the sequence $(\boldsymbol{a})_n$, then,*

$$P(s_r|s_0,(\boldsymbol{a})_n) = \phi(s_r)^\top \mathcal{M}\phi(s_0).$$

Proof Here we use proof by induction. Let $\boldsymbol{X}(s_0,(\boldsymbol{a})_n) = \{s_n|n \in \mathbb{N}\}$ be a random process with a fixed starting point s_0 where $(\boldsymbol{a})_n$ is a sequence of actions and $s_r \sim P_{\boldsymbol{a}_r}(s_r|s_{r-1})$. We show that

$$\begin{aligned}
P(s_2|s_0,(\boldsymbol{a})_n) &= \int_{s_1\in S} P_{\boldsymbol{a}_2}(s_2|s_1)P_{\boldsymbol{a}_1}(s_1|s_0)ds_1 \\
&= \int_{s_1\in S}\left(\sum_{k,j=1}^{\infty}\phi_k(s_2)\mathcal{Z}_{kj}^{(\boldsymbol{a}_2)}\phi_j(s_1)\right)\left(\sum_{m,\ell=1}^{\infty}\phi_m(s_1)\mathcal{Z}_{m\ell}^{(\boldsymbol{a}_1)}\phi_\ell(s_0)\right)ds_1 \\
&= \sum_{k,j=1}^{\infty}\phi_k(s_2)\mathcal{Z}_{kj}^{(\boldsymbol{a}_2)}\sum_{m,\ell=1}^{\infty}\mathcal{Z}_{m\ell}^{(\boldsymbol{a}_1)}\phi_\ell(s_0)\int_{s_1\in S}\phi_j(s_1)\phi_m(s_1)ds_1 \\
&= \phi(s_2)\mathcal{Z}^{(\boldsymbol{a}_2)}\mathcal{Z}^{(\boldsymbol{a}_1)}\phi(s_0),
\end{aligned}$$

where the last step follows since the integral on line three equals one if $j = m$ and zero otherwise. We note that the interchange of the integral and infinite summation is justified by Section 3.7 in [5], since the coefficients $\mathcal{Z}_{ij}^{(a)}$ are absolutely summable, hence the infinite sum is upper bounded by a constant for all s_2, s_1, s_0. If there is an infinite matrix $\mathcal{M}$ such that $P(s_r|s_0,(\boldsymbol{a})_n) = \sum_{m,\ell=1}^{\infty}\phi_m(s_r)\mathcal{M}_{m\ell}\,\phi_\ell(s_0)$, then

$$\begin{aligned}
P(s_{r+1}|s_0,(\boldsymbol{a})_n) &= \int_{s_r\in S} P_{\boldsymbol{a}_{r+1}}(s_{r+1}|s_r)P(s_r|s_0,(\boldsymbol{a})_n)ds_r \\
&= \int_{s_r\in S}\left(\sum_{k,j=1}^{\infty}\phi_k(s_{r+1})\mathcal{Z}_{kj}^{(\boldsymbol{a}_{r+1})}\phi_j(s_r)\right)\left(\sum_{m,\ell=1}^{\infty}\phi_m(s_r)\mathcal{M}_{m\ell}\phi_\ell(s_0)\right)ds_r \\
&= \sum_{k,j=1}^{\infty}\phi_k(s_{r+1})\mathcal{Z}_{kj}^{(\boldsymbol{a}_{r+1})}\sum_{m,\ell=1}^{\infty}\mathcal{M}_{m\ell}\phi_\ell(s_0)\int_{s_r\in S}\phi_j(s_r)\phi_m(s_r)ds_r \\
&= \phi(s_{r+1})\mathcal{Z}^{(\boldsymbol{a}_{r+1})}\mathcal{M}\phi(s_0).
\end{aligned}$$

We can then conclude the statement of the lemma by induction.

■

A.2 Proof of Proposition 1

Proof By Lemma 9, given a fixed sequence of actions $(\boldsymbol{a})_n$, the r-th state s_r under this sequence of actions starting from state s_0 has a distribution that can be represented over the basis $\{\phi_n(s)\}$. In other words, there exists an infinite matrix $\mathcal{M}$ such that,

$$P(s_r|s_0,(\boldsymbol{a})_n) = \sum_{i,j=1}^{\infty}\phi_i(s_r)\mathcal{M}_{ij}\phi_j(s_0).$$

We now define the projection $R'(s)$ of reward function $R(s)$ to the span of $\{\phi_n(s)\}$. Let $R'(s) = \sum_{k=1}^{\infty} \alpha_k \phi_k(s)$ such that,

$$\alpha_k = \int_{s \in S} \phi_k(s) R(s) ds$$

For any $r \in \mathbb{N}$, $r > 0$, starting state s_0, and action sequence $(\boldsymbol{a})_n$, we can write

$$\begin{aligned}
E[R(s_r)|s_0, (\boldsymbol{a}_n)] &= \int_{s_r \in S} R(s_r) P(s_r|s_0, (\boldsymbol{a})_n) ds_r \\
&= \int_{s_r \in S} R(s_r) \left(\sum_{i,j=1}^{\infty} \phi_i(s_r) \mathcal{M}_{ij} \phi_j(s_0) \right) ds_r \\
&= \sum_{i,j=1}^{\infty} \mathcal{M}_{ij} \phi_j(s_0) \int_{s_r \in S} R(s_r) \phi_i(s_r) ds_r \\
&= \sum_{i,j=1}^{\infty} \mathcal{M}_{ij} \phi_j(s_0) \alpha_i \\
&= \sum_{i,j=1}^{\infty} \mathcal{M}_{ij} \phi_j(s_0) \int_{s \in S} \alpha_i \phi_i(s_r) \phi_i(s_r) ds_r \\
&= \int_{s \in S} \left(\sum_{k=1}^{\infty} \alpha_k \phi_k(s_r) \right) \left(\sum_{i,j=1}^{\infty} \phi_i(s_r) \mathcal{M}_{ij} \phi_j(s_0) \right) ds_r \\
&= \int_{s_r \in S} R'(s_r) P(s_r|s_0, (\boldsymbol{a})_n) ds_r \\
&= E[R'(s_r)|s_0, (\boldsymbol{a}_n)].
\end{aligned}$$

Therefore, the expected reward under any sequence of actions for reward R is the same as for the projected reward R' for any state s_r where $r > 0$. The reward at the starting state, $R(s_0)$ does not depend on the policy. Therefore, the value of $R(s_0)$ does not change whether a policy is optimal or not. We can conclude that, regarding optimality of π, reward function R and R' are equivalent. ■

A.3 Proof of Lemma 2

Proof For notational simplicity, let $\mathcal{T} \equiv \mathcal{Z}^{(a_1)}$. Recall that the value function is defined as,

$$V^{\pi}(s_0) = R(s_0) + \sum_{r=1}^{\infty} \gamma^r E[R(s_r)|s_0, \pi]$$

As motivated by Proposition 1, we only consider reward function R which is a (potentially infinite) sum of basis functions $\phi_n(s)$. Since in our case $\pi \equiv a_1$, we can apply Lemma 9 with action sequence $(\boldsymbol{a})_n$ such that $\boldsymbol{a}_n = a_1$ for all $n \in \mathbb{N}$, then

$$\begin{aligned}
V^{\pi}(s_0) &= R(s_0) + \sum_{r=1}^{\infty} \gamma^r E[R(s_r)|s_0, \pi] \\
&= R(s_0) + \sum_{r=1}^{\infty} \gamma^r \int_{s_r \in S} R(s_r) P(s_r|s_0) ds_r \\
&= \alpha^{\top} \phi(s_0) + \sum_{r=1}^{\infty} \gamma^r \alpha^{\top} \mathcal{T}^r \phi(s_0) \\
&= \sum_{r=0}^{\infty} \gamma^r \alpha^{\top} \mathcal{T}^r \phi(s_0)
\end{aligned}$$

We use the convention that $[\mathcal{T}^0]_{ij} = 1$ when $i = j$ and $[\mathcal{T}^0]_{ij} = 0$ otherwise. Now, define action sequence $(\boldsymbol{a})_n$ such that $\boldsymbol{a}_1 = a$ and $\boldsymbol{a}_n = a_1$ for all $n > 1$. Let $\boldsymbol{X}(s_0, (\boldsymbol{a})_n) = \{s_n | n \in \mathbb{N}\}$ be a random process with a fixed starting point s_0 where $(\boldsymbol{a})_n$ is a sequence of actions and $s_r \sim P_{\boldsymbol{a}_r}(s_r|s_{r-1})$. This expected discounted reward of this random process with parameters s_0 and a equals the function $Q^\pi(s_0, a)$, which is equivalent to the following,

$$\begin{aligned}
Q^\pi(s_0, a) &= R(s_0) + \gamma E[V^\pi(s_1)|s_0, a] \\
&= \alpha^\top \phi(s_0) + \gamma \int_{s_1 \in S} V^\pi(s_1) P_a(s_1|s_0) ds_1 \\
&= \alpha^\top \phi(s_0) + \gamma \left(\sum_{r=0}^{\infty} \gamma^r \alpha^\top \mathcal{T}^r \right) \mathcal{Z}^{(a)} \phi(s_0) \\
&= \alpha^\top \phi(s_0) + \sum_{r=0}^{\infty} \gamma^{r+1} \alpha^\top \mathcal{T}^r \mathcal{Z}^{(a)} \phi(s_0).
\end{aligned}$$

Therefore, the strict Bellman optimality condition can be written as, for all $a \in A \setminus \{\boldsymbol{a}_1\}$,

$$\begin{aligned}
& Q^\pi(s_0, a_1) - Q^\pi(s_0, a) > 0 \\
\iff & \alpha^\top \phi(s_0) + \sum_{r=0}^{\infty} \gamma^{r+1} \alpha^\top \mathcal{T}^r \mathcal{Z}^{(a_1)} \phi(s_0) - \alpha^\top \phi(s_0) + \sum_{r=0}^{\infty} \gamma^{r+1} \alpha^\top \mathcal{T}^r \mathcal{Z}^{(a)} \phi(s_0) > 0 \\
\iff & \sum_{r=0}^{\infty} \gamma^r \alpha^\top \mathcal{T}^r \mathcal{T} \phi(s_0) - \sum_{r=0}^{\infty} \gamma^r \alpha^\top \mathcal{T}^r \mathcal{Z}^{(a)} \phi(s_0) > 0 \\
\iff & \sum_{r=0}^{\infty} \gamma^r \alpha^\top \mathcal{T}^r (\mathcal{T} - \mathcal{Z}^{(a)}) \phi(s_0) > 0 \\
\iff & \alpha^\top \left(\sum_{r=0}^{\infty} \gamma^r \mathcal{T}^r (\mathcal{T} - \mathcal{Z}^{(a)}) \right) \phi(s_0) > 0
\end{aligned}$$

Note that the last step interchanges the explicit and implicit summations by again using absolute summability of α and $\mathcal{F}^{(a)}$. Substituting the definition of $\mathcal{F}^{(a)}$ into the last line allows us to conclude the following.

$$\alpha^\top \mathcal{F}^{(a)} \phi(s) > 0 \quad \forall\, s \in S,\ a \in A \setminus \{a_1\}.$$

■

A.4 Proof of Theorem 3

Proof For a fixed action a, let $\mathcal{D}$ be a joint distribution of $(s', \bar{s})$ such that $\bar{s}$ is distributed uniformly over S and, given $\bar{s}$, $s' \sim P_a(\cdot|\bar{s})$. In words, we sample the starting state $\bar{s}$ from a uniform distribution and then sample the one-step transition s'. Note that,

$$\begin{aligned}
\underset{\mathcal{D}}{E}[\phi_i(s')\phi_j(\bar{s})] &= \int_{-1}^{1} \int_{-1}^{1} P_a(s'|\bar{s}) P_a(\bar{s})\ \phi_i(s')\phi_j(\bar{s}) ds' d\bar{s} \\
&= \frac{1}{2} \int_{-1}^{1} \int_{-1}^{1} P_a(s'|\bar{s})\ \phi_i(s')\phi_j(\bar{s}) ds' d\bar{s} \\
&= \frac{1}{2} \int_{-1}^{1} \int_{-1}^{1} \left(\sum_{p,q=1}^{\infty} \phi_p(s') \mathcal{Z}_{pq}^{(a)} \phi_q(\bar{s}) \right) \phi_i(s)\phi_j(\bar{s}) ds' d\bar{s} \\
&= \frac{1}{2} \int_{-1}^{1} \int_{-1}^{1} \mathcal{Z}_{ij}^{(a)}\ \phi_i(s')^2 \phi_j(\bar{s})^2 ds' d\bar{s} \\
&= \frac{1}{2} \mathcal{Z}_{ij}^{(a)}.
\end{aligned}$$

Therefore, $\widehat{\mathcal{Z}}^{(a)} = \frac{2}{n}\sum_{r=1}^{n}\phi(s'_r)\phi(\bar{s}_r)^\top$ computed by Algorithm 1 is an unbiased estimator of $\mathcal{Z}^{(a)}$ thus $[^k\widehat{\mathcal{Z}}^{(a)}]$ is an unbiased estimator of $[^k\mathcal{Z}^{(a)}]$. We drop the superscript (a) in the rest of the proof, since the same argument applies for all $a \in A$.

We can then use Hoeffding's inequality to get concentration of measure for each entry in the truncated matrix. Since $\phi_i(s')\phi_j(\bar{s}) \in [-1, 1]$, we have

$$P\left(\left|\frac{2}{n}\sum_{r=1}^{n}\phi_i(s'_r)\phi_j(\bar{s}_r) - [^k\mathcal{Z}]_{ij}\right| \geq \varepsilon\right) \leq 2e^{\frac{-n\varepsilon^2}{8}}.$$

Then we can use subadditivity of measure to bound the maximum difference across all entries of $[^k\mathcal{Z}]$.

$$P\Big(\max_{i,j\in\mathbb{N}}\big|\frac{2}{n}\sum_{r=1}^{n}[\phi(s'_r)\phi(\bar{s}_r)]_{ij} - [^k\mathcal{Z}]_{ij}\big| \geq \varepsilon\Big) \leq 2k^2e^{\frac{-n\varepsilon^2}{8}}.$$

Lastly, since we want a bound on the error of the induced infinity norm and not element wise error, we use the following

$$\|[^k\mathcal{Z}]\|_\infty \leq k\max_{i,j\in\mathbb{N}}\left|[^k\mathcal{Z}]_{ij}\right|.$$

Therefore, the induced infinity norm error of $\widehat{\mathcal{Z}}$ is less than ε if the element wise error is less than ε/k. That is,

$$\begin{aligned}P\left(\left\|\frac{2}{n}\sum_{r=1}^{n}\phi(s'_r)\phi(\bar{s}_r) - [^k\mathcal{Z}]\right\|_\infty \geq \varepsilon\right) &\leq P\left(\max_{i,j\in\mathbb{N}}\left|\frac{2}{n}\sum_{r=1}^{n}[\phi(s'_r)\phi(\bar{s}_r)]_{ij} - [^k\mathcal{Z}]_{ij}\right| \geq \frac{\varepsilon}{k}\right)\\ &\leq 2k^2e^{\frac{-n\varepsilon^2}{8k^2}}.\end{aligned}$$

Now we find the sample complexity for fixed δ and ε. This is equivalent to $\delta \geq 2k^2e^{\frac{-n\varepsilon^2}{8k^2}}$, which is also equivalent to $n \geq \frac{8k^2}{\varepsilon^2}\log(\frac{2k^2}{\delta})$.

Then, we can bound $\|\widehat{\mathcal{Z}} - \mathcal{Z}\|_\infty$ by the triangle inequality with high probability,

$$\begin{aligned}\|\widehat{\mathcal{Z}} - \mathcal{Z}\|_\infty &= \|\widehat{\mathcal{Z}} - ([^\infty_k\mathcal{Z}] + [^k\mathcal{Z}])\|_\infty\\ &\leq \|[^\infty_k\mathcal{Z}]\|_\infty + \|\widehat{\mathcal{Z}} - [^k\mathcal{Z}]\|_\infty\\ &\leq 2\varepsilon,\end{aligned}$$

which concludes our proof. ■

A.5 Proof of Lemma 4

Proof For notational simplicity, we drop the superscript (a) and denote $\mathcal{T} \equiv \mathcal{Z}^{(a_1)}$. Recall that,

$$\mathcal{F} = \sum_{r=0}^{\infty}\gamma^r\mathcal{T}^r(\mathcal{T} - \mathcal{Z}).$$

First we bound the error of $\widehat{\mathcal{T}}^{r+1}$ as follows.

$$\begin{aligned}&\|\widehat{\mathcal{T}}^{r+1} - \mathcal{T}^{r+1}\|_\infty\\ &\quad= \|\widehat{\mathcal{T}}^r\widehat{\mathcal{T}} - \mathcal{T}^r\mathcal{T}\|_\infty\\ &\quad= \|\widehat{\mathcal{T}}^r\widehat{\mathcal{T}} - \widehat{\mathcal{T}}^r\mathcal{T} + \widehat{\mathcal{T}}^r\mathcal{T} - \mathcal{T}^r\mathcal{T}\|_\infty\\ &\quad\leq \|\widehat{\mathcal{T}}^r\widehat{\mathcal{T}} - \widehat{\mathcal{T}}^r\mathcal{T}\|_\infty + \|\widehat{\mathcal{T}}^r\mathcal{T} - \mathcal{T}^r\mathcal{T}\|_\infty\\ &\quad\leq \|\widehat{\mathcal{T}}^r\|_\infty\|\widehat{\mathcal{T}} - \mathcal{T}\|_\infty + \|\mathcal{T}\|_\infty\|\widehat{\mathcal{T}}^r - \mathcal{T}^r\|_\infty\\ &\quad\leq \|\widehat{\mathcal{T}}\|^r_\infty\|\widehat{\mathcal{T}} - \mathcal{T}\|_\infty + \|\mathcal{T}\|_\infty\|\widehat{\mathcal{T}}^r - \mathcal{T}^r\|_\infty\\ &\quad\leq \Delta^r\varepsilon + \Delta\|\widehat{\mathcal{T}}^r - \mathcal{T}^r\|_\infty.\end{aligned}$$

Define the recurrence relation $\boldsymbol{T}(r+1) = \Delta^r \varepsilon + \Delta \boldsymbol{T}(r)$ and $\boldsymbol{T}(1) = \varepsilon$. Solving this recursive formula gives the following bound,

$$\|\widehat{\mathcal{T}}^{r+1} - \mathcal{T}^{r+1}\|_\infty \leq \boldsymbol{T}(r+1) \leq (r+1)\varepsilon\Delta^r.$$

We can use the inequality above to prove the bound on the induced infinity norm error of $\widehat{\mathcal{F}}$ as follows.

$$\begin{aligned}
\|\widehat{\mathcal{F}} - \mathcal{F}\|_\infty &= \Big\| \sum_{r=0}^{\infty} \gamma^r [\widehat{\mathcal{T}}^r(\widehat{\mathcal{T}} - \widehat{\mathcal{Z}}) - \mathcal{T}^r(\mathcal{T} - \mathcal{Z})] \Big\|_\infty \\
&= \sum_{r=0}^{\infty} \gamma^r \|\widehat{\mathcal{T}}^{r+1} - \widehat{\mathcal{T}}^r \widehat{\mathcal{Z}} - \mathcal{T}^{r+1} + \mathcal{T}^r \mathcal{Z}\|_\infty \\
&\leq \sum_{r=0}^{\infty} \gamma^r (\|\widehat{\mathcal{T}}^{r+1} - \mathcal{T}^{r+1}\|_\infty + \|\widehat{\mathcal{T}}^r \widehat{\mathcal{Z}} - \mathcal{T}^r \mathcal{Z}\|_\infty) \\
&\leq \sum_{r=0}^{\infty} \gamma^r (\varepsilon(r+1)\Delta^r + \|\widehat{\mathcal{T}}^r\|_\infty \|\widehat{\mathcal{Z}} - \mathcal{Z}\|_\infty \\
&\qquad + \|\widehat{\mathcal{T}}^r - \mathcal{T}^r\|_\infty \|\mathcal{Z}\|_\infty) \\
&\leq \sum_{r=0}^{\infty} \gamma^r (\varepsilon(r+1)\Delta^r + \Delta^r \varepsilon + r\Delta^r \varepsilon) \\
&\leq \varepsilon \left(\sum_{r=0}^{\infty} 2\gamma^r (r+1)\Delta^r \right) \\
&\leq 2\varepsilon \left(\frac{1}{1-\gamma\Delta} + \frac{\gamma\Delta}{(1-\gamma\Delta)^2} \right) \\
&\leq 2\varepsilon \frac{1}{(1-\gamma\Delta)^2}.
\end{aligned}$$

Where the second to last step follows from applying the standard formula for solving arithmetico-geometric series. ■

A.6 Proof of Lemma 5

Proof The basis functions $\{\phi_n(s)\}$ are assumed to be continuously differentiable; the trigonometric basis is an example of such basis. Therefore, $\widehat{\alpha}^\top \mathcal{F}\phi(s)$ is ρ-Lipschitz if the absolute value of its derivative is bounded by ρ, i.e.

$$|\widehat{\alpha}^\top \mathcal{F}\phi'(s)| \leq \rho, \quad \forall\, s \in S.$$

Since there exists $\bar{s} \in \bar{S}$ for all $s \in S$ such that $|s - \bar{s}| < c$, if $\widehat{\alpha}^\top \mathcal{F}\phi(s)$ is ρ-Lipschitz and $\widehat{\alpha}^\top \mathcal{F}\phi(\bar{s}) \geq c\rho$ then,

$$\widehat{\alpha}^\top \mathcal{F}\phi(\bar{s}) \geq c\rho, \quad \forall\, \bar{s} \in \bar{S},$$

implies that for every $s \in S$ there exists an $\bar{s} \in \bar{S}$ such that,

$$\begin{aligned}
&\widehat{\alpha}^\top \mathcal{F}\phi(s) \geq \widehat{\alpha}^\top \mathcal{F}\phi(\bar{s}) - c\rho, \\
&\Rightarrow \widehat{\alpha}^\top \mathcal{F}\phi(s) \geq 0, \quad \forall s \in S.
\end{aligned}$$

We now bound $\|\widehat{\alpha}\|_1$ in order to assess the maximum effect the error in estimating $\mathcal{F}$ can have on the Bellman Optimality Criteria $\widehat{\alpha}^\top \mathcal{F}\phi(s)$. Since the IRL problem is β-separable, there exists $\bar{\alpha}$ such that $\|\bar{\alpha}\|_1 = 1$ and $\bar{\alpha}^\top \mathcal{F}\phi(s) \geq \beta$ for all $s \in S$.

Let $G > 0$ and $\tilde{\alpha} = G\bar{\alpha}$. Then $\|\tilde{\alpha}\|_1 = G$ and

$$\begin{aligned}\tilde{\alpha}^\top \widehat{\mathcal{F}}\phi(\bar{s}) &= \tilde{\alpha}^\top(\widehat{\mathcal{F}} - \mathcal{F} + \mathcal{F})\phi(\bar{s}) \\ &= \tilde{\alpha}^\top \mathcal{F}\phi(\bar{s}) + \tilde{\alpha}^\top(\widehat{\mathcal{F}} - \mathcal{F})\phi(\bar{s}) \\ &\geq \tilde{\alpha}^\top \mathcal{F}\phi(\bar{s}) - \|\tilde{\alpha}^\top\|_1 \|(\widehat{\mathcal{F}} - \mathcal{F})\phi(\bar{s})\|_\infty \\ &\geq G(\beta - \varepsilon).\end{aligned}$$

Therefore, if $G = 1/(\beta - \varepsilon)$ then $\tilde{\alpha}^\top \widehat{\mathcal{F}}\phi(\bar{s}) \geq 1$ and thus $\|\widehat{\alpha}\|_1 \leq 1/(\beta - \varepsilon)$. Therefore,

$$\begin{aligned}\widehat{\alpha}^\top \mathcal{F}\phi(\bar{s}) &= \widehat{\alpha}^\top(\mathcal{F} - \widehat{\mathcal{F}})\phi(\bar{s}) + \widehat{\alpha}^\top \widehat{\mathcal{F}}\phi(\bar{s}) \\ &\geq \widehat{\alpha}^\top \widehat{\mathcal{F}}\phi(\bar{s}) - \|\widehat{\alpha}\|_1 \|(\mathcal{F} - \widehat{\mathcal{F}})\phi(\bar{s})\|_\infty \\ &\geq 1 - \frac{\varepsilon}{\beta - \varepsilon}.\end{aligned}$$

We can upper bound $|\widehat{\alpha}^\top \mathcal{F}\phi'(s)|$ as follows,

$$|\widehat{\alpha}^\top \mathcal{F}\phi'(s)| \leq \|\widehat{\alpha}\|_1 \|\mathcal{F}\phi'(s)\|_\infty \leq \frac{\rho}{\beta - \varepsilon}.$$

Then, we can guarantee that $\widehat{\alpha}^\top \mathcal{F}\phi(s) > 0$ for all $s \in S$ if,

$$1 - \frac{\varepsilon}{\beta - \varepsilon} > \frac{c\rho}{\beta - \varepsilon} \iff \varepsilon < \frac{\beta - c\rho}{2}.$$

■

A.7 Proof of Theorem 6

Proof By Theorem 3, condition (iii) of Theorem 6 and the condition on n, calling Algorithm 1 with parameters (a, n, k) returns an estimate $\widehat{\mathcal{Z}}^{(a)}$ such that with probability at least $1 - \frac{\delta}{|A|}$,

$$\|\mathcal{Z}^{(a)} - \widehat{\mathcal{Z}}^{(a)}\|_\infty \leq \frac{\beta - c\rho}{4}\left(\frac{1}{2} - \gamma\Delta\right)^2.$$

Applying the union bound guarantees that the above inequality holds for all $a \in A$ with probability at least $1 - \delta$.

Note that $\|\widehat{\mathcal{Z}}^{(a)}\| \leq \Delta + \frac{1}{2\gamma}$ is satisfied for all $a \in A$ since $n \geq 32k^2 \log \frac{2k^2|A|}{\delta}$. Then by Lemma 4,

$$\begin{aligned}\|\widehat{\mathcal{F}}^{(a)} - \mathcal{F}^{(a)}\|_\infty &< \frac{2(\beta - c\rho)}{4} \frac{(\frac{1}{2} - \gamma\Delta)^2}{(1 - \gamma(\Delta + \frac{1}{2\gamma}))^2} \\ &< \frac{\beta - c\rho}{2}, \ \ \forall\, a \in A \setminus \{a_1\}.\end{aligned}$$

The minimization problem at Step 4 of Algorithm 2 fulfills the assumptions of Lemma 5, and thus it returns $\widehat{\alpha}$ such that,

$$\widehat{\alpha}^\top \mathcal{F}^{(a)}\phi(s) > 0, \quad \forall\, s \in S,\ a \in A \setminus \{a_1\}.$$

Therefore, Algorithm 2 will return a reward function such that a_1 is an optimal policy for the MDP with probability at least $1 - \delta$.

The time complexity of Algorithm 2 is dominated by solving the linear program. Since $\widehat{\mathcal{F}}$ has all zeros beyond the k-th column and row, each infinite-matrix $\widehat{\mathcal{F}}$ can be treated as a $k \times k$ matrix. Therefore, the constraints given by

$$\alpha^\top \widehat{\mathcal{F}}^{(a)}\phi(\bar{s}) \geq 1, \ \ \forall \bar{s} \in \bar{S}, \quad a \in A \setminus \{a_1\}.$$

correspond to $|\bar{S}|(|A| - 1)$ constraints in k variables. Since $|\bar{S}| = \lceil 2/c \rceil$, there are $\mathcal{O}(\frac{k|A|}{c})$ constraints in total. ■

A.8 Proof of Theorem 8

Proof

By the conditions of the Theorem, we have

$$P_a(s_1|s_0) = \prod_{j=1}^{T} P_a(s_1^{(j)}|s_0^{(j)}).$$

We use proof by induction to show that $P_a(s_r|s_0) = \prod_{j=1}^{T} P_a(s_r^{(j)}|s_0^{(j)})$ for all $r \in \mathbb{N}$. First, we prove the inductive step.

$$\begin{aligned}
P(s_{r+1}|s_0) &= \int_{s_r \in S} P(s_{r+1}|s_r)P(s_r|s_0)ds_r \\
&= \int_{s_r^{(1)}...s_r^{(T)}} \left(\prod_{j=1}^{T} P(s_{r+1}^{(j)}|s_r^{(j)})\right)\left(\prod_{j=1}^{T} P(s_r^{(j)}|s_0^{(j)})\right) ds_r^{(1)}...ds_r^{(T)} \\
&= \prod_{j=1}^{T} \int_{s_r^{(j)}} P(s_{r+1}^{(j)}|s_r^{(j)})P(s_r^{(j)}|s_0^{(j)})ds_1^{(j)} \\
&= \prod_{j=1}^{T} P(s_{r+1}^{(j)}|s_0^{(j)}).
\end{aligned}$$

The base case of $r = 1$ is guaranteed by the conditions of the Theorem. Therefore, we conclude that $P_a(s_r|s_0) = \prod_{j=1}^{T} P_a(s_r^{(j)}|s_0^{(j)})$ for all $r \in \mathbb{N}$.

Next, we use the previous decomposition of $P_a(s_r|s_0)$ to rewrite the Bellman Optimality condition. Recall that R is strictly Bellman-optimal if for all $a \in A \setminus \{a_1\}$ and $s_0 \in S$,

$$E[V^\pi(s_1)|s_0, a_1] > E[V^\pi(s_1)|s_0, a].$$

We can rearrange each side of the above inequality by the following.

$$\begin{aligned}
E[V^\pi(s_1)|s_0, a] &= \sum_{r=1}^{\infty} \gamma^r \int_{s_r \in S} R(s_r)P_{a_1}(s_r|s_1)P_a(s_1|s_0)Ts_r \\
&= \sum_{r=1}^{\infty} \gamma^r \int_{s_r^{(1)}...s_r^{(T)}} \left(\sum_{i=1}^{T} R^{(i)}(s^{(i)})\right)\left(\prod_{j=1}^{T} P_{a_1}(s_r^{(j)}|s_1^{(j)})P_a(s_1^{(j)}|s_0^{(j)})\right) ds_r^{(1)}...ds_r^{(T)} \\
&= \sum_{r=1}^{\infty} \gamma^r \int_{s_r^{(1)}...s_r^{(T)}} \sum_{i=1}^{T} \left(R^{(i)}(s^{(i)}) \prod_{j=1}^{T} P_{a_1}(s_r^{(j)}|s_1^{(j)})P_a(s_1^{(j)}|s_0^{(j)})\right) ds_r^{(1)}...ds_r^{(T)} \\
&= \sum_{r=1}^{\infty} \gamma^r \sum_{i=1}^{T} \int_{s_r^{(1)}...s_r^{(T)}} R^{(i)}(s^{(i)}) \prod_{j=1}^{T} P_{a_1}(s_r^{(j)}|s_1^{(j)})P_a(s_1^{(j)}|s_0^{(j)})ds_r^{(1)}...ds_r^{(T)} \\
&= \sum_{r=1}^{\infty} \gamma^r \sum_{i=1}^{T} \int_{s_r^{(i)}} R^{(i)}(s^{(i)})P_{a_1}(s_r^{(i)}|s_1^{(i)})P_a(s_1^{(i)}|s_0^{(i)})ds_r^{(i)} \\
&= \sum_{i=1}^{T} \sum_{r=1}^{\infty} \gamma^r \int_{s_r^{(i)}} R^{(i)}(s^{(i)})P_{a_1}(s_r^{(i)}|s_1^{(i)})P_a(s_1^{(i)}|s_0^{(i)})ds_r^{(i)}.
\end{aligned}$$

The second-to-last step above follows since the integral of any probability distribution equals 1.

Since each 1-dimensional IRL problem is solved by $R^{(j)}(s^{(j)})$,

$$\sum_{r=1}^{\infty}\gamma^r\int_{s_r^{(j)}}R^{(j)}(s_r^{(j)})P_{a_1}(s_r^{(j)}|s_1^{(j)})P_{a_1}(s_1^{(j)}|s_0^{(j)})ds_r^{(j)}$$
$$>\sum_{r=1}^{\infty}\gamma^r\int_{s_r^{(j)}}R(s_r^{(j)})P_{a_1}(s_r^{(j)}|s_1^{(j)})P_{a}(s_1^{(j)}|s_0^{(j)})ds_r^{(j)}.$$

This implies that,

$$\sum_{i=1}^{T}\sum_{r=1}^{\infty}\gamma^r\int_{s_r^{(j)}}R^{(j)}(s_r^{(j)})P_{a_1}(s_r^{(j)}|s_1^{(j)})P_{a_1}(s_1^{(j)}|s_0^{(j)})ds_r^{(j)}$$
$$>\sum_{i=1}^{T}\sum_{r=1}^{\infty}\gamma^r\int_{s_r^{(j)}}R(s_r^{(j)})P_{a_1}(s_r^{(j)}|s_1^{(j)})P_{a}(s_1^{(j)}|s_0^{(j)})ds_r^{(j)}$$
$$\Rightarrow$$
$$E[V^{\pi}(s_1)|s_0,a_1]>E[V^{\pi}(s_1)|s_0,a].$$

Therefore, reward function $R(s)$ is strictly Bellman optimal for the total IRL problem.

■

B Fourier instantiation

We prove Lemma 7 by splitting the derivations into two lemmas. First we state stronger conditions on the infinite-matrix $\mathcal{Z}$ which are sufficient to satisfy the representation assumptions of Theorem 6. These stronger conditions will be easier to verify using known proof techniques for Fourier series.

Lemma 10 *Let* $0<\Delta<\frac{1}{2\gamma}$. *If,*

$$|\mathcal{Z}_{ij}|<\frac{\Delta}{\zeta(3)i^3j^3},\quad \textit{and}\quad |\phi_n'(s)|\le Cn,\ \ C>0,$$

then

$$\|\mathcal{F}\phi'(s)\|_\infty<\frac{4C\Delta\zeta(2)}{\zeta(3)},\qquad \|[^{\infty}_{k}\mathcal{Z}]\|_\infty<\varepsilon,\qquad \|\mathcal{Z}\|_\infty<\Delta,$$

with truncation parameter k on the order of $k\in\mathcal{O}\left(\sqrt{\frac{\Delta}{\varepsilon}}\right)$, where $\zeta(r)=\sum_{n=1}^{\infty}\frac{1}{n^r}$ is the Riemann zeta function and $\varepsilon>0$.

Proof By the condition of the lemma, $|\mathcal{Z}_{ij}|<\frac{\Delta}{\zeta(3)i^3j^3}$. First we bound, $\|\mathcal{Z}\|_\infty$,

$$\begin{aligned}\|\mathcal{Z}\|_\infty&=\max_{i\in\mathbb{N}}\sum_{j=1}^{\infty}|\mathcal{Z}_{ij}|\\&<\max_{i\in\mathbb{N}}\sum_{j=1}^{\infty}\frac{\Delta}{\zeta(3)i^3j^3}\\&=\frac{\Delta}{\zeta(3)}\sum_{j=1}^{\infty}\frac{1}{j^3}\\&=\frac{\zeta(3)\Delta}{\zeta(3)}\\&=\Delta.\end{aligned}$$

Therefore, we can guarantee that $\|\mathcal{Z}\|_\infty < \Delta$. Next we bound $\|[_k^\infty \mathcal{Z}]\|_\infty$. First, we define the Hurwitz zeta function, $H_s(x)$.

$$H_s(x) = \sum_{n=0}^{\infty} \frac{1}{(n+x)^s} \tag{7}$$

Recall that $[_k^\infty \mathcal{Z}]_{ij} = 0$ for all $i, j \leq k$. Therefore,

$$\begin{aligned}
||[_k^\infty \mathcal{Z}]||_\infty &= \max\left\{ \max_{i\in\{1,\ldots,k\}} \sum_{j=k+1}^{\infty} |\mathcal{Z}_{i,j}|, \quad \max_{i\in\{k+1,k+2,\ldots\}} \sum_{j=1}^{\infty} |\mathcal{Z}_{i,j}| \right\} \\
&< \max\left\{ \sum_{j=k+1}^{\infty} \frac{\Delta}{j^3\zeta(3)}, \quad \sum_{j=1}^{\infty} \frac{\Delta}{(k+1)^3 j^3 \zeta(3)} \right\} \\
&= \max\left\{ \frac{\Delta}{\zeta(3)} \sum_{j=k+1}^{\infty} \frac{1}{j^3}, \quad \frac{\Delta}{(k+1)^3\zeta(3)} \sum_{j=1}^{\infty} \frac{1}{j^3} \right\} \\
&= \max\left\{ \frac{\Delta}{\zeta(3)} H_3(k+1), \quad \frac{\Delta}{(k+1)^3} \right\}
\end{aligned}$$

We can bound the Hurwitz zeta function by combining the following two inequalities where $\psi(x)$ represents the digamma function. The first inequality comes from Theorem 3.1 in [4] and the second comes from Equation 2.2 in [2].

$$\begin{aligned}
H_{s+1}(x) &< \frac{1}{s}\exp(-s\psi(x)) \\
\log x - \frac{1}{x} \leq \psi(x) &\leq \log x - \frac{1}{2x}
\end{aligned}$$

This gives

$$\begin{aligned}
H_3(k+1) &\leq \frac{1}{2}\exp\left(-2\left(\log(k+1) - \frac{1}{k+1}\right)\right) \\
&= \frac{1}{2} e^{-2\log(k+1)} e^{\frac{2}{k+1}} \\
&= \frac{1}{2(k+1)^2} e^{\frac{2}{(k+1)}}.
\end{aligned}$$

We can then bound the truncation error as,

$$\begin{aligned}
||[_k^\infty \mathcal{Z}]||_\infty &< \max\left\{ \frac{\Delta}{\zeta(3)} H_3(k+1), \frac{\Delta}{(k+1)^3} \right\} \\
&\leq \max\left\{ \frac{\Delta e}{2\zeta(3)(k+1)^2}, \frac{\Delta}{(k+1)^3} \right\}.
\end{aligned}$$

Therefore, $k \in \mathcal{O}(\sqrt{\Delta/\varepsilon})$ is sufficient to guarantee that $||[_k^\infty \mathcal{Z}]||_\infty < \varepsilon$. Finally, we bound $\|\mathcal{F}\phi'(s)\|_\infty$. Recall that,

$$\mathcal{F} = \sum_{r=0}^{\infty} \gamma^r \mathcal{T}^r (\mathcal{T} - \mathcal{Z})$$

Then,

$$\begin{aligned}
\|\mathcal{F}\phi'(s)\|_\infty &= \left\| \left[\sum_{r=0}^{\infty} \gamma^r \mathcal{T}^r (\mathcal{T} - \mathcal{Z}) \right] \phi'(s) \right\|_\infty \\
&= \left\| \sum_{r=0}^{\infty} \gamma^r \mathcal{T}^r (\mathcal{T} - \mathcal{Z}) \phi'(s) \right\|_\infty \\
&\leq \sum_{r=0}^{\infty} \gamma^r \|\mathcal{T}^r (\mathcal{T} - \mathcal{Z}) \phi'(s)\|_\infty \\
&\leq \sum_{r=0}^{\infty} \gamma^r \|\mathcal{T}^r\|_\infty \|(\mathcal{T} - \mathcal{Z}) \phi'(s)\|_\infty \\
&\leq \|(\mathcal{T} - \mathcal{Z}) \phi'(s)\|_\infty \sum_{r=0}^{\infty} \gamma^r \|\mathcal{T}^r\|_\infty
\end{aligned}$$

We bound $\|(\mathcal{T} - \mathcal{Z})\phi'(s)\|_\infty$ and $\sum_{r=0}^{\infty} \gamma^r \|\mathcal{T}^r\|_\infty$ separately. By the conditions of the lemma, $\|\phi_n'(s)\| \leq Cn$. Then using the fact that $\|\mathcal{Z}\|_\infty, \|\mathcal{T}\|_\infty < \Delta$.

$$\begin{aligned}
\|(\mathcal{T} - \mathcal{Z})\phi'(s)\|_\infty &= \max_{i \in \mathbb{N}} \left| \sum_{j=1}^{\infty} [\mathcal{T} - \mathcal{Z}]_{ij} \phi_j'(s) \right| \\
&< \max_{i \in \mathbb{N}} \left| \sum_{j=1}^{\infty} \frac{2\Delta}{\zeta(3) i^3 j^3} Cj \right| \\
&= \frac{2C\Delta}{\zeta(3)} \left| \sum_{j=1}^{\infty} \frac{1}{j^2} \right| \\
&= \frac{2C\Delta\zeta(2)}{\zeta(3)}
\end{aligned}$$

Then by the conditions of the lemma, $\|\mathcal{T}\|_\infty < \Delta < \frac{1}{2\gamma}$, which implies,

$$\begin{aligned}
\sum_{r=0}^{\infty} \gamma^r \|\mathcal{T}^r\|_\infty &\leq \sum_{r=0}^{\infty} \gamma^r \|\mathcal{T}\|_\infty^r \\
&< \sum_{r=0}^{\infty} \gamma^r \frac{1}{(2\gamma)^r} \\
&= \frac{1}{1 - \frac{1}{2}} = 2
\end{aligned}$$

Combining these previous results gives,

$$\|\mathcal{F}\phi'(s)\|_\infty \leq \|(\mathcal{T} - \mathcal{Z})\phi'(s)\|_\infty \sum_{r=0}^{\infty} \gamma^r \|\mathcal{T}^r\|_\infty < \frac{4C\Delta\zeta(2)}{\gamma\zeta(3)}$$

■

The next lemma states conditions on the partial derivatives of the transition functions so that the infinite-matrix representations over a trigonometric basis fulfill the conditions of Lemma 10.

Lemma 11 *If $P(\tilde{s}|s) = \sum_{i,j=1}^{\infty} \phi_i(\tilde{s}) \mathcal{Z}_{ij} \phi_j(s)$ where $\{\phi_n(s)\}$ are the trigonometric basis functions, and*

$$\left| \frac{\partial^6}{\partial \tilde{s}^3 \partial s^3} P(\tilde{s}|s) \right| < \frac{\pi^6 \Delta}{\zeta(3)},$$

then,

$$|\mathcal{Z}_{ij}| < \frac{\Delta}{\zeta(3) i^3 j^3}.$$

Proof

We defined the basis of trigonometric functions $\{\phi_n(s)\}$ as

$$\phi_n(s) = \cos\left(\lfloor n/2 \rfloor \pi s\right) \quad \text{for all odd } n,$$
$$\phi_n(s) = \sin\left((n/2)\pi s\right) \quad \text{for all even } n,$$

where $\lfloor . \rfloor$ is the floor function.

First we directly represent the partial derivative in terms of the trigonometric basis. We define $\mathcal{Z}'_{ij}$ as

$$\begin{aligned}\mathcal{Z}'_{ij} &\equiv \int_{s_1, s_0 \in S} \frac{\partial^6}{\partial s_1^3 \partial s_0^3} P(s_1|s_0)\phi_i(s_1)\phi_j(s_0) ds_1 ds_0 \\ &\Rightarrow \frac{\partial^6}{\partial s_1^3 \partial s_0^3} P(s_1|s_0) = \sum_{i,j=1}^{\infty} \phi_i(s_1)\mathcal{Z}'_{ij}\phi_j(s_0).\end{aligned}$$

We use the bound on the sixth order partial derivative to bound the values of $\mathcal{Z}'$ as follows.

$$\begin{aligned}|\mathcal{Z}'_{ij}| &= \left| \int_{s_1, s_0 \in S} \frac{\partial^6}{\partial s_1^3 \partial s_0^3} P(s_1|s_0)\phi_i(s_1)\phi_j(s_0) ds_1 ds_0 \right| \\ &< \frac{\pi^6 \Delta}{\zeta(3)} \left| \int_{s_1, s_0 \in S} \phi_i(s_1)\phi_j(s_0) ds_1 ds_0 \right| \\ &\leq \frac{\pi^6 \Delta}{\zeta(3)}.\end{aligned}$$

Next we represent the sixth order partial derivative using the entries of $\mathcal{Z}$ by differentiating the series representation of $P(s'|s)$.

$$P(s_1|s_0) = \sum_{i,j=1}^{\infty} \phi_i(s_1)\mathcal{Z}_{ij}\phi_j(s_0) \Rightarrow$$
$$\frac{\partial^6}{\partial s_1^3 \partial s_0^3} P(s_1|s_0) = \sum_{i,j=1}^{\infty} \frac{\partial^3}{\partial s_1^3}\phi_i(s_1)\mathcal{Z}_{ij}\frac{\partial^3}{\partial s_0^3}\phi_j(s_0).$$

If n is odd then,

$$\frac{\partial^3}{\partial s^3}\phi_n(s) = \frac{\partial^3}{\partial s^3}\cos(\lfloor n/2 \rfloor \pi s) = (\pi n)^3 \sin(\lfloor n/2 \rfloor \pi s) = (\pi n)^3 \phi_{n-1}(s).$$

If n is even then,

$$\frac{\partial^3}{\partial s^3}\phi_n(s) = \frac{\partial^3}{\partial s^3}\sin((n/2)\pi s) = -(\pi n)^3 \cos((n/2)\pi s) = -(\pi n)^3 \phi_{n+1}(s).$$

Therefore we can map the entries of $\mathcal{Z}$ to $\mathcal{Z}'$. The exact mapping is not important since all entries of $\mathcal{Z}'$ are bounded by $\frac{\pi^6 \Delta}{\zeta(3)}$. Finally, we have,

$$\begin{aligned}\pi^6 i^3 j^3 |\mathcal{Z}_{ij}| = |\mathcal{Z}'_{i\pm1, j\pm1}| &< \frac{\pi^6 \Delta}{\zeta(3)} \\ \Rightarrow |\mathcal{Z}_{ij}| &< \frac{\Delta}{\zeta(3) i^3 j^3}.\end{aligned}$$

■

Combining the statements of Lemma 10 and Lemma 11 along with the fact that $|\frac{d}{ds}\cos(\lfloor n/2\rfloor \pi s)|, |\frac{d}{ds}\sin((n/2)\pi s)| \leq \pi n$ constitutes the proof of Lemma 7.

We can strengthen the assumption on Δ in Lemma 7 from $\Delta < \frac{1}{2\gamma}$ to $\Delta < \frac{1}{4\gamma}$ to achieve a simple sample complexity for Algorithm 2 which only depends on the probability of failure δ and the separability measure β.

Corollary 12 *If the IRL problem is β-separable and*

$$\left|\frac{\partial^6}{\partial \tilde{s}^3 \partial s^3} P_a(\tilde{s}|s)\right| < \frac{\pi^6}{4\gamma\zeta(3)}, \ \forall a \in A,$$

then with $\mathcal{O}\left(\frac{1}{\beta^3}\log\frac{1}{\beta\delta}\right)$ samples, Algorithm 2 with a trigonometric basis outputs a correct reward function R with probability at least $1-\delta$.

Proof

By Lemma 7, we have the following bounds when substituting in $\Delta < \frac{1}{4\gamma}$.

$$(i)\ \|\mathcal{F}\phi'(s)\|_\infty < \frac{\pi\zeta(2)}{\gamma\zeta(3)}, \quad (ii)\ \|\mathcal{Z}\|_\infty < \frac{1}{4\gamma}, \quad (iii)\ \|[\mathcal{Z}]_k^\infty\|_\infty < \varepsilon, \text{ with } k \in \mathcal{O}\left(\sqrt{\frac{1}{\varepsilon}}\right)$$

Theorem 6 then guarantees that Algorithm 2 will return a correct reward function with probability at least $1-\delta$ for some $\gamma \in (0,1)$, $c > 0$, $k \in \mathcal{O}\left(\sqrt{\frac{1}{\varepsilon}}\right)$, and $n \in \mathcal{O}\left(\frac{1}{(\beta-c\rho)^2}k^2\log\frac{k}{\delta}\right)$.

To see this, let

$$\varepsilon = \frac{\beta - c\rho}{8}\left(\frac{1}{2} - \gamma\Delta\right)^2 = \frac{\beta - c\rho}{128}, \quad \text{and} \quad \rho = \frac{\pi\zeta(2)}{\gamma\zeta(3)}.$$

We can simplify the sample complexity result by setting c such that $c\rho = \frac{\beta}{2}$ and using $k \in \mathcal{O}\left(\sqrt{\frac{1}{\varepsilon}}\right)$ to obtain,

$$\begin{aligned}
\mathcal{O}\left(\frac{1}{(\beta-c\rho)^2}k^2\log\frac{k}{\delta}\right) &= \mathcal{O}\left(\frac{1}{(\beta-c\rho)^2}\frac{1}{\beta-c\rho}\log\frac{1}{(\beta-c\rho)\delta}\right) \\
&= \mathcal{O}\left(\frac{1}{(\beta)^2}\frac{1}{\beta}\log\frac{1}{\beta\delta}\right) \\
&= \mathcal{O}\left(\frac{1}{\beta^3}\log\frac{1}{\beta\delta}\right).
\end{aligned}$$

which concludes the proof. ■

C Experiments

In order to verify our theoretical results, we test our algorithm on simple randomly generated IRL problems. To accomplish this, we randomly generate IRL problems with polynomial transition functions. Using polynomial transition functions has two main advantages. First, we can obtain simple closed form solutions when generating the coefficient matrix $\mathcal{Z}^{(a)}$. Second, the transition function is infinitely differentiable, meaning the truncation error rapidly decreases with k. We conducted our experiments on a 64-core AMD Epyc 7662 'Rome' processor with 256 GB memory and coded our experiments using Python 3.

C.1 Generating transition functions

First, we describe a way to generate a random polynomial which is a valid probability distribution function over $S = [-1, 1]$. Let $\mathcal{P}_r = a(x-b)^{2r}$ where $a, b \sim$Uniform$(0,1)$ denote a polynomial with variable x. Notice that for all $r \in \mathbb{N}$, $\mathcal{P}_r$ is non-negative over S. Therefore, we can construct a

non-negative (even degree) polynomial $\mathcal{P} = \sum_{r=1}^{d/2} \mathcal{P}_r$. Re-normalizing $\mathcal{P}$ so that it integrates to 1 over S then makes $\mathcal{P}$ a valid probability density function since it integrates to one and is non-negative.

To generate a transition function, $P(s'|s)$, create two random polynomial distributions $\mathcal{P}_a$ and $\mathcal{P}_b$ as described above. Then let $P(s'|s) = (1 - s^2)\mathcal{P}_a(s') + s^2\mathcal{P}_b(s')$. For each fixed s, $P(s'|s)$ is a weighted average of two probability density functions, and is therefore a probability density function.

C.2 Sampling the coefficient matrix (Algorithm 1) (Figure 2)

In order to test Algorithm 1 and the guarantees of Theorem 3, we must compute $\mathcal{Z}$ and $\widehat{\mathcal{Z}}$. Specifically, to create Figure 2, we first generate a fixed random transition function. Next, we implement Algorithm 1 to compute $\widehat{\mathcal{Z}}$ for a specified truncation size k, where we compute the one-step transition $s'_r \sim P_a(\cdot|\bar{s}_r)$ using inverse transform sampling with eight bits of precision. We cannot compute the infinite matrix $\mathcal{Z}^{(a)}$ completely, so instead we restrict ourselves to measuring the error $\|[^k\mathcal{Z}] - \widehat{\mathcal{Z}}\|_\infty$. The i, j-th entry of $\mathcal{Z}^{(a)}$ is given by the following formula.

$$\begin{aligned}\mathcal{Z}_{ij} &= \int_{-1}^{1}\int_{-1}^{1}((1-s^2)\mathcal{P}_a(s') + s^2\mathcal{P}_b(s'))\phi_i(s')\phi_j(s)ds'ds \\ &= \int_{-1}^{1}\left((1-s^2)\int_{-1}^{1}\mathcal{P}_a(s')\phi_i(s')ds'\right)\phi_j(s)ds + \int_{-1}^{1}\left(s^2\int_{-1}^{1}\mathcal{P}_b(s')ds'\phi_i(s')\right)\phi_j(s)ds.\end{aligned}$$

We then use the fact that a polynomial is a sum of monomials. This allows us to give a simple recursive specification of the integral of a polynomial with each trigonometic basis function using integration by parts.

C.3 Solving the IRL problem (Algorithm 2) (Figure 3)

Next, we conduct experiments testing our ability to solve a randomly generated IRL problem. In these experiments, we use $\gamma = 0.7$ and $|A| = 3$, where each transition function is as described above. We compute each $\widehat{\mathcal{Z}}^{(a)}$ using Algorithm 1 and the compute the reward vector $\widehat{\alpha}$ using Algorithm 2. We classify the returned $\widehat{\alpha}$ as "correct" or "incorrect" by checking if $\widehat{\alpha}^T\widehat{\mathcal{F}}^{(a)}\phi(\bar{s}) > 0$ for all $\bar{s} \in \bar{S}$ where $\bar{S}$ is a covering set over $[-1, 1]$ of size 100. If this inequality does not hold for any $\bar{s} \in \bar{S}$, then the reward vector is classified as "incorrect". To generate Figure 3, we repeat the above process 320 times for each value of n and each value of k.

Additionally, we verified that we implement Algorithm 2 correctly by checking the empirical expected reward on a set of 100 points in $[-1, 1]$ for several IRL problems. To compute the empirical expected reward at state s_0, we sample a sequence of point $s_1 ... s_6$ such that $s_r \sim P_a(\cdot|s_{r-1})$ using inverse transform sampling, as described above. We average the discounted reward of these 6 points over 6000 samples and find that we succeed in generating a reward vector $\widehat{\alpha}$ such that the expected reward of the first action is higher than the expected reward of any other action at each starting point s_0.

C.4 Effect of β-separability

As expected from the sample complexity result of Theorem 6, we found the β-separability had a significant impact on the sample complexity of Algorithm 2 (see Figure 4 below).

We find the separability measure β of a randomly generated IRL problem by running Algorithm 2 using the exact $[^k\mathcal{Z}]$ matrices with $k = 11$. Using the exact matrices removes any error introduced by sampling, and since the transition functions are infinitely differentiable, the truncation error is minimal. Since $\widehat{\alpha}\mathcal{F}^{(a)}\phi(s) \geq 1$ holds approximately, we can approximate β as $\beta = \frac{1}{\|\widehat{\alpha}\|_\infty}$.

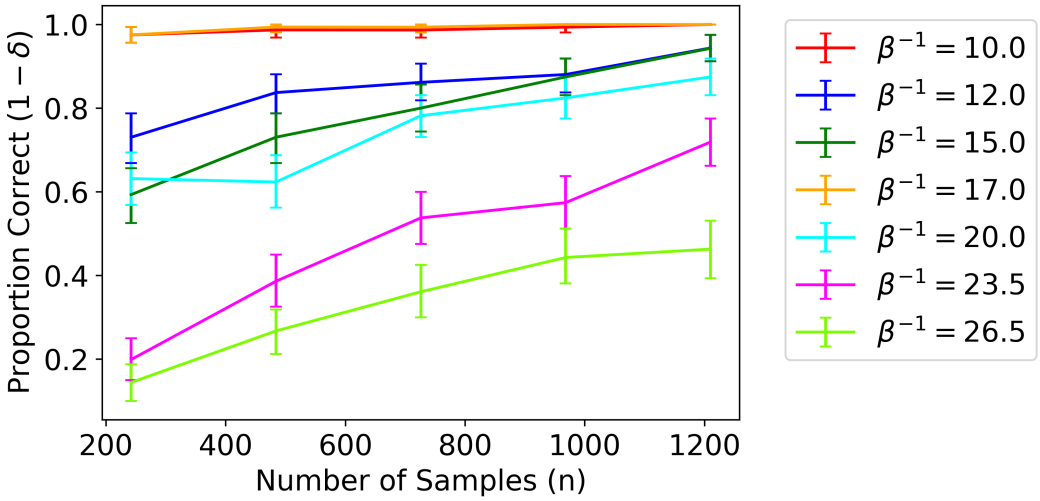

Figure 4: This graph shows the proportion of out of of 160 trials at each value of n where the returned reward vector is approximately Bellman optimal versus the number of samples. We plot the results from multiple generated IRL problems across a range of values for β^{-1}. We observe that β has a significant impact on the samples needed in Algorithm 2. Error bars represent 95% confidence and are computed by bootstrapping out of 120 trials.