A Lower Bound for the Sample Complexity of Inverse Reinforcement Learning

Abi Komanduru [1] Jean Honorio [2]

Abstract

Inverse reinforcement learning (IRL) is the task of finding a reward function that generates a desired optimal policy for a given Markov Decision Process (MDP). This paper develops an information-theoretic lower bound for the sample complexity of the finite state, finite action IRL problem. A geometric construction of β-strict separable IRL problems using spherical codes is considered. Properties of the ensemble size as well as the Kullback-Leibler divergence between the generated trajectories are derived. The resulting ensemble is then used along with Fano's inequality to derive a sample complexity lower bound of $O(n \log n)$, where n is the number of states in the MDP.

1. Introduction

Reinforcement learning (RL) focuses on the well studied problem of finding an optimal policy for a given Markov Decision Process (MDP) with a known reward function. Inverse Reinforcement Learning (IRL) (Ng & Russel, 2000) considers a MDP with known optimal policy and aims to find a reward function that generates the desired optimal policy. It is well known that the choice of such reward function is not necessarily unique. The IRL problem occurs in situations where the actions of an *expert*, which represent the optimal policy, are known or can be observed and are to be replicated through the proper choice of a reward function. Examples include cases such as apprenticeship learning.

Two major formulations of the IRL problem have been proposed. The first is the standard MDP formulation considered by (Ng & Russel, 2000). This is the formulation that is considered in this paper and for which the results are derived. The second is the linearly-solvable MDP (LMDP) formulation of (Dvijotham & Todorov, 2010). As noted in (Dvijotham & Todorov, 2010), while the standard MDP problem can be embedded in LMDP, solutions to standard MDP problems based on standard MDPs are guaranteed to generate the desired Bellman optimal policy given the true transition probabilities whereas the LMDP formulation does not. Both formulations are used successfully in practice. Methods to solve the standard MDP formulations include the methods presented in (Ng & Russel, 2000), (Abbeel & Ng, 2004), Multiplicative Weights for Apprenticeship Learning (Syed et al., 2008), Bayesian Estimation IRL (Ramachandran & Amir, 2007), Maximum Margin Planning (Ratliff et al., 2006), Hybrid IRL (Neu & Szepesvári, 2007) and the L1 SVM formulation(Komanduru & Honorio, 2019). Examples of methods for solving the LMDP formulation are Maximum-Entropy IRL (Ziebart et al., 2008) and Gaussian Process IRL (Levine et al., 2011).

As shown in (Komanduru & Honorio, 2019), various solutions to the standard MDP problem can fail to result in a reward that uniquely generates the desired optimal policy. This failure can render such reward function solutions useless in the case where the goal is to replicate the policy of the expert and not just simply achieve similar values for the value function. With this in mind, (Komanduru & Honorio, 2019) derived an upper bound of $O(\frac{n^2}{\beta^2} \log(nk))$ for the sample complexity of the inverse reinforcement learning problem. In this paper, we derive an information-theoretic lower bound for the sample complexity of the standard MDP IRL problem when the transition probabilities are estimated from observed trajectories. In this case, a sample is an observation of tuples consisting of the previous state, action and resultant state from the observed trajectories. The derived sample complexity is a bound on the number of samples with respect to the goal of recovering a reward function that correctly generates the desired optimal policy.

We use Fano's inequality (Cover & Thomas, 2006) to prove our result through a careful construction of an ensemble. The use of restricted ensembles is customary for information-theoretic lower bounds (Santhanam & Wainwright, 2012), (Wang et al., 2010), (Tandon et al., 2014). To the best of our knowledge, no such information-theoretic sample complexity lower bound exists for the recovery of the reward function with the stated properties in the case of standard MDP IRL.

In the following section, we review the basic notation of the

[1] Purdue University, Indiana, USA [2]Purdue University, Indiana USA. Correspondence to: Abi Komanduru <akomandu@purdue.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Inverse Reinforcement Learning problem, the conditions for Bellman optimality of the optimal policy for standard MDP and the notion of β-strict separability. In Section 3, we describe the geometric construction of the ensemble using spherical codes. Section 4 derives bounds for the cardinality and the KL divergence within the ensemble and culminates with the $O(n \log n)$ lower bound for the sample complexity of inverse reinforcement learning. Appendix B provides results from simulated experiments using various solutions methods to support our sample complexity bound.

2. Preliminaries and Notation

Consider the standard Markov Decision Process $(S, A, \{P_a\}, \gamma, R)$, where

- S is a finite set of n states.
- $A = \{a_1, \ldots, a_k\}$ is a set of k actions.
- $P_a \in [0,1]^{n \times n}$ are the state transition probabilities for action a. We use $P_a(i) \in [0,1]^n$ to represent the state transition probabilities for action a in the i-th state or more simply, the i-th row of the transition probability matrix P_a
- $\gamma \in [0,1]$ is the discount factor.
- $R : S \to \mathbb{R}$ is the reward function.

In our representation, we consider the P_a to be right stochastic matrices, i.e.,

$P_a(i,j) \geq 0 \; \forall \; i,j$ and $\sum_j P_a(i,j) = 1 \; \forall i$

where $P_a(i,j)$ are the entries of $P_a(i)$ and represent the transition probability of going from state i to state j when taking action a.

We assume the reward function to depend only on the state instead of the state and the action. This assumption is also made for the prior results in (Ng & Russel, 2000).

Throughout this paper, we represent the 1-norm with $|| \cdot ||_1$ and the 2-norm with $|| \cdot ||_2$.

Given a standard MDP, a policy is defined as a map $\pi : S \to A$. Given a policy π, we can define two functions.

The first is the *value function* at a state s_1 which is defined as

$$V^\pi(s_1) = \mathbb{E}\big[R(s_1) + \gamma R(\tau(s_1)) + \gamma^2 R(\tau(\tau(s_1))) + \ldots \mid \pi\big]$$

where $\tau(s)$ represents the trajectory under policy π. The second is the *Q function* which is defined as

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

The *Bellman Optimality equation* states that a policy $\pi^*(s)$ is an optimal policy for an MDP if and only if

$$\pi^*(s) \in \arg\max_{a \in A} Q^{\pi^*}(s,a), \quad s \in S$$

The **Inverse Reinforcement Learning problem** for a standard MDP is posed as the problem of finding the reward function R that generates a desired optimal policy π^* given a MDP (without reward R) and the known optimal policy π^* to be generated.

(Ng & Russel, 2000) prove that for a finite-state MDP with reward R, and for $\pi^* \equiv a_1$, the Bellman optimality equation is equivalent to the following condition:

$$(P_{a_1}(i) - P_a(i))(I - \gamma P_{a_1})^{-1} R \geq 0 \qquad \forall \; a \in A \setminus a_1, i = 1, \ldots, n \quad (2.1)$$

It can also be shown that $\pi^* \equiv a_1$ is the unique optimal policy if the above inequality is strict. We note that this condition is necessary and sufficient for the policy to be optimal for the reward.

This condition also forms the basis for the β-strict separability condition defined in (Komanduru & Honorio, 2019), which we will make use of in our construction. We reproduce the aforementioned condition here for convenience.

Definition 2.1 (**β-Strict Separability**). *Let $\beta > 0$. An inverse reinforcement learning problem (S, A, P_a, γ) with optimal policy $\pi^* \equiv a_1$ satisfies β-strict separability if and only if there exists a reward function $R^* : S \to \mathbb{R}$ that satisfies Bellman optimality strictly. More formally,*

$$\|R^*\|_1 = 1$$

and

$$(P_{a_1}(i) - P_a(i))(I - \gamma P_{a_1})^{-1} R^* \geq \beta > 0 \qquad \forall a \in A \setminus a_1, i = 1, \ldots, n$$

There are various formulations and solution methods for the standard MDP Inverse Reinforcement Learning problem such as those presented in (Ng & Russel, 2000), (Ramachandran & Amir, 2007), (Syed et al., 2008) and (Komanduru & Honorio, 2019) to list a few. Our concern in this paper is not the particular method of solution. Instead we seek to provide an information-theoretic lower bound for the sample complexity of the IRL problem. To achieve this, we use Fano's inequality (Cover & Thomas, 2006) along with the construction of an ensemble.

Throughout this paper we will use $\mathcal{F}$ to represent MDP (without reward R) problems of the specified construction as described in Section 3, the collection of which is denoted by F.

We also represent the unit sphere in $\mathbb{R}^n$ with $\mathcal{S}^{n-1}$. That is

$$\mathcal{S}^{n-1} = \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\}$$

We also utilize the concept of spherical codes in this paper. A spherical code with parameters $(n, N, \cos\theta)$ represents a set of N points (say $y_i, i = 1, \ldots, N$) on the unit sphere $\mathcal{S}^{n-1} \subset \mathbb{R}^n$ such that the angle between the unit vectors from the origin to any two distinct points is at least θ, i.e.,

$$\langle y^i, y^j \rangle \le \cos\theta, \quad i \ne j$$

where $\langle \cdot, \cdot \rangle$ represents the dot product on $\mathbb{R}^n$.

Given these preliminaries, we will now proceed to the construction of the ensemble F along with its properties.

First, in Section 3, we describe the geometric construction of the ensemble of IRL problem sets along with the resulting number of problems constructed. We start the geometric construction of the problem set by specifying the desired properties of the IRL problems in the set. This includes the presence of a common optimal policy, exclusivity of the corresponding reward with respect to other problems in the set, and β-strict separability. We then relate the transition probabilities and reward pairs of these problems to the facets (represented by $\mathcal{Y}$) of a spherical code on $\mathcal{S}^{n-2}$. We also show how the above allows for construction of rewards that are exclusively Bellman optimal for their corresponding problems within the problem set.

Next, in Section 4, we provide further analysis of the constructed problem set by providing the conditions for β-strict separability of the generated problems, the cardinality of the set and an upper bound for the KL divergence between the densities of the transition probabilities. Finally, we use these results along with Fano's inequality to come up with the sample complexity lower bound.

3. Geometric Construction

In this section we aim to construct a set of inverse reinforcement learning problems with the intention of applying Fano's inequality to obtain a lower bound for the sample complexity. The geometry of the construction of these problems provides a lower bound of the number of such problems that can be constructed which allows for a Fano's style approach to bounding the sample complexity.

Consider a set of n-state 2-action IRL problems $F = \{\mathcal{F}^i\}$ where each problem i consists of two possible actions, $a_1^i = a_1$ and a_2^i with corresponding transition probabilities $P_{a_1} \in [0,1]^{n\times n}$ and $P_{a_2^i} \in [0,1]^{n\times n}$. Let $\gamma \in (0,1)$ be fixed between all the problems. Let $R^i \in \mathbb{R}^n$ be the reward function for problem i that results in action a_1 as the optimal policy. Further, let P_{a_1}, the transition probability under the desired optimal action a_1, be fixed between the set of problems and be given as follows

$$P_{a_1} = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}$$

We construct the problem pairs $(\mathcal{F}^i, R^i)$ such that the following two relations, which follow from the Bellman Optimality condition (Eq 2.1) in matrix form, hold

$$\left(P_{a_1} - P_{a_2^i}\right)\left(I - \gamma P_{a_1}\right)^{-1} R^i \succeq \mathbf{0}$$

and

$$\left(P_{a_1} - P_{a_2^i}\right)\left(I - \gamma P_{a_1}\right)^{-1} R^j \not\succeq \mathbf{0} \quad i \ne j$$

That is, reward R^i results in a_1 being the Bellman optimal action only for its corresponding problem set $\mathcal{F}^i$. Here the notation $\succeq$ represents the entrywise $\geq$ relation. The notation $\mathbf{0}$ is used to represent a vector of zeros.

Furthermore we want to enforce β- strict separability in each problem set. That is, for each problem set $\mathcal{F}^i$, we have $||R^i||_1 = 1$ and

$$\left(P_{a_1}(j) - P_{a_2^i}(j)\right)\left(I - \gamma P_{a_1}\right)^{-1} R^i \ge \beta > 0 \qquad \forall j = 1, \ldots, n$$

We also constrain the transition probabilities of the non-optimal action $P_{a_2^i}$ such that the probabilities of transitioning from each state lie in an ε-ball around the point $\begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \ldots & \frac{1}{n} \end{bmatrix}^\top$, which also corresponds to the rows of P_{a_1} by construction, i.e.,

$$\|P_{a_1}(j) - P_a(j)\|_2 < \varepsilon, \quad \forall j = 1 \ldots n$$

We now look at the construction of such sets of problems $\mathcal{F}^i$

Let $P_a(i)$ represent the i-th row of P_a. We notice that $P_a(i)$ belongs to the following set

$$G_n^1 := \left\{ x \mid x \in \mathbb{R}^n, \sum_{i=1}^n x_i = 1 \right\}$$

Now notice that if $P_a(i), P_{a_1}(j) \in G_n^1$ then $P_{a_1}(i) - P_a(j)$ belongs to the hyperplane $H_n = \{x \in \mathbb{R}^n \mid \sum_{i=1}^n x_i = 0\}$.

We also notice that there exists an invertible rotation matrix $\Pi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that the following holds

$$\Pi(x) = \begin{bmatrix} y \\ 0 \end{bmatrix}, \quad x \in H_n,\ y \in \mathbb{R}^{n-1}$$

or alternatively

$$\text{Proj}_n(\Pi(x)) = \mathbf{0}, \quad x \in H_n$$

where Proj_n represents the projection on to the e_n dimension, i.e., projection on to the orthogonal subspace of the unit vector e_n. Here $\{e_i\}_{i=1,\dots,n}$ represent the canonical basis of $\mathbb{R}^n$.

It is also of note that none of the translation, rotation or projection operations described above change the 2-norm of $P_{a_1}(i) - P_a(j)$. Thus, through composition, we can come up with a one-to-one mapping between the points $\{P_a(i) \mid ||P_{a_1}(i) - P_a(j)||_2 < \varepsilon\}$ and the points $\{x \mid x \in \mathbb{R}^{n-1},\ ||x||_2 < \varepsilon\}$

Now we describe a construction of each problem $\mathcal{F}^i$ by specifying the second transition matrix $P_{a_2^i}$ through $P_{a_1} - P_{a_2^i}$ and the corresponding reward R^i by constructing their equivalents in $\mathbb{R}^{n-1}$.

Consider the unit sphere in $\mathbb{R}^{n-1}$ given by $\mathcal{S}^{n-2}$. Consider the set of points defining a spherical code $(n-1, N, \cos\theta)$ with minimum angle θ on $\mathcal{S}^{n-2}$ such that N is maximal. From (Jenssen et al., 2018) we have the result

$$N \geq (1 + o(1))\sqrt{2\pi(n-1)}\frac{\cos\theta}{\sin^{n-2}\theta} \tag{3.1}$$

To form the transition probabilities and the corresponding rewards we wish to consider the facets $\mathcal{Y}$ of the simplicial polytope formed by the spherical code. Since the convex polytope formed by the maximal spherical code can be simplicially decomposed (Edmonds, 1970), the resulting simplicial decomposition can be used to form a simplicial polytope with the same vertices with the condition that the interiors of the facets, a $n-2$ simplex, are pairwise disjoint. The number of facets of such a simplicial polytope are lower bounded by (Barnette, 1971)

$$|\mathcal{Y}| \geq (n-2)N - (n-1)(n-3) \tag{3.2}$$

We denote the elements of $\mathcal{Y}$ as $\mathcal{Y}^i$, such that any pair of points $y \in \mathcal{Y}^i$ are neighbors with respect to the spherical code and

$$|\mathcal{Y}^i \cap \mathcal{Y}^j| \leq n-2, \quad i \neq j$$

To form the set of problems, we consider the pairwise disjoint cones formed by the vertices in each $\mathcal{Y}^i$ and the origin. The corresponding rewards are formed from the centroids of each $\mathcal{Y}^i$. The resulting geometry from the disjoint interiors of the $\mathcal{Y}^i$ ensures that each reward function only results in a_1 as the optimal action for the corresponding problem.

For every $\mathcal{Y}^i$ we denote the centroid of $\mathcal{Y}^i$ as follows

$$\bar{y}^i = \frac{1}{n-1}\sum_{y \in \mathcal{Y}^i} y \tag{3.3}$$

We also consider the following hyperplanes in $\mathbb{R}^{n-1}$ passing through the elements of the leave-one-out set of $\mathcal{Y}^i$ and the origin as defined by the corresponding normal vectors p_j^i. We further impose the constraint that the norm of each p_j^i is constant across all i and j. Formally, $p_j^i \in \mathbb{R}^{n-1}$ is defined by the following conditions:

$$p_j^{i\top} y_k^i = 0, \quad j \neq k, \quad 1 \leq k \leq n$$

$$p_j^{i\top} \bar{y}^i > 0$$

$$||p_j^i||_2 = \varepsilon$$

Notice that p_j^i is an element of the null space of

$$y^i := \begin{bmatrix} y_1^i & \dots & y_{j-1}^i & y_{j+1}^i & \dots & y_n^i & \mathbf{0} \end{bmatrix}^\top$$

where the exponent $\top$ represents transpose, such that $\bar{y}^i$ lies in the interior of the cone formed by the hyperplanes. We also notice that as a result of the construction of the set $\mathcal{Y}^i$ and the hyperplanes defined by p_j^i

$$\begin{bmatrix} p_1^{i\top} \\ \vdots \\ p_{n-1}^{i\top} \end{bmatrix} \bar{y}^k \succeq \mathbf{0} \iff i = k$$

If $\hat{y}^i$ is the unit vector in direction of $\bar{y}^i$, then it follows that

$$\begin{bmatrix} p_1^{i\top} \\ \vdots \\ p_{n-1}^{i\top} \end{bmatrix} \hat{y}^k \succeq \mathbf{0} \iff i = k$$

and also since the centroid $\bar{y}^i$ of points on the sphere lies in the interior of the sphere since the sphere is a convex shape, while $\hat{y}^i$ lies on the surface of the sphere. An example of such a spherical code formation along with visualization of the hyperplanes and centroid involved for a single facet is provided in Figure 1.

Lemma 3.1. *Let p_j^i and $\bar{y}^i$ be as described above. Let θ be the minimum angle between any pair of points $y_j^i \in \mathcal{Y}$. Let $\hat{y}^i = \bar{y}^i/||\bar{y}^i||_2$ be the unit vector along $\bar{y}^i$. Then we have*

$$\frac{p_j^{i\top}\hat{y}^i}{||p_j^i||_2} \geq \frac{2\sin\theta/2}{\sqrt{2(n-2)(1+(n-2)\cos\theta)}}$$

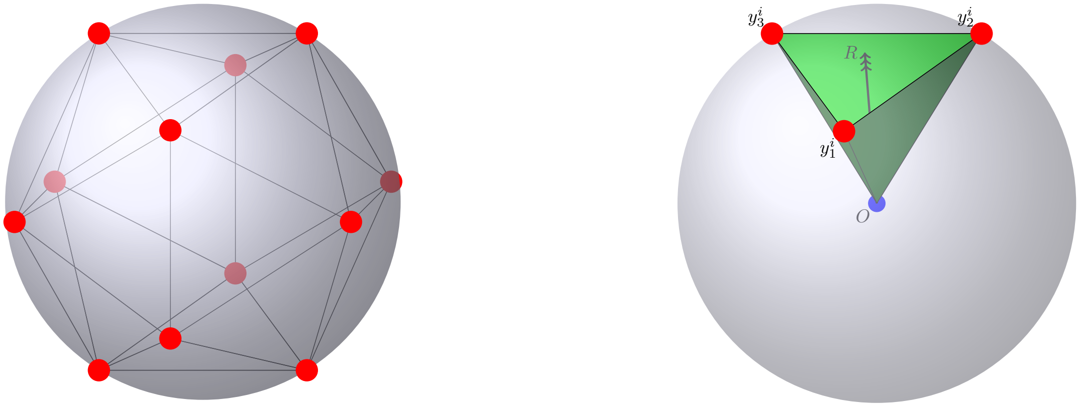

Figure 1. **Left:** An example graphical visualization a spherical code (kissing number arrangement) on the sphere $\mathcal{S}^2$. The arrangement forms a regular icosahedron. The vertices (elements of $\mathcal{Y}$) are marked in red and the corresponding ridges of the polytope are outlined in black. **Right:** An example graphical visualization of a facet ($\mathcal{Y}^i$) from the configuration on the left used in the formation of a single problem. The hyperplanes formed from the leave-one-out set of $\mathcal{Y}^i$ and the origin are shown as the green planes (with normals p_j^i) passing through the origin (blue). The reward direction corresponding to $\hat{y}^i$ is shown with the black arrow labeled R

Proof. We notice that the distance of $\bar{y}^i$ to each hyperplane is greater than or equal to the minimum radius of the ball inside the $n-2$ simplex formed by $\mathcal{Y}^i$. We can show this from the inner product of p_j^i and $\bar{y}^i$ as well.

Consider $p_j^{i\top}\bar{y}^i$, using the expression for the centroid given in Equation 3.3, we get

$$p_j^{i\top}\bar{y}^i = p_j^{i\top}\left[\frac{1}{n-1}\sum_{k=1}^{n-1} y_k^i\right]$$

since we know from the construction of p_j^i that $p^{i\top}y_k^i = 0,\ j \neq k$. Therefore we have,

$$p_j^{i\top}\bar{y}^i = \frac{1}{n-1}p_j^{i\top}y_j^i$$

We also notice that the projection of the ray from the origin to the vertex of the $n-2$ simplex, y_j^i, orthogonal to the hyperplane forming the opposing side of the $n-2$ simplex, the normal vector of which is p_j^i, is at least the height of the $n-2$ simplex which is given by $s\sqrt{\frac{n-1}{2(n-2)}}$, where $s \geq 2\sin(\theta/2)$ is the edge length of the simplex $\mathcal{Y}^i$. That is

$$\frac{p_j^{i\top}y_j^i}{\|p_j^i\|_2} \geq 2\sin(\theta/2)\sqrt{\frac{n-1}{2(n-2)}}$$

$$\implies \frac{p_j^{i\top}\bar{y}^i}{\|p_j^i\|_2} \geq \frac{1}{n-1}2\sin(\theta/2)\sqrt{\frac{n-1}{2(n-2)}}$$

$$= \frac{2\sin\theta/2}{\sqrt{2(n-2)(n-1)}}$$

Now we also have from Equation 3.3 and the fact that y_k^i are points on a maximal spherical code with minimum angle θ. As a result $\langle y_j^i, y_k^i\rangle \leq \cos\theta$ and thus

$$\begin{aligned}
||\bar{y}^i||_2^2 &= \left\|\frac{1}{n-1}\sum_{k=1}^{n-1} y_k^i\right\|_2^2 \\
&= \frac{1}{(n-1)^2}\left[\sum_{k=1}^{n-1}\|y_k^i\|_2^2 + 2\sum_{1\leq j<k\leq n-1}\langle y_j^i, y_k^i\rangle\right] \\
&\leq \frac{1}{(n-1)^2}\left[n-1+2\frac{(n-1)(n-2)}{2}\cos\theta\right] \\
&= \frac{1+(n-2)\cos\theta}{n-1}
\end{aligned}$$

$$\implies \frac{p_j^{i\top}\hat{y}^i}{||p_j^i||_2} =$$

$$\frac{p_j^{i\top}\bar{y}^i}{||p_j^i||_2||\bar{y}^i||_2} \geq \frac{2\sin\theta/2}{\sqrt{2(n-2)(1+(n-2)\cos\theta)}}$$

□

We will use the p_j^i's to construct the matrix $P_{a_1} - P_{a_2^i}$ and $\bar{y}^i$ to form the corresponding reward R^i through the following transformations

$$P_{a_1}(j) - P_{a_2^i}(j) = \left(\Pi^\top\begin{bmatrix}p_j^i\\0\end{bmatrix}\right)^\top, \quad 1 \leq j \leq n-1$$

$$P_{a_1}(n) - P_{a_2^i}(n) = \left(\Pi^\top\begin{bmatrix}p_1^i\\0\end{bmatrix}\right)^\top$$

$$R^i = \frac{(I - \gamma P_{a_1})\,\Pi^\top \begin{bmatrix}\hat{y}^i \\ 0\end{bmatrix}}{\left\|(I - \gamma P_{a_1})\,\Pi^\top \begin{bmatrix}\hat{y}^i \\ 0\end{bmatrix}\right\|_1} \tag{3.4}$$

Since there are only $n-1$ vectors p_j^i while $P_{a_1} - P_{a_2^i}$ is an $n \times n$ matrix, we use p_1^i for the final row so that the interior of the transformed $n-2$ simplex in $\mathbb{R}^{n-1}$ remains the same.

Notice that from the above construction, we have the following conditions being met

$$\left(P_{a_1} - P_{a_2^i}\right)(I - \gamma P_{a_1})^{-1} R^i \succeq 0$$

and

$$\left(P_{a_1} - P_{a_2^i}\right)(I - \gamma P_{a_1})^{-1} R^j \not\succeq 0 \quad i \neq j$$

Next we will use the β-strict separability condition to form a relation between β, $\varepsilon = \|p_j^i\|_2$ and the angle θ that generates the spherical code.

4. Analysis of Geometric Construction and Sample Complexity of IRL

The previous section described the geometric construction of the ensemble of IRL problems. We now analyze the geometric construction described in the previous section in order to find the conditions for β-strict separability of the generated problems. We also derive bounds on the cardinality of the problem ensemble constructed as well as the KL divergence between the trajectories generated by different problems in the ensemble. Finally we use these results along with Fano's inequality (Cover & Thomas, 2006), to derive an information-theoretic lower bound for the sample complexity of the IRL problem.

Lemma 4.1. *Consider a IRL problem and reward pair constructed as in Equation 3.4. If the following relation between β, ε and θ holds*

$$\sin^2\frac{\theta}{2} = \frac{n(n-1)(n-2)\beta^2}{2\varepsilon^2 + 2n(n-2)^2\beta^2}$$

then the problem is β-strict separable.

Proof. Note that

$$(P_{a_1}(j) - P_{a_2^i}(j))\,(I - \gamma P_{a_1})^{-1} R^i =$$

$$\left(\Pi^\top \begin{bmatrix}p_j^i \\ 0\end{bmatrix}\right)^\top (I - \gamma P_{a_1})^{-1} \frac{(I - \gamma P_{a_1})\,\Pi^\top \begin{bmatrix}\hat{y}^i \\ 0\end{bmatrix}}{\left\|(I - \gamma P_{a_1})\,\Pi^\top \begin{bmatrix}\hat{y}^i \\ 0\end{bmatrix}\right\|_1}$$

$$= \frac{\begin{bmatrix}p_j^i \\ 0\end{bmatrix}^\top \begin{bmatrix}\hat{y}^i \\ 0\end{bmatrix}}{\left\|(I - \gamma P_{a_1})\,\Pi^\top \begin{bmatrix}\hat{y}^i \\ 0\end{bmatrix}\right\|_1} \geq \frac{\begin{bmatrix}p_j^i \\ 0\end{bmatrix}^\top \begin{bmatrix}\hat{y}^i \\ 0\end{bmatrix}}{\max_y \left\|(I - \gamma P_{a_1})\,\Pi^\top \begin{bmatrix}y \\ 0\end{bmatrix}\right\|_1}$$

$$\geq \beta, \quad ||y||_2 = 1$$

Notice that since Π is a rotation matrix $||\Pi||_2 = 1$. Additionally from the construction of $(I - \gamma P_{a_1})$, we have $||\,(I - \gamma P_{a_1})\,||_2 = 1$. Now consider

$$\max_{y,\ ||y||_2 = 1} \left\|(I - \gamma P_{a_1})\,\Pi^\top \begin{bmatrix}y \\ 0\end{bmatrix}\right\|_1$$

The above equation represents the norm of $(I - \gamma P_{a_1})\,\Pi^\top$ as an operator from $(\mathbb{R}^n, \|\cdot\|_2)$ to $(\mathbb{R}^n, \|\cdot\|_1)$. By duality, this is the same as the norm of the adjoint $\Pi\,(I - \gamma P_{a_1})^\top$ as an operator from $(\mathbb{R}^n, \|\cdot\|_\infty)$ to $(\mathbb{R}^n, \|\cdot\|_2)$. Using this we get

$$\max_{||y||_2 = 1} \left\|(I - \gamma P_{a_1})\,\Pi^\top \begin{bmatrix}y \\ 0\end{bmatrix}\right\|_1$$

$$= \max_{||v||_\infty = 1} \left\|\Pi\,(I - \gamma P_{a_1})^\top v\right\|_2 = \sqrt{n}$$

This gives us

$$(P_{a_1}(j) - P_{a_2^i}(j))\,(I - \gamma P_{a_1})^{-1} R^i \geq \frac{p_j^{i\top}\hat{y}^i}{\sqrt{n}} = \frac{\varepsilon p_j^{i\top}\hat{y}^i}{||p_j^i||_2\sqrt{n}}$$

Substituting the result of Lemma 3.1, we get

$$\frac{\varepsilon p_j^{i\top}\hat{y}^i}{||p_j^i||_2\sqrt{n}} \geq \frac{2\varepsilon\sin\theta/2}{\sqrt{2n(n-2)(1+(n-2)\cos\theta)}}$$

Now substituting

$$\sin\frac{\theta}{2} = \sqrt{\frac{n(n-1)(n-2)\beta^2}{2\varepsilon^2 + 2n(n-2)^2\beta^2}}$$

and using the trigonometric formula $\cos\theta = 1 - 2\sin^2\theta/2$ we get

$$(P_{a_1}(j) - P_{a_2^i}(j))\,(I - \gamma P_{a_1})^{-1} R^i \geq \beta$$

Since $\|R^i\|_1 = 1$ by construction, the problem is β-strict separable □

We now proceed to use this construction to find a lower bound for the number of such probability matrix - reward function pairs as well as an upper bound on the KL divergence between the corresponding probability matrices.

Theorem 4.1. *Given a construction of β-strict separable IRL problems and reward function pairs in Equation 3.4, where the angle θ of the spherical code used to generate the problems satisfies Lemma 4.1, the minimum number of such problem-reward pairs for a given n, ε and β is*

$$|\mathcal{Y}| \geq (n-2)\left((1+o(1))\sqrt{2\pi(n-1)}\frac{\varepsilon^2 - n(n-2)\beta^2}{\varepsilon^2 + n(n-2)^2\beta^2} \times \left(\frac{\varepsilon^2 + n(n-2)^2\beta}{\sqrt{n(n-1)(n-2)\left(2\varepsilon^2 + n(n-2)(n-3)\beta^2\right)}} \right)^{n-2} \right) - (n-1)(n-3)$$

Proof. We start with the result of Lemma 4.1.

$$\sin\frac{\theta}{2} = \sqrt{\frac{n(n-1)(n-2)\beta^2}{2\varepsilon^2 + 2n(n-2)^2\beta^2}}$$

From the trigonometric identities $\cos^2\frac{\theta}{2} + \sin^2\frac{\theta}{2} = 1$, $\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}$ and $\cos\theta = 1 - 2\sin^2\frac{\theta}{2}$ we get

$$\sin\theta = \frac{\beta\sqrt{n(n-1)(n-2)\left(2\varepsilon^2 + n(n-2)(n-3)\beta^2\right)}}{\varepsilon^2 + n(n-2)^2\beta^2}$$

$$\cos\theta = \frac{\varepsilon^2 - n(n-2)\beta^2}{\varepsilon^2 + n(n-2)^2\beta^2}$$

Substituting the above into Equation 3.1 and subsequently into Equation 3.2 we get

$$|\mathcal{Y}| \geq (n-2)\left((1+o(1))\sqrt{2\pi(n-1)}\frac{\varepsilon^2 - n(n-2)\beta^2}{\varepsilon^2 + n(n-2)^2\beta^2} \times \left(\frac{\varepsilon^2 + n(n-2)^2\beta^2}{\beta\sqrt{n(n-1)(n-2)\left(2\varepsilon^2 + n(n-2)(n-3)\beta^2\right)}} \right)^{n-2} \right) - (n-1)(n-3)$$

□

It is of note that the bounds of ε from the β-strict separability condition as well as the condition that the of the minimum ball contained in the probability simplex we have the following result

Lemma 4.2. *Consider a IRL problem and reward pair constructed as in Equation 3.4 from an $(n-1, N, \cos\theta)$ spherical code such that Lemma 4.1 holds. Then*

$$\frac{1}{\sqrt{(n-1)(n)}} \geq \varepsilon \geq \sqrt{n-2}\beta \tag{4.1}$$

and the lower bound corresponds to a $n-1$ simplex.

Proof. The upper bound

$$\frac{1}{\sqrt{(n-1)(n)}} \geq \varepsilon$$

is straightforwardly obtained from the condition that $P_{a_2^i}(j)$ is contained in a ball of radius ε around $P_{a_1}(j)$ which is located at the center of the probability simplex. The bound represents the maximum radius ball that fits within the $n-1$ probability simplex with side length $\sqrt{2}$.

The minimum can be found by noticing that the convex $n-1$ dimensional polytope in $\mathbb{R}^{n-1}$ formed by the maximal spherical code must be at least $n-1$-vertex-connected by Balinski's theorem (Balinski et al., 1961). Thus the minimum number of vertices (which give the minimum possible $\cos\theta$) must be n. This minimum is achieved in the form of an $n-1$ simplex with vertices on $\mathcal{S}^{n-2}$ with $\theta = \cos^{-1}\frac{-1}{n-1}$

From this minimum case we have

$$\cos\theta = \frac{\varepsilon^2 - n(n-2)\beta^2}{\varepsilon^2 + n(n-2)^2\beta^2} \geq -\frac{1}{n-1}$$

rearranging and simplifying gives

$$\implies \varepsilon^2 \geq (n-2)\beta^2$$

The solution to which gives the lower bound of the lemma

□

Now we apply the result from Equation (3) of (Borade & Zheng, 2008) to bound the KL divergence D of two columns of the transition probability matrices P and Q from two different problems where the columns lie within a ball of radius ε around $P_{a1}(i) = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \ldots \frac{1}{n} \end{bmatrix}$

Lemma 4.3. *Let $P(i)$ and $Q(j)$ be the columns i and j of the transition probability matrices of such problem-reward pairs as described above constructed using Equation 3.4. Then*

$$D(P(i)||Q(j)) \leq \frac{2\varepsilon^2 n}{1 - n\varepsilon}$$

Theorem 4.2. *Let P and Q be two transition probability matrices of such problem-reward pairs as described above constructed using Equation 3.4. Consider the m-length trajectories drawn from each transition probability and let $p^{(m)}$ and $q^{(m)}$ represent the corresponding probability distributions of the trajectory m-tuples. Furthermore, let the trajectories start in each state with equal probability. Then we have*

$$D(p^{(m)}||q^{(m)}) \leq (m-1)\frac{2\varepsilon^2 n}{1 - n\varepsilon}$$

Proof. We use an approach based on Theorem 1 of (Rached et al., 2004). Notice that since the columns of P and Q

are within the ε ball around $\begin{bmatrix}\frac{1}{n} & \frac{1}{n} & \dots \frac{1}{n}\end{bmatrix}$ and $\varepsilon < \frac{1}{n}$, none of the elements of P and Q are 0. Thus we have P is absolutely continuous with respect to Q.

Now we have

$$D(p^{(m)}||q^{(m)}) = p(I+P+P^2+\ldots+P^{m-2})V+D(p||q)$$

Where $p = q = \begin{bmatrix}\frac{1}{n} & \frac{1}{n} & \dots \frac{1}{n}\end{bmatrix}$ represent the initial distributions of states for the trajectories and V is given by

$$V = \begin{bmatrix} D(P(1)||Q(1)) \\ D(P(2)||Q(2)) \\ \vdots \\ D(P(n)||Q(n)) \end{bmatrix} \leq \mathbb{1}\frac{2\varepsilon^2 n}{1-n\varepsilon} \quad \text{(by Lemma 4.3)}$$

Also notice that $D(p||q) = 0$. By construction of P and Q, we can write P as $P = P_{a_1} + \varepsilon U$. Since P is also a transition probability matrix, U will be a matrix whose columns are unit vectors whose components sum to 0. That is $\mathbb{1}^T U = \begin{bmatrix}1 & 1 & \dots & 1\end{bmatrix} U = 0^T \implies P_{a_1} U = 0I$

Substituting into the expression for $D(p^{(m)}||q^{(m)})$, and simplifying, we get

$$D(p^{(m)}||q^{(m)}) = \begin{bmatrix}\frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n}\end{bmatrix} \Big(I + P_{a_1} + \varepsilon U + \\ (P_{a_1}+\varepsilon U)^2 + \ldots + (P_{a_1}+\varepsilon U)^{m-2} \Big) V$$

Since every term with U will either be multiplied with P_{a_1} or $\begin{bmatrix}\frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n}\end{bmatrix}$ from the left, all of the terms with U will end up being 0.

$$\begin{aligned} \implies D(p^{(m)}||q^{(m)}) &= \begin{bmatrix}\frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n}\end{bmatrix} \\ &\quad (I + (m-2)P_{a_1})\, V \\ &\leq \begin{bmatrix}\frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n}\end{bmatrix} (I + (m-2)P_{a_1})\, \mathbb{1}\frac{2\varepsilon^2 n}{1-n\varepsilon} \\ &= \begin{bmatrix}\frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n}\end{bmatrix} \mathbb{1}(1+(m-2))\frac{2\varepsilon^2 n}{1-n\varepsilon} \\ &= (m-1)\frac{2\varepsilon^2 n}{1-n\varepsilon} \end{aligned}$$

□

Notice that in the case of multiple actions, if we assume the actions are chosen randomly, thereby creating an extended P and Q matrix for state transition of dimensions $nk \times nk$, the KL bound as well as the result of Theorem 4.2 do not change.

We will now use Fano's inequality to form the lower bound.

Theorem 4.3. *Consider the set of constructed problem reward pairs $F = \{\mathcal{F}^i, R^i\}$ constructed as described above. Let $\mathcal{F}$ be uniform on F. Let Z represent the m sample trajectory generated from $\mathcal{F}$ and let $\hat{\mathcal{F}}$ be an estimator of $\mathcal{F}$ from Z. Then for any Markov chain $\mathcal{F} \to Z \to \hat{\mathcal{F}}$, we have*

$$\mathbb{P}(\hat{\mathcal{F}} \neq \mathcal{F}) \geq 1 - \frac{(m-1)\frac{2\varepsilon^2 n}{1-n\varepsilon} + \log 2}{\log \eta}$$

where

$$\eta = (n-2)\left(\sqrt{2\pi(n-1)}\frac{\varepsilon^2 - n(n-2)\beta^2}{\varepsilon^2 + n(n-2)^2\beta^2} \times \\ \left(\frac{\varepsilon^2 + n(n-2)^2\beta^2}{\beta\sqrt{n(n-1)(n-2)\left(2\varepsilon^2 + n(n-2)(n-3)\beta^2\right)}}\right)^{n-2}\right) \\ - (n-1)(n-3)$$

Proof. We start with Fano's inequality

$$\mathbb{P}(\hat{\mathcal{F}} \neq \mathcal{F}) \geq 1 - \frac{I(\mathcal{F};Z) + \log 2}{\log |F|}$$

Notice that

$$\begin{aligned} I(\mathcal{F};Z) &\leq \max_{\mathcal{F},\mathcal{F}'} D\left(P_{Z|\mathcal{F}}(\cdot|\mathcal{F})||P_{Z|\mathcal{F}}(\cdot|\mathcal{F}')\right) \\ &= \max_{p^{(m)},q^{(m)}} D(p^{(m)}||q^{(m)}) \\ &\leq (m-1)\frac{2\varepsilon^2 n}{1-n\varepsilon} \quad \text{(by Theorem 4.2)} \end{aligned}$$

We also know that the number of such problem-reward pairs $|F|$ is the number of facets $|\mathcal{Y}|$. From Theorem 4.1, we have

$$|F| \geq (n-2)\left(\sqrt{2\pi(n-1)}\frac{\varepsilon^2 - n(n-2)\beta^2}{\varepsilon^2 + n(n-2)^2\beta^2} \times \\ \left(\frac{\varepsilon^2 + n(n-2)^2\beta^2}{\beta\sqrt{n(n-1)(n-2)\left(2\varepsilon^2 + n(n-2)(n-3)\beta^2\right)}}\right)^{n-2}\right) \\ - (n-1)(n-3)$$

Substituting these results in Fano's inequality gives us the result of the Theorem. □

Corollary 4.1. *Consider the set of constructed problem reward pairs $F = \{\mathcal{F}^i, R^i\}$ constructed as described above. Let $\mathcal{F}$ be uniform on F. Let n be large and consider the case*

$$\frac{1}{\sqrt{2n(n-1)}} = \varepsilon = \sqrt{n-2}\beta$$

Let Z represent the m sample trajectory generated from $\mathcal{F}$ and let $\hat{\mathcal{F}}$ be an estimator of $\mathcal{F}$ from Z with

$$m \leq (n-1)(0.5\log n - \log 2)\left(1 - \sqrt{\frac{n}{2(n-1)}}\right) + 1$$

Then for any Markov chain $\mathcal{F} \rightarrow Z \rightarrow \hat{\mathcal{F}}$, we have

$$\mathbb{P}(\hat{\mathcal{F}} \neq \mathcal{F}) \geq 0.5$$

This result gives us the lower bound for the sample complexity in the case of large n on the order of $O(n\log n)$.

A similar result can be found for the case where just the lower bound of $\varepsilon \geq \sqrt{n-2}\beta$ is satisfied, which from Lemma 4.2 results in the case of an $n-1$ simplex.

Corollary 4.2. *Consider the set of constructed problem reward pairs $F = \left\{\mathcal{F}^i, R^i\right\}$ constructed as described above. Let $\mathcal{F}$ be uniform on F. Let n be large and consider the case*

$$\frac{1}{\sqrt{2n(n-1)}} \geq \varepsilon = \sqrt{n-2}\beta$$

Let Z represent the m sample trajectory generated from $\mathcal{F}$ and let $\hat{\mathcal{F}}$ be an estimator of $\mathcal{F}$ from Z with

$$m \leq \frac{(0.5\log n - \log 2)}{2(n-2)n\beta^2}\left(1 - n\sqrt{n-2}\beta\right) + 1$$

Then for any Markov chain $\mathcal{F} \rightarrow Z \rightarrow \hat{\mathcal{F}}$, we have

$$\mathbb{P}(\hat{\mathcal{F}} \neq \mathcal{F}) \geq 0.5$$

This result gives us the lower bound for the sample complexity in the case of large n on the order of $O(\frac{\log n}{n^2\beta^2})$. The results of our experimental validation of this bound for different values of n and various solution methods is described in Appendix B and can be seen in Figure B.1 and Figure B.2. The bound of $O(\frac{\log n}{n^2\beta^2})$ from Corollary 4.2 approaches the bound of $O(n\log n)$ from Corollary 4.1 as β approaches the upper bound of $\frac{1}{\sqrt{2n(n-2)(n-1)}}$. It is also important to note that the more general lower bound is the one provided in Theorem 4.3.

5. Discussion

A lower bound of sample complexity is a statement representing a high probability of failing to recover the correct solution in the case where sufficient samples are not provided. The results presented in this paper show, by construction, the existence of an entire family of problem sets for any number of dimensions where problems within each set have distinct solutions that are incompatible with other members of the set. Rotating and perturbing the spherical codes on $\mathcal{S}^{n-2}$ can result in a different set of problems, resulting in this family of problem sets being "dense". The interpretation of the lower bound result is that with very little knowledge of the underlying problem, there is a high probability of not recovering the correct solution without a sufficient number of samples since, independent of the algorithm used to solve, one will not be able to correctly distinguish the "true" problem among the elements of the corresponding problem set.

Although there is a gap in the bound of $O(\frac{\log n}{n^2\beta^2})$ from Corollary 4.2 with respect to the $O(\frac{n^2\log n}{\beta^2})$ of (Komanduru & Honorio, 2019), we note that non-sparse $\frac{\log n}{\beta^2}$ part of the bound is captured by our lower bound estimate. An estimation of the lower bound in the sparse case may further help reduce the gap between the lower bound and the upper bound of the sample complexity of Inverse Reinforcement Learning.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning*, New York, NY, USA, 2004. ACM. doi: 10.1145/1015330.1015430.

Balinski, M. L. et al. On the graph structure of convex polyhedra in n-space. *Pacific Journal of Mathematics*, 11 (2):431–434, 1961.

Barnette, D. W. The minimum number of vertices of a simple polytope. *Israel Journal of Mathematics*, 10(1): 121–125, 1971.

Borade, S. and Zheng, L. Euclidean information theory. In *2008 IEEE International Zurich Seminar on Communications*, pp. 14–17. IEEE, 2008.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2006.

Dvijotham, K. and Todorov, E. Inverse optimal control with linearly-solvable MDPs. In *International Conference on Machine Learning*, pp. 335–342, 2010.

Edmonds, A. L. Simplicial decompositions of convex polytopes. *Pi Mu Epsilon Journal*, 5(3):124–128, 1970.

Jenssen, M., Joos, F., and Perkins, W. On kissing numbers and spherical codes in high dimensions. *Advances in Mathematics*, 335:307–321, 2018.

Komanduru, A. and Honorio, J. On the correctness and sample complexity of inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 2019.

Levine, S., Popovic, Z., and Koltun, V. Nonlinear inverse reinforcement learning with Gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 19–27, 2011.

Neu, G. and Szepesvári, C. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Conference in Uncertainty in Artificial Intelligence*, pp. 295–302. AUAI Press, 2007.

Ng, A. Y. and Russel, S. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 663 – 670, 2000.

Rached, Z., Alajaji, F., and Campbell, L. L. The Kullback-Leibler divergence rate between Markov sources. *IEEE Transactions on Information Theory*, 50(5):917–921, 2004.

Ramachandran, D. and Amir, E. Bayesian inverse reinforcement learning. *International Joint Conference on Artificial Intelligence*, 51(61801):1–4, 2007.

Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. A. Maximum margin planning. In *International Conference on Machine Learning*, pp. 729–736. ACM, 2006.

Santhanam, N. P. and Wainwright, M. J. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

Syed, U., Bowling, M., and Schapire, R. Apprenticeship learning using linear programming. In *International Conference on Machine Learning*, pp. 1032–1039. ACM, 2008.

Tandon, R., Shanmugam, K., Ravikumar, P. K., and Dimakis, A. G. On the information theoretic limits of learning Ising models. *Advances in Neural Information Processing Systems*, 27:2303–2311, 2014.

Wang, W., Wainwright, M. J., and Ramchandran, K. Information-theoretic bounds on model selection for Gaussian Markov random fields. In *2010 IEEE International Symposium on Information Theory*, pp. 1373–1377. IEEE, 2010.

Ziebart, B. D., Maas, A., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. *AAAI Conference on Artificial Intelligence*, 2008.