# On the Correctness and Sample Complexity of Inverse Reinforcement Learning

Abi Komanduru Jean Honorio

*akomandu@purdue.edu* *jhonorio@purdue.edu*

Purdue University, West Lafayette - IN, 47907.

## INTRODUCTION

Inverse reinforcement learning (IRL) is the problem of finding a reward function that generates a given optimal policy for a given Markov Decision Process. Often, in situations including apprenticeship learning, the reward function is unknown but optimal policy can be observed through the actions of an *expert*. It is well known that such a reward function is not necessarily unique.

Previous approaches include: linear programming [1], Hybrid IRL [2], Maximum Margin Planning [3], Multiplicative Weights for Apprenticeship Learning [4] and Bayesian estimation [5]. Linear MDP approaches include Maximum Entropy IRL [6] and Gaussian Process IRL [7]

**Our contributions:**

- Algorithmic-independent geometric analysis of the IRL problem with finite states and actions.
- We show a sample complexity of $O(d^2 \log(nk))$ for $n$ states and $k$ actions and transition probability matrices with at most $d$ nonzeros per row, to recover a reward function that satisfies Bellman's optimality condition with respect to the true transition probabilities.

## PRELIMINARIES

The formulation of the IRL problem is based on a standard Markov Decision Process (MDP) $(S, A, \{P_{sa}\}, \gamma, R)$, 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 $a$. $P_a$'s are right stochastic.
- $\gamma \in [0,1]$ is the discount factor.
- $R : S \to \mathbb{R}$ is the reinforcement or reward function to be determined.

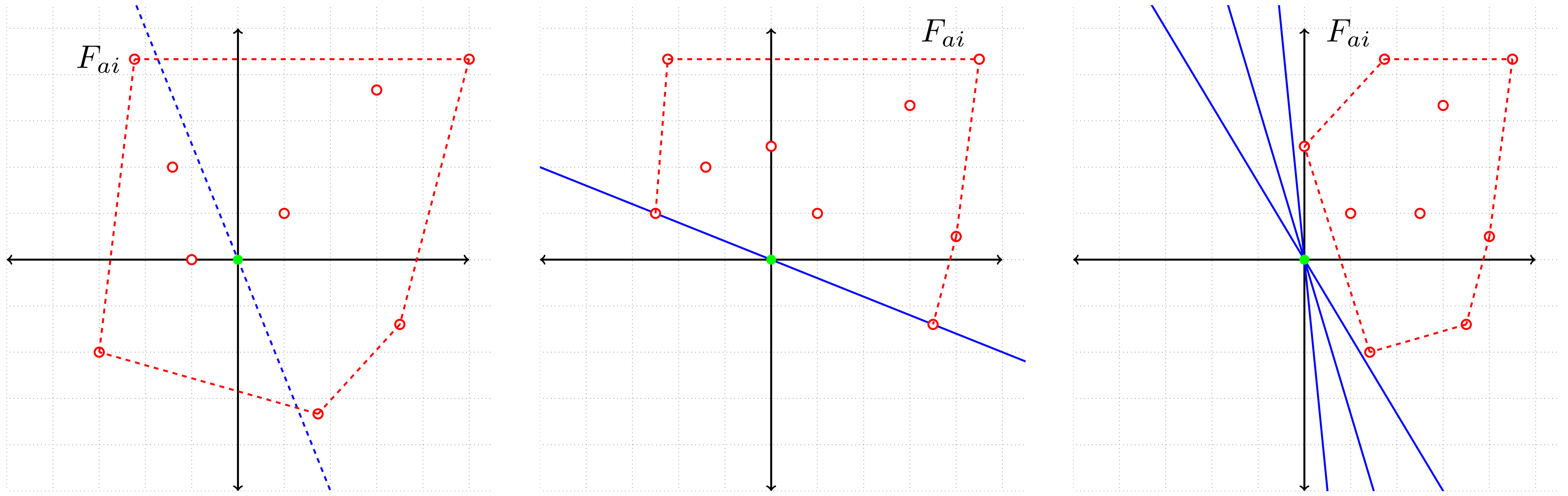
The **Bellman optimality equation** is equivalent to the following condition:

$$F_{ai} \equiv (P_{a_1}(i) - P_a(i))(I - \gamma P_{a_1})^{-1} R \geq 0 \quad \forall i = 1, \ldots, n;\ a \neq a_1$$

An inverse reinforcement learning problem $\{S, A, P_a, \gamma\}$ satisfies $\beta$**-strict separability** if and only if there exists a $\{\beta, R^*\}$ such that

$$\|R^*\|_1 = 1 \quad \text{and} \quad F_{ai}^T R^* \geq \beta > 0 \quad \forall a \in A \setminus a_1, i = 1, \ldots, n$$

## GEOMETRIC INTERPRETATION




The problem of Inverse Reinforcement Learning, then is equivalent to the problem of **finding such a separating hyperplane passing through the origin** for the points $\{F_{ai}\}$. There is an $R$ for which the policy $\pi = a_1$ is strictly optimal iff there exists a hyperplane for which all the points $\{F_{ai}\}$ are strictly on one side.

**Optimization Problem – L1 SVM Formulation**

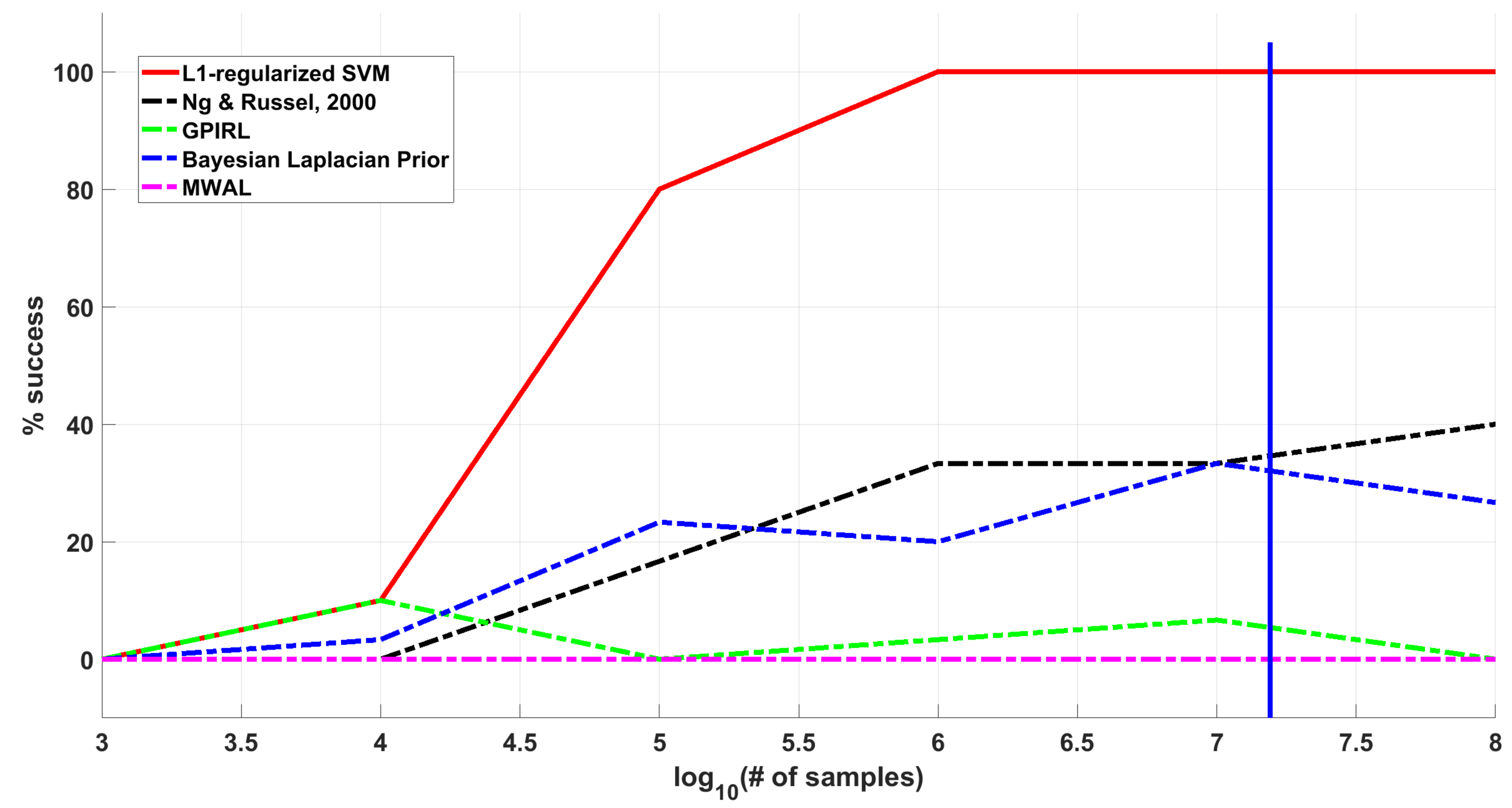
$$\begin{aligned} &\underset{R}{\text{minimize}}\ \|R\|_1 \\ &\text{subject to } \hat{F}_{ai}^T R \geq 1 \quad \forall a \in A \setminus a_1 \quad i = 1, \ldots, n \end{aligned}$$

## RESULTS AND VALIDATION

**Theorem 1 (Main Result)** *Let $\{S, A, P_a, \gamma\}$ be an inverse reinforcement learning problem that is $\beta$- strictly separable. Let the transition probability matrices $P_a$ have at most $d \in \{1, \ldots, n\}$ non-zero elements per row. Let every state be reachable from the starting state in one step with probability at least $\alpha$. Let $\hat{R}$ be the solution to the optimization problem with $\hat{F}_{ai}$ with transition probability matrices $\hat{P}_a$ that are maximum likelihood estimates of $P_a$ formed from m samples where*

$$m \geq \frac{64}{\alpha \beta^2} \left( \frac{(d-1)\gamma + 1}{(1-\gamma)^2} \right)^2 \log \frac{4nk}{\delta}$$

*Then with probability at least $(1-\delta)$, we have $F_{ai}^T \hat{R} \geq 0 \quad \forall a \in A \setminus a_1, i = 1, \ldots, n$.*




Empirical probability of success versus number of samples for an inverse reinforcement learning problem performed with $n = 7$ states and $k = 7$ actions using both our L1-regularized support vector machine formulation, the linear programming formulation proposed in [1], Multiplicative Weights for Apprenticeship Learning [4] , Bayesian IRL with Laplacian prior [5] and Gaussian Process IRL [7]. The vertical blue line represents the sample complexity for our method, as stated the theorem

## DISCUSSION

The result of the theorem shows that the number of samples required to solve a $\beta$-strict separable inverse reinforcement learning problem and obtain a reward that generates the desired optimal policy is on the order of $m \in O\left(\frac{n^2}{\beta^2} \log(nk)\right)$ or $m \in O\left(\frac{d^2}{\beta^2} \log(nk)\right)$ in the sparse case.

In practical applications, however, it may be difficult to determine if an inverse reinforcement learning problem is $\beta$-strict separable (Regime 3) or not. In this case, the result of equation (**??**) can be used as a witness to determine that the obtained $\hat{R}$ satisfies Bellman's optimality condition with respect to the true transition probability matrices with high probability.

Let $\hat{R}$ be the solution to the optimization problem with $\hat{F}_{ai}$ with transition probability matrices $\hat{P}_a$ that are maximum likelihood estimates of $P_a$, which have at most $d \in \{1, \ldots, n\}$ non-zero elements per row, formed from $m$ samples and let

$$\varepsilon = 2\sqrt{\frac{4}{\alpha m} \log \frac{4nk}{\delta}} \cdot \frac{(d-1)\gamma + 1}{(1-\gamma)^2}$$

If $\|\hat{R}\|_1 \ll \frac{1}{\varepsilon}$, then with probability at least $(1-\delta)$, we have $F_{ai}^T \hat{R} \geq 0 \quad \forall a \in A \setminus a_1, i = 1, \ldots, n$.

## 1 CONCLUDING REMARKS

The L1-regularized support vector formulation along with the geometric interpretation provide a useful way of looking at the inverse reinforcement learning problem with strong, formal guarantees. Possible future work on this problem includes extension to the inverse reinforcement learning problem with continuous states by using sets of basis functions or feature vectors.

## References

[1] A. Y. Ng and S.J. Russel. Algorithms for inverse reinforcement learning. In *ICML 2000*, pages 663 – 670, 2000.

[2] Gergely Neu and Csaba Szepesvári. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *UAI 2007*, pages 295–302. AUAI Press, 2007.

[3] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *ICML 2006*, pages 729–736. ACM, 2006.

[4] U. Syed, M. Bowling, and R. Schapire. Apprenticeship learning using linear programming. In *ICML 2008*, pages 1032–1039. ACM, 2008.

[5] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. *IJCAI 2007*, 51(61801):1–4, 2007.

[6] Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. *AAAI 2008*.

[7] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Nonlinear inverse reinforcement learning with gaussian processes. In *NeurIPS 2011*, pages 19–27, 2011.

[8] J. Kiefer A. Dvoretzky and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[9] P. Wolfe. Finding the nearest point in a polytope. *Mathematical Programming*, 11(1):128–149, 1976.

[10] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. ICML '04, pages 1–, New York, NY, USA, 2004. ACM.

[11] Ji Zhu, Saharon Rosset, Robert Tibshirani, and Trevor J Hastie. 1-norm support vector machines. In *Advances in neural information processing systems*, pages 49–56, 2004.

[12] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[13] Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[14] Han Liu, Larry Wasserman, John D Lafferty, and Pradeep K Ravikumar. Spam: Sparse additive models. In *Advances in Neural Information Processing Systems*, pages 1201–1208, 2008.

[15] Hadi Daneshmand, Manuel Gomez-Rodriguez, Le Song, and Bernhard Scholkopf. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In *International Conference on Machine Learning*, pages 793–801, 2014.

[16] K. Dvijotham and E. Todorov. Inverse optimal control with linearly-solvable mdps. In *ICML 2010*, pages 335–342, 2010.

[17] S. Arora and P. Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *arXiv:1806.06877*, 2018.