Our paper presents an geometric interpretation for inverse reinforcement learning (IRL) with finite states and actions, as well as a corresponding L1-regularized Support Vector Machine formulation with formal guarantees in sample complexity and with regard to Bellman optimality. We thank the reviewers for their feedback and for bringing several issues and improvements to our attention. Our main contribution and focus in this paper is theoretical, which the reviewers seem to agree with, and the experiments serve primarily to validate our theory.

Following comments from the reviewers, we will move The proof of Theorem 5.1 to the appendix and add content discussing the background of IRL. In addition to the linear programming method [5] and the Bayesian IRL method [6] presented in the original submission, several other approaches to the IRL problem exist. These include Hybrid IRL [4], Maximum Margin Planning [7], Maximum Entropy IRL (MaxEnt) [9], and Gaussian Process IRL (GPIRL) [3]. A survey of other approaches to the inverse reinforcement learning problem can be found in [1]. While the problem in our paper is formulated in the form of a standard Markov Decision Process (MDP), several approaches instead consider the linearly-solvable MDP (LMDP) formulation presented in [2]. As mentioned in [2] (in particular Section 2.6), the problem formulation of the standard MDP and LMDP is different. Given the true transition probabilities, methods such as Multiplicative Weights for Apprenticeship Learning (MWAL) [8] and [5] are guaranteed to recover the true optimal policy where as methods that use LMDP to solve the same problem are not. To confirm this, we have used GPIRL [3] code available at `https://graphics.stanford.edu/projects/gpirl/` to solve the same synthetic experiments presented in Figure 3 of Section 8 of our paper and found the rewards from GPIRL were not Bellman optimal as per the definition in section 3 line 75 of our paper. Further comparisons with GPIRL, Bayesian IRL and MWAL are shown the Figure 1. We note that about 30% of the rewards returned by MWAL were the trivial solution $R = 0$ which were not counted as successes. The result presented in our paper immediately impacts algorithms that use standard MDP models more it impacts than algorithms that use LMDP models such as MaxEnt and GPIRL, as the objective of the LMDP-based algorithms is different. In the case of standard MDP problems it readily provides a sample complexity and a formal guarantee with respect to Bellman optimality, which is not provided by any of the other methods.

The formulation provided in our paper is a nonparametric approach as compared to approaches that use features derived from states. It also makes no assumptions on the sparseness of the transitions. Our experiments reflect this as well, as the transition probabilities are drawn from a uniform distribution with no sparseness assumptions and would be more difficult to than sparse cases. In contrast, tasks like gridworld tend to have have sparse transitions probabilities or feature transformations that reduce dimensions. We provide a method and a guarantee on optimality that does not require sparseness and does not depend on feature selection, which is a problem with other methods.

Definition 4.1 leads to our paper considering only Regime 3 cases. From a practical perspective this is not a loss. All problems within Regime 1 have only one feasible solution, $R \equiv 0$ which does not provide any information with respect to Bellman optimality as no policy is preferred over the other with this reward. Problems in Regime 2 require an infinite number of samples to both ascertain that the problem is indeed in Regime 2 and to solve the problem as perturbations can lead to the estimated problem falling under Regime 1 or Regime 3 as mentioned in line 134 of our paper. As stated in line 130, this type of assumption has been made in other areas as well.

While our paper does state the parameters and objective of the finite state MDP IRL problem from lines 39-76 as well as an initial formulation, our paper does not aim to derive the results and formulation of [5] from scratch. As suggested, we will present this problem as a standalone problem and explain the origin of the F matrix from Bellman Optimality.

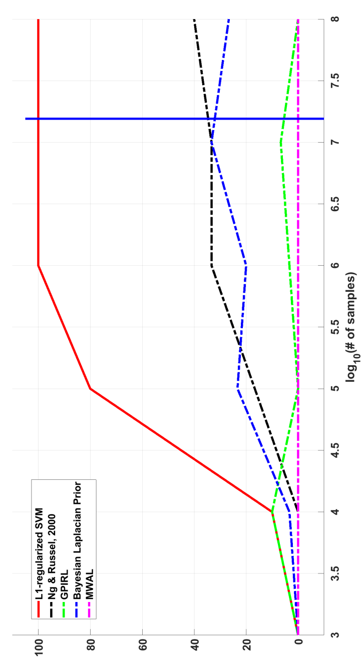

Figure 1:

References

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

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

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

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

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

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

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

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

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