A. Omitted Proofs

A.1. Proof of Lemma 4.3

Lemma A.1. *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*

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

Proof. Using Equation (3) of (Borade & Zheng, 2008), if $Q(j) = P(i) + J$, then we have

$$D(P(i)||Q(j)) \leq \frac{1}{2}||J||^2_{P(i)}$$

where

$$||J||^2_{P(i)} = \sum_{j=1}^{n} \frac{J_j^2}{P_j(i)}$$

Since both $P(i)$ and $Q(j)$ lie in the ball of radius ε around $\left[\frac{1}{n} \quad \frac{1}{n} \quad \dots \frac{1}{n}\right]$, we have $\max ||J||_2 = 2\varepsilon$ and $\min P_i(j) = \frac{1}{n} - \varepsilon$ and thus

$$||J||^2_P \leq \frac{\max ||J||_2^2}{\min P_i(j)} = \frac{4\varepsilon^2}{1/n - \varepsilon}$$

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

□

A.2. Proof of Corollary 4.1

Corollary A.1. *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*

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

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 $\mathcal{F} \to Z \to \hat{\mathcal{F}}$, we have

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

Proof. From Lemma 4.2, we know that the case $\varepsilon = \sqrt{n-2}\beta$ corresponds to the spherical code being an $n-1$-simplex which has n facets. Since the number of problems is the number of facets of the convex polytope, substituting $|F| = n$ in the proof of Theorem 4.3 along with $\varepsilon = \frac{1}{\sqrt{2n(n-1)}}$ gives us

$$\begin{aligned}\mathbb{P}(\hat{\mathcal{F}} \neq \mathcal{F}) &\geq 1 - \frac{(m-1)\frac{2\varepsilon^2 n}{1-n\varepsilon} + \log 2}{\log |F|} \\ &= 1 - \frac{(n-1)(0.5\log n - \log 2)\frac{1}{n-1} + \log 2}{\log n} \\ &= 1 - 0.5 = 0.5\end{aligned}$$

□

A.3. Proof of Corollary 4.2

Corollary A.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 $\mathcal{F} \to Z \to \hat{\mathcal{F}}$, we have

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

Proof. From Lemma 4.2, we know that the case $\varepsilon = \sqrt{n-2}\beta$ corresponds to the spherical code being an $n-1$-simplex which has n facets. Since the number of problems is the number of facets of the convex polytope, substituting $|F| = n$ in the proof of Theorem 4.3 along with $\varepsilon = \sqrt{n-2}\beta$ gives us

$$\begin{aligned}\mathbb{P}(\hat{\mathcal{F}} \neq \mathcal{F}) &\geq 1 - \frac{(m-1)\frac{2n(n-2)\beta^2}{1-n\sqrt{n-2}\beta} + \log 2}{\log |F|} \\ &= 1 - \frac{(0.5\log n - \log 2) + \log 2}{\log n} \\ &= 1 - 0.5 = 0.5\end{aligned}$$

□

B. Simulated Experiments

With permission from the authors, we apply our results to the simulated experiment cases performed in (Komanduru & Honorio, 2019) with a similar metric of percentage of trials where the estimated reward function generates the desired optimal strategy. The choice of this metric reflects the nature of our result: the correct identification of $\mathcal{F}$ from Theorem 4.3 is equivalent to the correct identification of the facet which contains the set of rewards that generate the optimal policy. This is in contrast to other methods that use closeness in the value function generated by the estimated reward function as their metric. We consider the scenario presented in their work (MDP with $n = 7$ states, $k = 7$ actions, $\gamma = 0.1$ and $\beta \approx 0.0032$) using the L1-regularized SVM (Komanduru & Honorio, 2019), the method of (Ng & Russel, 2000), Multiplicative Weights for Apprenticeship Learning from (Syed et al., 2008), Bayesian IRL with Laplacian prior from (Ramachandran & Amir, 2007) and Gaussian Process IRL from (Levine et al., 2011). The results are presented in Figure B.1. We also consider another case with $n = 5$ states, $k = 5$ actions, $\gamma = 0.1$ and $\beta \approx 0.0056$ similar to the case presented in (Komanduru & Honorio, 2019). The results for this case can be seen in Figure B.2.

In both cases, we observe that the performance of all the methods tested is abysmal ($< 50\%$ success) when the number samples is below or close to our predicted lower bound. The performance only starts to improve in various methods when the samples are well above the bound we present. This visibly supports our sample complexity lower bound of $O(\frac{\log n}{n^2\beta^2})$

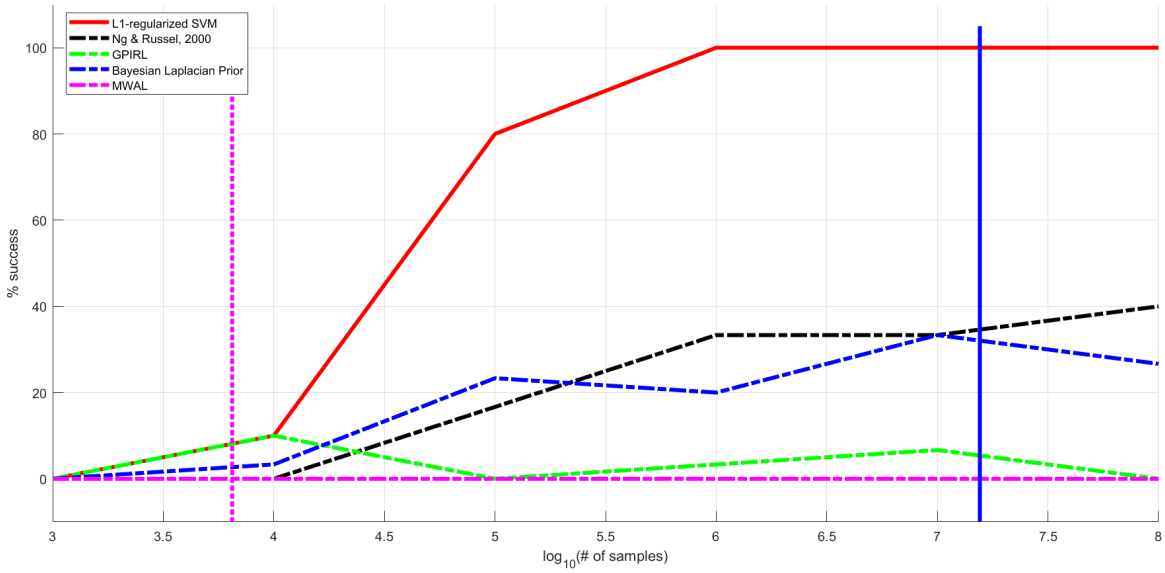

Figure B.1. Empirical probability of success versus log number of samples for an IRL problem with $n = 7$ states, $k = 7$ actions, $\gamma = 0.1$ and $\beta \approx 0.0032$) using the L1-regularized SVM (Komanduru & Honorio, 2019), the method of (Ng & Russel, 2000), Multiplicative Weights for Apprenticeship Learning from (Syed et al., 2008), Bayesian IRL with Laplacian prior from (Ramachandran & Amir, 2007) and Gaussian Process IRL from (Levine et al., 2011). The vertical magenta line represents the lower bound sample complexity from Corollary 4.2. The vertical blue line represents the sample complexity upper bound from (Komanduru & Honorio, 2019).

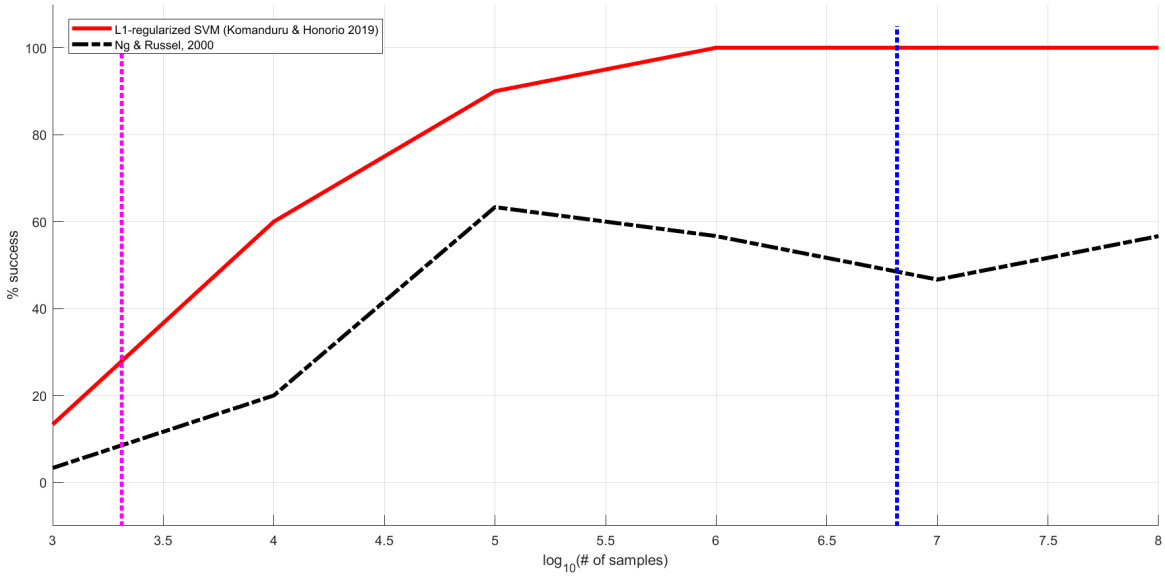

Figure B.2. Empirical probability of success versus log number of samples for an IRL problem with $n = 7$ states, $k = 7$ actions, $\gamma = 0.1$ and $\beta \approx 0.0056$) using the L1-regularized SVM (Komanduru & Honorio, 2019) and the method of (Ng & Russel, 2000). The vertical magenta line represents the lower bound sample complexity from Corollary 4.2. The vertical blue line represents the sample complexity upper bound from (Komanduru & Honorio, 2019).