Supplementary Material to ICML04 Submission Apprenticeship Learning via Inverse Reinforcement Learning

Abstract

In this document, we give more elaborate proofs for the theorems in the submitted paper. We also have a section on a different interpretation of the projection algorithm.

A. Proofs of Theoretical Results

Due to space constraints, in (Abbeel & Ng, 2004) we gave a full proof only for the case of $\hat{\mu}_E \in M$. Here we give proofs for the more general case, i.e. we have not necessarily that $\hat{\mu}_E \in M$. First however, we give a more extensive proof of Lemma 3 in (Abbeel & Ng, 2004), which was proved very densely there.

A.1. Extended Proof of Lemma 3

Figure 1 may be helpful for conveying geometric intuition.

Proof of Lemma 3

Proof. For simplicity of notation, we let the origin of our coordinate system coincide with $\bar{\mu}^{(i)}$. Then

$$\frac{(\tilde{\mu}^{(i+1)} - \hat{\mu}_E) \cdot (\tilde{\mu}^{(i+1)} - \hat{\mu}_E)}{\hat{\mu}_E \cdot \hat{\mu}_E} \tag{1}$$

$$= \frac{\mu^{(i+1)} \cdot \mu^{(i+1)} - \frac{(\mu^{(i+1)} \cdot \hat{\mu}_E)^2}{\hat{\mu}_E \cdot \hat{\mu}_E}}{\mu^{(i+1)} \cdot \mu^{(i+1)}} \tag{2}$$

$$\leq \frac{\mu^{(i+1)} \cdot \mu^{(i+1)} - 2\hat{\mu}_E \cdot \mu^{(i+1)} + \hat{\mu}_E \cdot \hat{\mu}_E}{\mu^{(i+1)} \cdot \mu^{(i+1)}} \tag{3}$$

$$= \frac{(\mu^{(i+1)} - \hat{\mu}_E) \cdot (\mu^{(i+1)} - \hat{\mu}_E)}{(\mu^{(i+1)} - \hat{\mu}_E) \cdot (\mu^{(i+1)} - \hat{\mu}_E) + \hat{\mu}_E \cdot \hat{\mu}_E + 2(\mu^{(i+1)} - \hat{\mu}_E) \cdot \hat{\mu}_E} \tag{4}$$

$$\leq \frac{(\mu^{(i+1)} - \hat{\mu}_E) \cdot (\mu^{(i+1)} - \hat{\mu}_E)}{(\mu^{(i+1)} - \hat{\mu}_E) \cdot (\mu^{(i+1)} - \hat{\mu}_E) + \hat{\mu}_E \cdot \hat{\mu}_E} \tag{5}$$

$$\leq \frac{k^2/(1-\gamma)^2}{k^2/(1-\gamma)^2 + \hat{\mu}_E \cdot \hat{\mu}_E}, \tag{6}$$

where we used in order:

1. The definition of $\tilde{\mu}^{(i+1)} = \frac{\mu^{(i+1)} \cdot \hat{\mu}_E}{\mu^{(i+1)} \cdot \mu^{(i+1)}} \mu^{(i+1)}$, which gives for the numerator $(\tilde{\mu}^{(i+1)} - \hat{\mu}_E) \cdot (\tilde{\mu}^{(i+1)} - \hat{\mu}_E) = (\frac{\mu^{(i+1)} \cdot \hat{\mu}_E}{\mu^{(i+1)} \cdot \mu^{(i+1)}} \mu^{(i+1)} - \hat{\mu}_E) \cdot (\frac{\mu^{(i+1)} \cdot \hat{\mu}_E}{\mu^{(i+1)} \cdot \mu^{(i+1)}} \mu^{(i+1)} - \hat{\mu}_E) = \frac{(\mu^{(i+1)} \cdot \hat{\mu}_E)^2}{(\mu^{(i+1)} \cdot \mu^{(i+1)})^2} \mu^{(i+1)} \cdot \mu^{(i+1)} - 2\frac{\mu^{(i+1)} \cdot \hat{\mu}_E}{\mu^{(i+1)} \cdot \mu^{(i+1)}} \mu^{(i+1)} \cdot \hat{\mu}_E + \hat{\mu}_E \cdot \hat{\mu}_E = -\frac{(\mu^{(i+1)} \cdot \hat{\mu}_E)^2}{\mu^{(i+1)} \cdot \mu^{(i+1)}} + \hat{\mu}_E \cdot \hat{\mu}_E$. Using this expression for the numerator, and multiplying numerator and denominator by $\frac{\mu^{(i+1)} \cdot \mu^{(i+1)}}{\hat{\mu}_E \cdot \hat{\mu}_E}$ gives Equation (2).

2. The following inequalities are easily seen to be true:

$$\begin{aligned}
&(\mu^{(i+1)} \cdot \hat{\mu}_E - \hat{\mu}_E \cdot \hat{\mu}_E)^2 \geq 0 \\
&(\mu^{(i+1)} \cdot \hat{\mu}_E)^2 - 2(\mu^{(i+1)} \cdot \hat{\mu}_E)(\hat{\mu}_E \cdot \hat{\mu}_E) + (\hat{\mu}_E \cdot \hat{\mu}_E)^2 \geq 0 \\
&-(\mu^{(i+1)} \cdot \hat{\mu}_E)^2 \leq -2(\mu^{(i+1)} \cdot \hat{\mu}_E)(\hat{\mu}_E \cdot \hat{\mu}_E) + (\hat{\mu}_E \cdot \hat{\mu}_E)^2 \\
&\frac{-(\mu^{(i+1)} \cdot \hat{\mu}_E)^2}{\hat{\mu}_E \cdot \hat{\mu}_E} \leq -2(\mu^{(i+1)} \cdot \hat{\mu}_E) + \hat{\mu}_E \cdot \hat{\mu}_E.
\end{aligned}$$

   We used the last of these inequalities.

3. For the numerator, we just used $(\mu^{(i+1)} - \hat{\mu}_E) \cdot (\mu^{(i+1)} - \hat{\mu}_E) = \mu^{(i+1)} \cdot \mu^{(i+1)} - 2\hat{\mu}_E \cdot \mu^{(i+1)} + \hat{\mu}_E \cdot \hat{\mu}_E$. We rewrote the denominator as follows $\mu^{(i+1)} \cdot \mu^{(i+1)} = (\mu^{(i+1)} - \hat{\mu}_E + \hat{\mu}_E) \cdot (\mu^{(i+1)} - \hat{\mu}_E + \hat{\mu}_E) = (\mu^{(i+1)} - \hat{\mu}_E) \cdot (\mu^{(i+1)} - \hat{\mu}_E) + (\hat{\mu}_E \cdot \hat{\mu}_E) + 2(\mu^{(i+1)} - \hat{\mu}_E) \cdot \hat{\mu}_E$.

4. Since $\pi^{(i+1)} = \arg\max_\pi \hat{\mu}_E \cdot \mu(\pi)$ (recall the origin is at $\bar{\mu}^{(i)}$ for notational convenience), we have $\hat{\mu}_E \cdot \mu^{(i+1)} = \hat{\mu}_E \cdot \mu(\pi^{(i+1)}) \geq \hat{\mu}_E \cdot \hat{\mu}_E$, which implies

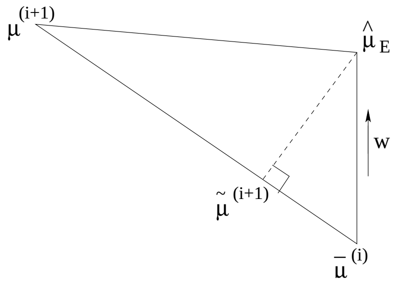

Figure 1. Progress in one iteration step.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

$2(\mu^{(i+1)} - \hat{\mu}_E) \cdot \hat{\mu}_E \geq 0$, which implies Equation (4).

5. All points considered lie in $M = [0, \frac{1}{1-\gamma}]^k$, so their norms are bounded by $k/(1-\gamma)$.

This proves the one step improvement equation of the Lemma.

It remains to prove that that $\tilde{\mu}^{(i+1)} = \lambda\mu^{(i+1)} + (1-\lambda)\bar{\mu}^{(i)}$, for some $\lambda \in [0,1]$. It is easily seen from the definition of $\tilde{\mu}^{(i+1)}$ that for $\lambda = \frac{\hat{\mu}_E \cdot \mu^{(i+1)}}{\mu^{(i+1)} \cdot \mu^{(i+1)}}$, we have $\tilde{\mu}^{(i+1)} = \lambda\mu^{(i+1)} + (1-\lambda)\bar{\mu}^{(i)}$. (Recall we still have $\bar{\mu}^{(i)}$ as our origin to simplify notation.) Since $\pi^{(i+1)} = \arg\max_\pi \hat{\mu}_E \cdot \mu(\pi)$, we have $\hat{\mu}_E \cdot \mu^{(i+1)} = \hat{\mu}_E \cdot \mu(\pi^{(i+1)}) \geq \hat{\mu}_E \cdot \hat{\mu}_E \geq 0$, which implies $\lambda \geq 0$.

Now to prove $\lambda \leq 1$, we start with Cauchy-Schwartz inequality

$$(\mu^{(i+1)} \cdot \hat{\mu}_E)^2 \leq (\mu^{(i+1)} \cdot \mu^{(i+1)})(\hat{\mu}_E \cdot \hat{\mu}_E), \tag{7}$$

which combined with $\hat{\mu}_E \cdot \hat{\mu}_E \leq \mu^{(i+1)} \cdot \hat{\mu}_E$ gives

$$(\mu^{(i+1)} \cdot \hat{\mu}_E)^2 \leq (\mu^{(i+1)} \cdot \mu^{(i+1)})(\hat{\mu}_E \cdot \mu^{(i+1)}), \tag{8}$$

from which we immediately get

$$\lambda = \frac{\hat{\mu}_E \cdot \mu^{(i+1)}}{\mu^{(i+1)} \cdot \mu^{(i+1)}} \leq 1 \tag{9}$$

□

A.2. More General Convergence Theorem

We first review some definitions from the proofs section in (Abbeel & Ng, 2004). Given a set of policies Π, we define $M = M(\Pi) = Co\{\mu(\pi) : \pi \in \Pi\}$ to be the convex hull of the set of feature expectations attained by policies $\pi \in \Pi$. Hence, given any vector of feature expectations $\tilde{\mu} \in M$, there is a set of policies $\pi_1, \ldots, \pi_n \in \Pi$ and mixture weights $\{\lambda_i\}_{i=1}^n$ ($\lambda_i \geq 0, \sum_{i=1}^n \lambda_i = 1$), so that $\tilde{\mu} = \sum_{i=1}^n \mu(\pi_i)$. Thus, given any point $\tilde{\mu} \in M$, by mixing together policies in Π, we can obtain a new policy whose feature expectations are exactly $\tilde{\mu}$. (Here, mixture policies are as defined in Section 2 of the paper.

We also define $M^{(i)} = Co\{\mu(\pi^{(j)}) : j = 0, \ldots, i\}$ to be the convex hull of the set of feature expectations of policies found after iterations $0, \ldots, i$ of our algorithm.

As mentioned previously, $\hat{\mu}_E$ is a noisy estimate of μ_E. Thus, it may not be a valid feature expectation vector for any policy; i.e., we do not necessarily have $\hat{\mu}_E \in M$. So rather than proving convergence to μ_E, we will instead consider a small ball with radius ρ centered around $\hat{\mu}_E$ and that intersects M, and prove convergence to this ball.

Lemmas 4-6 will establish properties for a single iteration of the algorithm, that will be useful for proving the main convergence theorem. In reading the proofs, Figure 2 may be helpful for conveying geometric intuition.

Lemma 4. *Let $\bar{\mu}^{(i)}, \hat{\mu}_E \in \mathbb{R}^k$, and $\rho \in \mathbb{R}$, with $\|\hat{\mu}_E - \bar{\mu}^{(i)}\|_2 \geq \rho$ be given,*
then the two following optimization problems

$$\min_{\mu:\|\hat{\mu}_E - \mu\|_2 \leq \rho} \|\mu - \bar{\mu}^{(i)}\|_2 \tag{10}$$

$$\min_{\mu:\|\hat{\mu}_E - \mu\|_2 \leq \rho} \mu \cdot (\hat{\mu}_E - \bar{\mu}^{(i)}) \tag{11}$$

have the same minimizing argument, which is given by

$$\mu_{\rho,i} = \frac{\rho}{\|\hat{\mu}_E - \bar{\mu}^{(i)}\|_2}\bar{\mu}^{(i)} + \frac{\|\hat{\mu}_E - \bar{\mu}^{(i)}\|_2 - \rho}{\|\hat{\mu}_E - \bar{\mu}^{(i)}\|_2}\hat{\mu}_E \tag{12}$$

We also have that

$$\exists \alpha > 0 \text{ such that } \bar{\mu}^{(i)} - \mu_{\rho,i} = \alpha(\bar{\mu}^{(i)} - \hat{\mu}_E) \tag{13}$$

Proof. This can be verified by solving each of the problems, which can be done by forming the Lagrangian, taking derivatives and setting to zero. The derivation is trivial but quite long, and thus omitted. Equation (13) follows immediately from Equation (12). □

Lemma 5. *Let there be given an MDP\R, features $\phi : S \mapsto [0,1]^k$, a set of policies Π, $\bar{\mu}^{(i)} \in M$, and $\rho \in \mathbb{R}$. Suppose $\|\hat{\mu}_E - \bar{\mu}^{(i)}\|_2 \geq \rho$, and that there exists some $\bar{\mu}_E \in M$ such that $\|\hat{\mu}_E - \bar{\mu}_E\|_2 \leq \rho$. Let $\pi^{(i+1)}$ be the optimal policy for the MDP\R with reward $R(s) = (\hat{\mu}_E - \bar{\mu}^{(i)}) \cdot \phi(s)$, and $\mu^{(i+1)} = \mu(\pi^{(i+1)})$. Further, let $\mu_{\rho,i} = \arg\min_{\mu:\|\hat{\mu}_E - \mu\|_2 \leq \rho} \|\mu - \bar{\mu}^{(i)}\|_2$. Then, we have that*

$$(\mu^{(i+1)} - \mu_{\rho,i}) \cdot (\bar{\mu}^{(i)} - \mu_{\rho,i}) \leq 0 \quad .$$

Proof. Since $\mu^{(i+1)} = \mu(\pi^{(i+1)})$, and $\pi^{(i+1)}$ is the optimal policy for the MDP\R with reward $R(s) = (\hat{\mu}_E - \bar{\mu}^{(i)}) \cdot \phi(s)$, we have

$$\mu^{(i+1)} = \arg\max\nolimits_{\mu \in M} (\hat{\mu}_E - \bar{\mu}^{(i)}) \cdot \mu.$$

This implies

$$(\hat{\mu}_E - \bar{\mu}^{(i)}) \cdot \mu^{(i+1)} \geq (\hat{\mu}_E - \bar{\mu}^{(i)}) \cdot \bar{\mu}_E \tag{14}$$

since $\bar{\mu}_E \in M$. Using the equivalent definition of $\mu_{\rho,i}$ as given by Equation (11) in Lemma 4 and $\|\bar{\mu}_E - \hat{\mu}_E\|_2 \leq \rho$, we have $(\hat{\mu}_E - \bar{\mu}^{(i)}) \cdot \bar{\mu}_E \geq (\hat{\mu}_E - \bar{\mu}^{(i)}) \cdot \mu_{\rho,i}$, which combined with Equation (14) gives

$$(\hat{\mu}_E - \bar{\mu}^{(i)}) \cdot \mu^{(i+1)} \geq (\hat{\mu}_E - \bar{\mu}^{(i)}) \cdot \mu_{\rho,i}$$

Simple manipulation gives

$$(\bar{\mu}^{(i)} - \hat{\mu}_E) \cdot (\mu^{(i+1)} - \mu_{\rho,i}) \leq 0$$

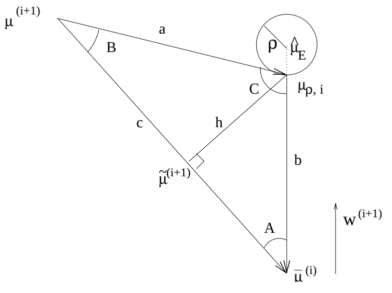

Figure 2. Triangle characterizing improvement for one iteration. $\bar{\mu}^{(i)} \in M^{(i)}, w^{(i+1)} = \alpha(\hat{\mu}_E - \bar{\mu}^{(i)}), \mu^{(i+1)} = \mu(\pi^{(i+1)})$, with $\pi^{(i+1)}$ the optimal policy for the given MDP\R, and $R(s) = w^{(i+1)} \cdot \phi(s)$. $\hat{\mu}_E$ is the estimate of the expert's feature expectations. $\mu_{\rho,i}$ is the point in the ρ-ball around $\hat{\mu}_E$ closest to $\bar{\mu}^{(i)}$. $\tilde{\mu}^{(i+1)}$ is the projection of $\mu_{\rho,i}$ onto the line through $\mu^{(i+1)}, \bar{\mu}^{(i)}.A, B, C$ denote the 3 angles of the triangle. a, b, c denote the vectors of each of the sides, i.e. $a = \mu_{\rho,i} - \mu^{(i+1)}, b = \bar{\mu}^{(i)} - \mu_{\rho,i}, c = \bar{\mu}^{(i)} - \mu^{(i+1)}$.

By using Equation (13) of Lemma 4 we get the desired result

$$(\bar{\mu}^{(i)} - \mu_{\rho,i}) \cdot (\mu^{(i+1)} - \mu_{\rho,i}) \leq 0$$

□

Lemma 6. *Let there be an MDP\R, features $\phi : S \mapsto [0,1]^k$, a set of policies Π, $\bar{\mu}^{(i)} \in M$, and $\rho \in \mathbb{R}$. Suppose $\|\hat{\mu}_E - \bar{\mu}^{(i)}\|_2 \geq \rho$ and there is some $\bar{\mu}_E \in M$ such that $\|\hat{\mu}_E - \bar{\mu}_E\|_2 \leq \rho$. Let $\pi^{(i+1)}$ be the optimal policy for the MDP\R with reward $R(s) = (\hat{\mu}_E - \bar{\mu}^{(i)}) \cdot \phi(s)$, and $\mu^{(i+1)} = \mu(\pi^{(i+1)})$. Further, let $\tilde{\mu}^{(i+1)}$ be the orthogonal projection of $\mu_{\rho,i} = \arg\min_{\mu:\|\hat{\mu}_E - \mu\|_2 \leq \rho} \|\mu - \bar{\mu}^{(i)}\|_2$ onto the line passing through the points $\mu^{(i+1)}$ and $\bar{\mu}^{(i)}$. We then have that*

$$\frac{\|\tilde{\mu}^{(i+1)} - \mu_{\rho,i}\|_2}{\|\bar{\mu}^{(i)} - \mu_{\rho,i}\|_2} \leq \frac{\sqrt{k}}{\sqrt{k + (1-\gamma)^2 \|\bar{\mu}^{(i)} - \mu_{\rho,i}\|_2^2}} \quad .$$

Proof. Consider the triangle formed by the three points $\bar{\mu}^{(i)}$, $\mu^{(i+1)}$ and $\mu_{\rho,i}$, and name the sides and angles as in Figure 2. (Note that a, b, and c are vectors; see figure caption.) Then we have

$$\begin{aligned} \frac{h}{\|\bar{\mu}^{(i)} - \mu_{\rho,i}\|_2} &= \frac{h}{\|b\|_2} && (15) \\ &= \sin A && (16) \\ &= \frac{\|a\|_2}{\|c\|_2} \sin C && (17) \\ &\leq \frac{a}{\sqrt{\|a\|_2^2 + \|b\|_2^2}}, && (18) \end{aligned}$$

where we used in order: the definition of b; the sin rule for right triangles; the sin rule for triangles[1]; $\sin C \leq 1$ and $\|c\|_2^2 = c^T c = (a+b)^T(a+b) = a^T a + b^T b + 2a^T b \geq \|a\|_2^2 + \|b\|_2^2$, where the inequality follows from $a^T b = (\mu^{(i+1)} - \mu_{\rho,i}) \cdot (\bar{\mu}^{(i)} - \mu_{\rho,i}) \leq 0$, which follows directly from Lemma 5. By taking the first derivative, we easily verify that $\frac{\|a\|_2}{\sqrt{\|a\|_2^2 + \|b\|_2^2}}$ is strictly increasing as a function of $\|a\|_2$ and thus $\frac{\|a\|_2}{\sqrt{\|a\|_2^2 + \|b\|_2^2}} \leq \frac{\sqrt{k}}{\sqrt{k + (1-\gamma)^2 \|b\|_2^2}}$ since $\sqrt{k}/(1-\gamma)$ is an upperbound on the distance between 2 points in M. This upperbound on the distance follows from $\phi \in [0,1]^k$ which implies $\mu \in \frac{1}{1-\gamma}[0,1]^k$. Combining this inequality with Equation (18) and $b = \bar{\mu}_i - \mu_{\rho,i}$ (by definition) gives the lemma. □

The above lemma describes how, starting from feature expectations $\bar{\mu}^{(i)} \in M$, we can pick $w = \hat{\mu}_E - \bar{\mu}^{(i)}$ such that the feature expectations $\mu^{(i+1)}$ of the corresponding optimal policy $\pi^{(i+1)}$ and $\bar{\mu}^{(i)}$ have a mixture $\tilde{\mu}^{(i+1)}$ such that $\|\tilde{\mu}^{(i+1)} - \mu_{\rho,i}\|_2 \leq c \|\bar{\mu}^{(i)} - \mu_{\rho,i}\|_2$, with $c < 1$. Convergence of each version of the algorithm is proved below, by showing how in each iteration, we achieve the improvement from the lemma.

For simplicity, in Section 3 of the paper we gave the algorithm assuming we are given the exact feature expectations μ_E. In the general case, the algorithm will use an estimate $\hat{\mu}_E$ instead, and convergence will be to a ball of radius ρ around $\hat{\mu}_E$, for any $\rho \geq \min_{\mu \in M} \|\mu - \hat{\mu}_E\|_2$. So now we let $t_{mm}^{(i)} = \min_{\nu \in M^{(i)}, \mu:\|\hat{\mu}_E - \mu\|_2 \leq \rho} \|\nu - \mu\|_2$ for the max-margin version, and $t_{proj}^{(i)} = \min_{\nu \in Co\{\bar{\mu}^{(i)}, \mu^{(i+1)}\}, \mu:\|\hat{\mu}_E - \mu\|_2 \leq \rho} \|\nu - \mu\|_2$. Note that $t_{proj}^{(i)} \geq t_{mm}^{(i)}$, because the $t_{proj}^{(i)}$ is defined using a minimize over a smaller domain ($Co\{\bar{\mu}^{(i)}, \mu^{(i+1)}\} \subseteq M^{(i+1)}$). Here, as before, we use CoZ to denote the convex hull of the set Z. Using our new definition of $t^{(i)}$, we can now state a more general version of the convergence theorem which can be seen to be the special case where $\hat{\mu}_E = \mu_E$ and thus we can choose $\rho = 0$ and the old and new definition of $t^{(i)}$ coincide.

Theorem 1. *Let an MDP\R, features $\phi : S \mapsto [0,1]^k$, any $\epsilon > 0$ and any $\rho \geq \min_{\mu \in M} \|\mu - \hat{\mu}_E\|_2$ be given. Then the apprenticeship learning algorithm (both max-margin and projection versions) will terminate with $t^{(i)} \leq \epsilon$ after at most*

$$n = O\left(\tfrac{k}{(1-\gamma)^2 \epsilon^2} \log \tfrac{k}{(1-\gamma)\epsilon}\right) \qquad (19)$$

iterations.

[1] For any triangle with sides a, b, c and opposite angles A, B, C the sin rule for triangles states $\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c}$.

Proof. We first show for both versions that we have geometric convergence at a rate $\frac{\sqrt{k}}{\sqrt{k+(1-\gamma)^2\epsilon^2}}$, and then use this to compute the required number of iterations. In the max-margin version, the computation of $w^{(i)}$ at every iteration, is easily seen to be equivalent to setting $w^{(i)} = \hat{\mu}_E - \bar{\mu}^{(i)}$, for $\bar{\mu}^{(i)} = \arg\min_{\mu \in M^{(i)}} \|\hat{\mu}_E - \mu\|_2$. Obviously $\bar{\mu}^{(i)} \in M^{(i)}$ as is required to apply Lemma 6 later. If we define $\mu_{\rho,i} = \arg\min_{\mu:\|\hat{\mu}_E - \mu\|_2 \leq \rho} \|\mu - \bar{\mu}^{(i)}\|_2$ as in Lemma 6, then we see that

$$t^{(i)}_{mm} = \|\mu_{\rho,i} - \bar{\mu}_i\|_2. \tag{20}$$

Define $\tilde{\mu}^{(i+1)}$ as in Lemma 6 and observe that $\tilde{\mu}^{(i+1)} \in M^{(i+1)}$, and using the definition of $t^{(i+1)}_{mm}$ we have

$$t^{(i+1)}_{mm} \leq \|\tilde{\mu}^{(i+1)} - \mu_{\rho,i}\|_2 \tag{21}$$

Combining Equations (20) and (21) gives

$$\frac{t^{(i+1)}_{mm}}{t^{(i)}_{mm}} \leq \frac{\|\tilde{\mu}^{(i+1)} - \mu_{\rho,i}\|_2}{\|\bar{\mu}_i - \mu_{\rho,i}\|_2}$$

Applying Lemma 6 gives

$$\frac{t^{(i+1)}_{mm}}{t^{(i)}_{mm}} \leq \frac{\sqrt{k}}{\sqrt{k + (1-\gamma)^2 \|\bar{\mu}^{(i)} - \mu_{\rho,i}\|_2^2}}$$

For the projection version, at every iteration we have $\bar{\mu}^{(i)} = \tilde{\mu}^{(i)} = \arg\min_{\mu \in Aff\{\bar{\mu}^{(i-1)}, \mu^{(i)}\}} \|\hat{\mu}_E - \mu\|_2 = \arg\min_{\mu \in Co\{\bar{\mu}^{(i)}, \mu^{(i+1)}\}} \|\hat{\mu}_E - \mu\|_2$, so obviously $\bar{\mu}^{(i)} \in M^{(i)}$ as is required to apply Lemma 6 later. Here, $Aff Z$ denotes the affine hull of Z; specifically, if Z is a set of two points, then this is the line through these 2 points.[2] The first equality holds because in step 2 of our algorithm we compute $\bar{\mu}^{(i-1)}$ exactly like we defined $\tilde{\mu}^{(i-1)}$. The second equality corresponds to the definition of orthogonal projection onto a line (and thus corresponds to our definition of $\tilde{\mu}^{(i)}$), the last equality follows because of Lemma 5.[3] If we define $\mu_{\rho,i} = \arg\min_{\mu:\|\hat{\mu}_E - \mu\|_2 \leq \rho} \|\mu - \bar{\mu}^{(i)}\|_2$ as in Lemma 6, then we see that

$$t^{(i)}_{proj} = \|\bar{\mu}_i - \mu_{\rho,i}\|_2 \tag{22}$$

Define $\tilde{\mu}^{(i+1)}$ as in Lemma 6, and observe that $\tilde{\mu}^{(i+1)} \in M^{(i+1)}$, and using the definition of $t^{(i+1)}_{proj}$ we have

$$t^{(i+1)}_{proj} \leq \|\tilde{\mu}^{(i+1)} - \mu_{\rho,i}\|_2 \tag{23}$$

Combining Equations (22) and (23) gives

$$\frac{t^{(i+1)}_{proj}}{t^{(i)}_{proj}} \leq \frac{\|\tilde{\mu}^{(i+1)} - \mu_{\rho,i}\|_2}{\|\bar{\mu}_i - \mu_{\rho,i}\|_2}$$

Applying Lemma 6 gives

$$\frac{t^{(i+1)}_{proj}}{t^{(i)}_{proj}} \leq \frac{\sqrt{k}}{\sqrt{k + (1-\gamma)^2 \|\bar{\mu}^{(i)} - \mu_{\rho,i}\|_2^2}}$$

So in both cases we get the same guarantee for improvement in every iteration. Throughout iterations this results into

$$t^{(i)} \leq \left(\frac{\sqrt{k}}{\sqrt{(1-\gamma)^2\epsilon^2 + k}}\right)^i t^{(0)} \leq \left(\frac{\sqrt{k}}{\sqrt{(1-\gamma)^2\epsilon^2 + k}}\right)^i \frac{\sqrt{k}}{1-\gamma}$$

where the last inequality follows from $M \subseteq \frac{1}{1-\gamma}[0,1]^k$, and so $\frac{\sqrt{k}}{1-\gamma}$ is an upper bound on the distance between any 2 points in M. So we have $t^{(i)} \leq \epsilon$ if and only if

$$\left(\frac{\sqrt{k}}{\sqrt{(1-\gamma)^2\epsilon^2 + k}}\right)^i \frac{\sqrt{k}}{1-\gamma} \leq \epsilon$$

which is equivalent to

$$i \geq \log \frac{\sqrt{k}}{(1-\gamma)\epsilon} / log \frac{\sqrt{(1-\gamma)^2\epsilon^2 + k}}{\sqrt{k}} = O\left(\frac{k}{(1-\gamma)^2\epsilon^2} \log \frac{k}{(1-\gamma)\epsilon}\right)$$

□

Note that in practice, convergence might be much faster than predicted by the upperbound in the theorem, since we have an improvement by at least a factor of $\frac{\sqrt{k}}{\sqrt{k+(1-\gamma)^2\|\bar{\mu}^{(i)} - \mu_{\rho,i}\|_2^2}}$ for every iteration, which is typically much better than the bound $\frac{\sqrt{k}}{\sqrt{k+(1-\gamma)^2\epsilon^2}}$ used in the proof.

Remark (using approximate RL algorithms). Sometimes, in each iteration of the algorithm we will be able to solve the MDP only approximately. It is straightforward to generalize our result to this setting. Assume on each iteration we can obtain an approximately optimal policy π such that $\|\mu(\pi) - \mu(\pi^*)\|_2 \leq \epsilon_1$, with π^* the optimal policy for that iteration. Then for any $\rho \geq \min_{\mu \in M} \|\mu - \hat{\mu}_E\|_2$ and any $\epsilon > 0$ our algorithm will converge to a policy $\tilde{\pi}$, such that $\|\hat{\mu}_E - \mu(\tilde{\pi})\|_2 \leq \rho + \epsilon_1 + \epsilon$ after at most the number of iterations given in Equation (19) of Theorem 1. This result is easily proved by changing the definition of the ρ-ball around $\hat{\mu}_E$ to a $(\rho + \epsilon_1)$-ball.

[2] More formally, $Aff Z = \{\sum_i \lambda_i z_i : z_i \in Z, \sum_i \lambda_i = 1, \lambda_i \in \mathbb{R}\}$.

[3] More precisely, Lemma 5 implies the 3 points $\mu^{(i+1)}, \mu_{\rho,i}, \bar{\mu}^{(i)}$ form an obtuse angle at $\mu_{\rho,i}$, which implies the orthogonal projection of $\mu_{\rho,i}$ onto $Aff\{\mu^{(i+1)}, \bar{\mu}^{(i)}\}$ falls into $Co\{\mu^{(i+1)}, \bar{\mu}^{(i)}\}$.

A.3. Sample Complexity

We now consider the number of samples from the expert required. In the paper we assumed $\hat{\mu}_E \in M$. Here we do not assume this, which leads to slightly different constant factors. The differences between this proof and the proof in the paper are the following: we take into account the possibility $\hat{\mu} \notin M$, the proof is a little less dense.

Theorem 2. *Let an MDP$\backslash$R, features $\phi : S \mapsto [0,1]^k$, and any $\epsilon > 0, \delta > 0$ be given. Suppose the apprenticeship learning algorithm (either max-margin or projection version) is run using an estimate $\hat{\mu}_E$ for μ_E obtained by m Monte Carlo samples. In order to ensure that with probability at least $1-\delta$ the algorithm terminates after at most a number of iterations n given by Equation (19), and outputs a policy $\tilde{\pi}$ so that for any true reward $R^*(s) = {w^*}^T\phi(s)$ ($\|w^*\|_1 \leq 1$)* [4] *we have*

$$E[\textstyle\sum_{t=0}^{\infty}\gamma^t R^*(s_t)|\tilde{\pi}] \geq E[\textstyle\sum_{t=0}^{\infty}\gamma^t R^*(s_t)|\pi_E]-\epsilon, \tag{24}$$

it suffices that

$$m \geq \tfrac{9k}{2(\epsilon(1-\gamma))^2}\log\tfrac{2k}{\delta}.$$

Proof. Recall $\phi \in [0,1]^k$ so $\mu \in [0, \frac{1}{1-\gamma}]^k$. Let μ_i denote the i'th component of μ then applying the Chernoff bound on the m-sample estimate $(1-\gamma)\hat{\mu}_i$ of $(1-\gamma)\mu_i \in [0,1]$ gives

$$P((1-\gamma)|\mu_i - \hat{\mu}_i| > \tau) \leq 2\exp(-2\tau^2 m). \tag{25}$$

Using Equation (25) for all components i and the union bound gives us

$$P(\exists i \in \{1\dots k\}.(1-\gamma)|\mu_i - \hat{\mu}_i| > \tau) \leq 2k\exp(-2\tau^2 m). \tag{26}$$

Now we subtract both sides of Equation (26) from 1, to find that

$$\begin{aligned} P(\neg\exists i \in \{1\dots k\}.(1-\gamma)|\mu_i - \hat{\mu}_i| > \tau) & \qquad (27)\\ = P((1-\gamma)\|\mu_E - \hat{\mu}_E\|_\infty \leq \tau) & \qquad (28)\\ \geq 1 - 2k\exp(-2\tau^2 m) & \qquad (29) \end{aligned}$$

Substituting $\tau = (1-\gamma)\epsilon/(3\sqrt{k})$ into Equation (29) gives

$$P(\|\mu_E - \hat{\mu}_E\|_\infty \leq \frac{\epsilon}{3\sqrt{k}}) \geq 1 - 2k\exp(-2(\frac{\epsilon(1-\gamma)}{3\sqrt{k}})^2 m), \tag{30}$$

where k is the dimension of the feature vectors ϕ and feature expectations μ, and m is the number of sample trajectories used for the estimate $\hat{\mu}_E$. So if we take $m \geq \frac{9k}{2(\epsilon(1-\gamma))^2}\log\frac{2k}{\delta}$ then with probability at least $(1-\delta)$ we have that

$$\|\mu_E - \hat{\mu}_E\|_\infty \leq \frac{\epsilon}{3\sqrt{k}}$$

Using $\|\cdot\|_2 \leq \sqrt{k}\|\cdot\|_\infty$ for k-dimensional space, we get

$$\|\mu_E - \hat{\mu}_E\|_2 \leq \frac{\epsilon}{3} \tag{31}$$

Since $\mu_E \in M$, Theorem 1 together with Equation (31) guarantee convergence to the ball of radius $\rho = \frac{\epsilon}{3}$ around $\hat{\mu}_E$. After sufficient iterations of the algorithm as specified in Equation (19), we have $t^{(i)} \leq \frac{\epsilon}{3}$, and thus the algorithm will return a policy $\tilde{\pi}$ such that $\tilde{\mu} = \mu(\tilde{\pi})$ satisfies

$$\|\tilde{\mu} - \hat{\mu}_E\|_2 \leq t^{(i)} + \rho \leq \frac{2\epsilon}{3} \tag{32}$$

Now (keeping in mind that $\|.\|_2 \leq \|.\|_1$ and so $\|w\|_1 \leq 1$ implies $\|w\|_2 \leq 1$) we can easily prove the result

$$\begin{aligned} &|E[\sum_{t=0}^{\infty}\gamma^t R^*(s^{(t)})|\tilde{\pi}] - E[\sum_{t=0}^{\infty}\gamma^t R^*(s^{(t)})|\pi_E]| \qquad (33)\\ &= \quad |(w^*)^T(\tilde{\mu} - \mu_E)|\\ &= \quad |(w^*)^T(\tilde{\mu} - \hat{\mu}_E + \hat{\mu}_E - \mu_E)|\\ &\leq \quad |(w^*)^T(\tilde{\mu} - \hat{\mu}_E)| + |(w^*)^T(\hat{\mu}_E - \mu_E)|\\ &\leq \quad \|w^*\|_2\|\tilde{\mu} - \hat{\mu}_E\|_2 + \|w^*\|_2\|\hat{\mu}_E - \mu_E\|_2\\ &\leq \quad 1(\frac{2\epsilon}{3} + \frac{\epsilon}{3}) \qquad \text{w.p. } (1-\delta) \text{ for } m \geq \frac{9k}{2(\epsilon(1-\gamma))^2}\log\frac{2k}{\delta}\\ &= \quad \epsilon \qquad \text{w.p. } (1-\delta) \text{ for } m \geq \frac{9k}{2(\epsilon(1-\gamma))^2}\log\frac{2k}{\delta} \end{aligned}$$

where we used in order the definition of w,μ; adding subtracting the same terms; the triangle inequality; Hölder's inequality for $p = q = 2$;[5] Equations (31), (32); and simplification. The last line directly implies the theorem.

□

Note that in case the underlying reward function R^* does not lie exactly in the span of basis functions, we have graceful degradation of performance. Let $R^*(s) = w^* \cdot \phi(s) + f(s)$, then it is easy to see from Equation (33) in the proof above, that performance degradation is bounded by $\frac{\|f\|_\infty}{1-\gamma}$.

[4] The rationale for the 1-norm constraint on w and ∞-norm constraint on ϕ is that these are dual norms, and thus we have $w \cdot \phi(s) \leq \|w\|_1\|\phi(s)\|_\infty$ (Hölder's inequality). So it corresponds to constraining the maximal reward $\max_s R(s) \leq 1$. Taking any 2 dual norms for w and ϕ would imply $max_s R(s) \leq 1$. Taking 2-norm constraints on w and ϕ results in exactly the same theorem.

[5] Hölder's inequality states that for any $p \geq 1, q \geq 1, 1/p + 1/q = 1$ and any x, y in some vector space we have $x \cdot y \leq \|x\|_p\|y\|_q$.

B. Alternative Interpretation of the Projection Algorithm

It is well-known that MDP's can be solved via linear programming, more specifically by solving the Bellman LP

$$\begin{aligned} &\min\nolimits_V e'V \qquad (34)\\ \text{s.t. } \forall s,a \;\; &V(s) \geq R(s) + \gamma \textstyle\sum_z P(z|s,a)V(z) \end{aligned}$$

where e is an $|S|$ dimensional vector of all ones. Although this LP is generally too big to solve exactly, there has been recent work on using this LP formulation to get approximate solutions (de Farias & Van Roy, 2003; Guestrin et al., 2003). The dual of this LP is

$$\begin{aligned} &\max\nolimits_\lambda \textstyle\sum_{s,a} \lambda(s,a)R(s) \qquad (35)\\ \text{s.t. } \forall s \;\; &\textstyle\sum_a \lambda(s,a) - \gamma \sum_{z,a} P(s|z,a)\lambda(z,a) = 1 \end{aligned}$$

The entry $\lambda(s,a)$ represents the expected frequency of the state-action pair s,a, the constraints ensure λ is consistent with the transition probabilities in the MDP. So we can write the feature expectations for a policy specified by λ explicitly as a function of λ

$$(\mu(\lambda))_k = \sum_{s,a} \lambda(s,a)\phi_k(s) \qquad (36)$$

We can explicitly formulate the problem of matching the expert's feature expectations μ_E as a QP

$$\begin{aligned} &\min\nolimits_\lambda \textstyle\sum_k (\mu_{E,k} - \sum_{s,a} \lambda(s,a)\phi_k(s))^2 \qquad (37)\\ \text{s.t. } \forall s \;\; &\textstyle\sum_a \lambda(s,a) - \gamma \sum_{z,a} P(s|z,a)\lambda(z,a) = 1 \end{aligned}$$

In practice, the above QP is typically too large to solve exactly, just like the Bellman LP and its dual. We will now derive an algorithm to solve the above QP, assuming we have access to a reinforcement learning algorithm, i.e. we assume we can get a solution to the Bellman LP (and/or its dual). The idea is to linearize the objective of (37) at the point of its current iterate, and then use the RL algorithm to solve the corresponding LP. Then the algorithm does a line search between the current iterate's point, and the solution to the LP (which is feasible, but not necessarily optimal for the QP). Then we linearize around the obtained point and iterate the above steps. [6] The linearized objective at a point $\lambda^{(i)}$ with corresponding feature expectations $\mu^{(i)}$ is easily computed to be

$$\sum_{s,a} \lambda(s,a) \sum_k (\mu_{E,k} - \mu_k^{(i)})\phi_k(s) \qquad (38)$$

So the corresponding LP is the dual of a Bellman LP, with reward $R(s) = \sum_k (\mu_{E,k} - \mu_k^{(i)})\phi_k(s) = (\mu_E - \mu^{(i)}) \cdot \phi(s)$. It is easily seen that the above algorithm corresponds to the *projection algorithm.* The projection corresponds to the line search, and choosing a reward weight vector w and finding the respective optimal policy corresponds to linearizing and solving the obtained LP.

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. *Proc. ICML.*

Censor, Y., & Zenios, S. (1997). *Parallel optimization: Theory, algorithms, and applications.* Oxford University Press.

de Farias, D. P., & Van Roy, B. (2003). The linear programming approach to approximate dynamic programming. *Operations Research, 51, No. 6.*

Guestrin, C., Koller, D., Parr, R., & Venkataraman, S. (2003). Efficient solution algorithms for factored mdps. *JAIR*, *19*, 399–468.

[6] Note our algorithm is an instantiation of the so called Frank-Wolfe algorithm (Censor & Zenios, 1997).