
A Game-Theoretic Approach to Apprenticeship Learning — Supplement

Umar Syed
 Computer Science Department
 Princeton University
 35 Olden St
 Princeton, NJ 08540-5233
 usyed@cs.princeton.edu

Robert E. Schapire
 Computer Science Department
 Princeton University
 35 Olden St
 Princeton, NJ 08540-5233
 schapire@cs.princeton.edu

1 The MWAL Algorithm

For reference, the MWAL algorithm from the main paper is repeated below.

Algorithm 1 The MWAL algorithm

- 1: **Given:** An MDP M and an estimate of the expert's feature expectations $\hat{\mu}_E$.
 - 2: Let $\beta = \left(1 + \sqrt{\frac{2 \ln k}{T}}\right)^{-1}$.
 - 3: Define $\tilde{\mathbf{G}}(i, \mu) \triangleq ((1 - \gamma)(\mu(i) - \hat{\mu}_E(i)) + 2)/4$, where $\mu \in \mathbb{R}^k$.
 - 4: Initialize $W^{(1)}(i) = 1$ for $i = 1, \dots, k$.
 - 5: **for** $t = 1, \dots, T$ **do**
 - 6: Set $w^{(t)}(i) = \frac{W^{(t)}(i)}{\sum_i W^{(t)}(i)}$ for $i = 1, \dots, k$.
 - 7: Compute an ϵ_P -optimal policy $\hat{\pi}^{(t)}$ for M with respect to reward function $R(s) = \mathbf{w}^{(t)} \cdot \boldsymbol{\phi}(s)$.
 - 8: Compute an ϵ_F -good estimate $\hat{\mu}^{(t)}$ of $\mu^{(t)} = \mu(\hat{\pi}^{(t)})$.
 - 9: $W^{(t+1)}(i) = W^{(t)}(i) \cdot \exp(\ln(\beta) \cdot \tilde{\mathbf{G}}(i, \hat{\mu}^{(t)}))$ for $i = 1, \dots, k$.
 - 10: **end for**
 - 11: Post-processing: Return the mixed policy $\bar{\psi}$ that assigns probability $\frac{1}{T}$ to $\hat{\pi}^{(t)}$, for all $t \in \{1, \dots, T\}$.
-

1.1 Differences between \mathbf{G} and $\tilde{\mathbf{G}}$

In the main paper, Algorithm 1 was motivated by appealing to the game matrix

$$\mathbf{G}(i, j) = \mu^j(i) - \mu_E(i),$$

where μ^j are the feature expectations of the j th deterministic policy. However, the algorithm actually uses

$$\tilde{\mathbf{G}}(i, \mu) = ((1 - \gamma)(\mu(i) - \hat{\mu}_E(i)) + 2)/4$$

The rationale behind each of the differences between \mathbf{G} and $\tilde{\mathbf{G}}$ follows.

- $\tilde{\mathbf{G}}$ depends on $\hat{\mu}_E$ instead of μ_E because μ_E is unknown and must be estimated. We account for the error of this estimate in the proof of Theorem 2.
- $\tilde{\mathbf{G}}$ is defined in terms of arbitrary feature expectations μ instead of μ^j because lines 7 and 8 of Algorithm 1 produce approximations, and hence $\hat{\mu}^{(t)}$ may not be the feature expectations of any deterministic policy. The results of Freund and Schapire [2] that we rely on are not affected by this change.

- $\tilde{\mathbf{G}}$ is shifted and scaled so that $\tilde{\mathbf{G}}(i, \boldsymbol{\mu}) \in [0, 1]$. This is necessary in order to directly apply the main result of Freund and Schapire [2].

The last point relies on a simplifying assumption. Recall that if $\boldsymbol{\mu}$ is a vector of feature expectations for some policy, then $\boldsymbol{\mu} \in [0, \frac{1}{1-\gamma}]^k$, because $\boldsymbol{\phi}(s) \in [0, 1]^k$ for all s . For simplicity, we will assume that this holds even if $\boldsymbol{\mu}$ is an *estimate* of a vector of feature expectations. (This is without loss of generality: if it does not hold, we can trim $\boldsymbol{\mu}$ so that it falls within the desired range without increasing the error in the estimate.) Therefore $(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_E) \in [\frac{-2}{1-\gamma}, \frac{2}{1-\gamma}]^k$, and hence $\tilde{\mathbf{G}}(i, \boldsymbol{\mu}) \in [0, 1]$.

2 Proof of Theorem 2

In this section we prove Theorem 2 from the main paper.

Theorem 2. *Given an MDP M , and m independent trajectories from an expert's policy π_E . Suppose we execute the MWAL algorithm for T iterations. Let $\bar{\boldsymbol{\psi}}$ be the mixed policy returned by the algorithm. Let ϵ_F and ϵ_P be the approximation errors from lines 7 and 8 of the algorithm. Let $H \geq (1/(1-\gamma)) \ln(1/(\epsilon_H(1-\gamma)))$ be the length of each sample trajectory. Let $\epsilon_R = \min_{\mathbf{w} \in \mathbb{S}^k} \max_s |R^*(s) - \mathbf{w} \cdot \boldsymbol{\phi}(s)|$ be the representation error of the features. Let $v^* = \max_{\boldsymbol{\psi} \in \Psi} \min_{\mathbf{w} \in \mathbb{S}^k} [\mathbf{w} \cdot \boldsymbol{\mu}(\boldsymbol{\psi}) - \mathbf{w} \cdot \boldsymbol{\mu}_E]$ be the game value. Then in order for*

$$V(\bar{\boldsymbol{\psi}}) \geq V(\pi_E) + v^* - \epsilon \quad (1)$$

to hold with probability at least $1 - \delta$, it suffices that

$$T \geq \frac{9 \ln k}{2(\epsilon'(1-\gamma))^2} \quad (2)$$

$$m \geq \frac{2}{(\epsilon'(1-\gamma))^2} \ln \frac{2k}{\delta} \quad (3)$$

$$(4)$$

where

$$\epsilon' \leq \frac{\epsilon - (2\epsilon_F + \epsilon_P + 2\epsilon_H + 2\epsilon_R/(1-\gamma))}{3}. \quad (5)$$

To prove Theorem 2, we will first need to prove several auxiliary results. Define

$$\tilde{\mathbf{G}}(\mathbf{w}, \boldsymbol{\mu}) \triangleq \sum_{i=1}^k \mathbf{w}(i) \cdot \tilde{\mathbf{G}}(i, \boldsymbol{\mu}).$$

Now we can directly apply the main result from Freund and Schapire [2], which we will call the MW Theorem.

MW Theorem. *At the end of the MWAL algorithm*

$$\frac{1}{T} \sum_{t=1}^T \tilde{\mathbf{G}}(\mathbf{w}^{(t)}, \hat{\boldsymbol{\mu}}^{(t)}) \leq \frac{1}{T} \min_{\mathbf{w} \in \mathbb{S}^k} \sum_{t=1}^T \tilde{\mathbf{G}}(\mathbf{w}, \hat{\boldsymbol{\mu}}^{(t)}) + \Delta_T$$

where

$$\Delta_T = \sqrt{\frac{2 \ln k}{T}} + \frac{\ln k}{T}.$$

Proof. Freund and Schapire [2]. □

The following corollary follows straightforwardly from the MW Theorem.

Corollary 1. *At the end of the MWAL algorithm*

$$\frac{1}{T} \sum_{t=1}^T [\mathbf{w}^{(t)} \cdot \hat{\boldsymbol{\mu}}^{(t)} - \mathbf{w}^{(t)} \cdot \hat{\boldsymbol{\mu}}_E] \leq \frac{1}{T} \min_{\mathbf{w} \in \mathbb{S}^k} \sum_{t=1}^T [\mathbf{w} \cdot \hat{\boldsymbol{\mu}}^{(t)} - \mathbf{w} \cdot \hat{\boldsymbol{\mu}}_E] + \Delta_T$$

The next lemma bounds the number of samples needed to make $\hat{\mu}_E$ close to μ_E .

Lemma 1. *Suppose the trajectory length $H \geq (1/(1-\gamma)) \ln(1/(\epsilon_H(1-\gamma)))$. For $\|\hat{\mu}_E - \mu_E\|_\infty \leq \epsilon + \epsilon_H$ to hold with probability at least $1 - \delta$, it suffices that*

$$m \geq \frac{2}{(\epsilon(1-\gamma))^2} \ln\left(\frac{2k}{\delta}\right)$$

Proof. This is a standard proof using Hoeffding's inequality, similar to that found in Abbeel and Ng [1]. However, care must be taken in one respect: $\hat{\mu}_E$ is *not* an unbiased estimate of μ_E , because the trajectories are truncated at H . So define

$$\mu_E^H \triangleq E \left[\sum_{t=0}^H \gamma^t \phi(s_t) \mid \pi_E, \theta, D \right].$$

Then we have,

$$\begin{aligned} \forall i \in [1, \dots, k] \quad & \Pr(|\hat{\mu}_E(i) - \mu_E^H(i)| \geq \epsilon) \leq 2 \exp(-m(\epsilon(1-\gamma))^2/2) \\ \Rightarrow \quad & \Pr(\exists i \in [1, \dots, k] \text{ s.t. } |\hat{\mu}_E(i) - \mu_E^H(i)| \geq \epsilon) \leq 2k \exp(-m(\epsilon(1-\gamma))^2/2) \\ \Rightarrow \quad & \Pr(\forall i \in [1, \dots, k], |\hat{\mu}_E(i) - \mu_E^H(i)| \leq \epsilon) \geq 1 - 2k \exp(-m(\epsilon(1-\gamma))^2/2) \\ \Rightarrow \quad & \Pr(\|\hat{\mu}_E - \mu_E^H\|_\infty \leq \epsilon) \geq 1 - 2k \exp(-m(\epsilon(1-\gamma))^2/2) \end{aligned}$$

We used in order: Hoeffding's inequality and $\mu_E^H \in [0, \frac{1}{1-\gamma}]^k$; the union bound; the probability of disjoint events; the definition of L_∞ norm.

It is not hard to show that $\|\mu_E^H - \mu_E\|_\infty \leq \epsilon_H$ (see Kearns and Singh [4], Lemma 2). Hence if $m \geq \frac{2}{(\epsilon(1-\gamma))^2} \ln(\frac{2k}{\delta})$, then with probability at least $1 - \delta$ we have

$$\|\hat{\mu}_E - \mu_E\|_\infty \leq \|\hat{\mu}_E - \mu_E^H\|_\infty + \|\mu_E^H - \mu_E\|_\infty \leq \epsilon + \epsilon_H.$$

□

The next lemma bounds the impact of “representation error”: it says that if $R^*(s)$ and $\mathbf{w}^* \cdot \phi(s)$ are not very different, then neither are $V(\psi)$ and $\mathbf{w}^* \cdot \mu(\psi)$.

Lemma 2. *If $\max_s |R^*(s) - \mathbf{w}^* \cdot \phi(s)| \leq \epsilon_R$, then $|V(\psi) - \mathbf{w}^* \cdot \mu(\psi)| \leq \frac{\epsilon_R}{1-\gamma}$ for every MDP/R M and mixed policy ψ .*

Proof.

$$\begin{aligned} & |V(\psi) - \mathbf{w}^* \cdot \mu(\psi)| \\ = & \left| E \left[\sum_{t=0}^{\infty} \gamma^t R^*(s_t) \right] - E \left[\sum_{t=0}^{\infty} \gamma^t \mathbf{w}^* \cdot \phi(s_t) \right] \right| \\ = & \left| \lim_{H \rightarrow \infty} E \left[\sum_{t=0}^H \gamma^t R^*(s_t) \right] - \lim_{H \rightarrow \infty} E \left[\sum_{t=0}^H \gamma^t \mathbf{w}^* \cdot \phi(s_t) \right] \right| \\ = & \left| \lim_{H \rightarrow \infty} E \left[\sum_{t=0}^H \gamma^t (R^*(s_t) - \mathbf{w}^* \cdot \phi(s_t)) \right] \right| \\ \leq & \lim_{H \rightarrow \infty} E \left[\sum_{t=0}^H \gamma^t |R^*(s_t) - \mathbf{w}^* \cdot \phi(s_t)| \right] \\ \leq & \frac{\epsilon_R}{1-\gamma} \end{aligned}$$

□

We are now ready to prove Theorem 2. The proof closely follows Section 2.5 of Freund and Schapire [2].

Proof of Theorem 2. Let $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^T \mathbf{w}^{(t)}$. Then we have

$$\begin{aligned} v^* &= \max_{\psi \in \Psi} \min_{\mathbf{w} \in \mathbb{S}^k} [\mathbf{w} \cdot \boldsymbol{\mu}(\psi) - \mathbf{w} \cdot \boldsymbol{\mu}_E] \\ &= \min_{\mathbf{w} \in \mathbb{S}^k} \max_{\psi \in \Psi} [\mathbf{w} \cdot \boldsymbol{\mu}(\psi) - \mathbf{w} \cdot \boldsymbol{\mu}_E] \end{aligned} \quad (6)$$

$$\begin{aligned} &\leq \min_{\mathbf{w} \in \mathbb{S}^k} \max_{\psi \in \Psi} [\mathbf{w} \cdot \boldsymbol{\mu}(\psi) - \mathbf{w} \cdot \hat{\boldsymbol{\mu}}_E] + \epsilon' + \epsilon_H \\ &\leq \max_{\psi \in \Psi} [\bar{\mathbf{w}} \cdot \boldsymbol{\mu}(\psi) - \bar{\mathbf{w}} \cdot \hat{\boldsymbol{\mu}}_E] + \epsilon' + \epsilon_H \end{aligned} \quad (7)$$

$$= \max_{\psi \in \Psi} \frac{1}{T} \sum_{t=1}^T [\mathbf{w}^{(t)} \cdot \boldsymbol{\mu}(\psi) - \mathbf{w}^{(t)} \cdot \hat{\boldsymbol{\mu}}_E] + \epsilon' + \epsilon_H \quad (8)$$

$$\leq \frac{1}{T} \sum_{t=1}^T \max_{\psi \in \Psi} [\mathbf{w}^{(t)} \cdot \boldsymbol{\mu}(\psi) - \mathbf{w}^{(t)} \cdot \hat{\boldsymbol{\mu}}_E] + \epsilon' + \epsilon_H$$

$$\leq \frac{1}{T} \sum_{t=1}^T [\mathbf{w}^{(t)} \cdot \boldsymbol{\mu}(\hat{\pi}^{(t)}) - \mathbf{w}^{(t)} \cdot \hat{\boldsymbol{\mu}}_E] + \epsilon_P + \epsilon' + \epsilon_H \quad (9)$$

$$\leq \frac{1}{T} \sum_{t=1}^T [\mathbf{w}^{(t)} \cdot \hat{\boldsymbol{\mu}}^{(t)} - \mathbf{w}^{(t)} \cdot \hat{\boldsymbol{\mu}}_E] + \epsilon_F + \epsilon_P + \epsilon' + \epsilon_H \quad (10)$$

$$\leq \frac{1}{T} \min_{\mathbf{w} \in \mathbb{S}^k} \sum_{t=1}^T [\mathbf{w} \cdot \hat{\boldsymbol{\mu}}^{(t)} - \mathbf{w} \cdot \hat{\boldsymbol{\mu}}_E] + \Delta_T + \epsilon_F + \epsilon_P + \epsilon' + \epsilon_H \quad (11)$$

$$\leq \frac{1}{T} \min_{\mathbf{w} \in \mathbb{S}^k} \sum_{t=1}^T [\mathbf{w} \cdot \boldsymbol{\mu}(\hat{\pi}^{(t)}) - \mathbf{w} \cdot \hat{\boldsymbol{\mu}}_E] + \Delta_T + 2\epsilon_F + \epsilon_P + \epsilon' + \epsilon_H \quad (12)$$

$$= \min_{\mathbf{w} \in \mathbb{S}^k} [\mathbf{w} \cdot \boldsymbol{\mu}(\bar{\psi}) - \mathbf{w} \cdot \hat{\boldsymbol{\mu}}_E] + \Delta_T + 2\epsilon_F + \epsilon_P + \epsilon' + \epsilon_H \quad (13)$$

$$\leq \min_{\mathbf{w} \in \mathbb{S}^k} [\mathbf{w} \cdot \boldsymbol{\mu}(\bar{\psi}) - \mathbf{w} \cdot \boldsymbol{\mu}_E] + \Delta_T + 2\epsilon_F + \epsilon_P + 2\epsilon' + 2\epsilon_H \quad (14)$$

$$\leq \mathbf{w}^* \cdot \boldsymbol{\mu}(\bar{\psi}) - \mathbf{w}^* \cdot \boldsymbol{\mu}_E + \Delta_T + 2\epsilon_F + \epsilon_P + 2\epsilon' + 2\epsilon_H \quad (15)$$

$$\leq V(\bar{\psi}) - V(\pi_E) + \Delta_T + 2\epsilon_F + \epsilon_P + 2\epsilon' + 2\epsilon_H + (2\epsilon_R)/(1 - \gamma) \quad (16)$$

In (6), we used von Neumann's minmax theorem. In (7), Lemma 1. In (8), the definition of $\bar{\mathbf{w}}$. In (9), the fact that $\hat{\pi}^t$ is ϵ_P -optimal w.r.t. $R(s) = \mathbf{w}^t \cdot \phi(s)$. In (10), the fact that $\hat{\boldsymbol{\mu}}^{(t)}$ is an ϵ_F -good estimate of $\boldsymbol{\mu}(\hat{\pi}^{(t)})$. In (11), Corollary 1. In (12), again the fact that $\hat{\boldsymbol{\mu}}^{(t)}$ is an ϵ_F -good estimate of $\boldsymbol{\mu}(\hat{\pi}^{(t)})$. In (13), the definition of $\bar{\psi}$. In (14), Lemma 1. In (15), we let $\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathbb{S}^k} \max_s |R^*(s) - (\mathbf{w} \cdot \phi(s))|$. In (16), Lemma 2.

Plugging in the choice for T into Δ_T and rearranging implies the theorem. \square

3 When transition function is unknown

We will employ several technical lemmas developed in Kearns and Singh [4] and Abbeel and Ng [5]. This is not a complete proof, but just a sketch of the main components of one.

For an MDP/R $M = (\mathcal{S}, \mathcal{A}, \gamma, \theta, \phi)$, suppose that we know $\theta(s, a, \cdot)$ exactly on a subset $Z \subseteq \mathcal{S} \times \mathcal{A}$. Then we can construct a estimate M_Z of M according to the following definition, which is similar to Definition 9 in Kearns and Singh [4].

Definition 1. Let $M = (\mathcal{S}, \mathcal{A}, \gamma, \theta, \phi)$ be a MDP/R, and let $Z \subseteq \mathcal{S} \times \mathcal{A}$. Then the induced MDP/R $M_Z = (\mathcal{S} \cup \{s_0\}, \mathcal{A}, \gamma, \theta_Z, \phi_Z)$ is defined as follows, where $\mathcal{S}_Z = \{s : (s, a) \in Z \text{ for some } a \in \mathcal{A}\}$:

- $\theta_Z(s_0, a, s_0) = 1$ for all $a \in \mathcal{A}$, i.e. s_0 is an absorbing state.
- If $(s, a) \in Z$ and $s' \in \mathcal{S}_Z$, then $\theta_Z(s, a, s') = \theta(s, a, s')$.

- If $(s, a) \in Z$, then $\theta_Z(s, a, s_0) = 1 - \sum_{s' \in \mathcal{S}_Z} \theta(s, a, s')$.
- If $(s, a) \notin Z$, then $\theta_Z(s, a, s_0) = 1$.
- $\phi_Z(s) = \phi(s)$ for all $s \in \mathcal{S}$, and $\phi_Z(s_0) = -\mathbf{1}$, where $-\mathbf{1}$ is the k -length vector of all -1 's.

The following lemma, due to Kearns and Singh [4] (Lemma 7), shows that M_Z is essentially a pessimistic estimate for M .

Lemma 3. Let $M = (\mathcal{S}, \mathcal{A}, \gamma, \theta, \phi)$ be a MDP/R where $\phi(s) \in [-1, 1]^k$, and let $Z \subseteq \mathcal{S} \times \mathcal{A}$. Then for all $\mathbf{w} \in \mathbb{S}^k$ and $\psi \in \Psi$, we have $\mathbf{w} \cdot \mu(\psi, M) \geq \mathbf{w} \cdot \mu(\psi, M_Z)$.

Proof. As above, let $\mathcal{S}_Z = \{s : (s, a) \in Z \text{ for some } a \in \mathcal{A}\}$. Also let $\mathcal{A}_Z = \{a : (s, a) \in Z \text{ for some } s \in \mathcal{S}\}$. All transitions in M_Z between states in \mathcal{S}_Z using an action in \mathcal{A}_Z are the same as in M , while all other transitions are routed to the absorbing state s_0 . Observing that $\phi(s_0) = -\mathbf{1}$ and $\phi(s) \succeq -\mathbf{1}$ for all s proves the lemma. \square

Definition 2. Let $M = (\mathcal{S}, \mathcal{A}, \gamma, \theta, \phi)$ be an MDP/R. Let H be the length of each sample trajectory from the expert's policy. Then we say a subset $Z \subseteq \mathcal{S} \times \mathcal{A}$ is (η, H) -visited by π_E in M if

$$Z = \left\{ (s, a) \mid \Pr(\exists t \in [1, \dots, H] \text{ such that } (s_t, a_t) = (s, a) \mid \pi_E, M) \geq \frac{\eta}{|\mathcal{S}||\mathcal{A}|} \right\}. \quad (17)$$

The following lemma, due to Abbeel and Ng [5], says that if $Z \subseteq \mathcal{S} \times \mathcal{A}$ is (η, H) -visited by π_E in M , then π_E has a similar value in M_Z as it does in M .

Lemma 4. Let $M = (\mathcal{S}, \mathcal{A}, \gamma, \theta, \phi)$ be a MDP/R, let $H \geq (1/(1-\gamma)) \ln(1/(\epsilon_H(1-\gamma)))$, and let $Z \subseteq \mathcal{S} \times \mathcal{A}$ be (η, H) -visited by π_E in M . Then for all $\mathbf{w} \in \mathbb{S}^k$

$$|\mathbf{w} \cdot \mu(\pi_E, M) - \mathbf{w} \cdot \mu(\pi_E, M_Z)| \leq \frac{\eta}{1-\gamma} + \epsilon_H. \quad (18)$$

Proof. By the definition of M_Z and the union bound, we have $\Pr(\{(s_t, a_t)\}_{t=1}^H \subseteq Z \mid \pi_E, M_Z) = \Pr(\{(s_t, a_t)\}_{t=1}^H \subseteq Z \mid \pi_E, M) \geq 1 - \eta$. Now suppose $\mathbf{w} \cdot \mu(\pi_E, M) \geq \mathbf{w} \cdot \mu(\pi_E, M_Z)$. Then

$$|\mathbf{w} \cdot \mu(\pi_E, M) - \mathbf{w} \cdot \mu(\pi_E, M_Z)| \quad (19)$$

$$= E \left[\sum_{t=0}^H \gamma^t \mathbf{w} \cdot \phi(s_t) \mid \pi_E, M \right] + E \left[\sum_{t=H+1}^{\infty} \gamma^t \mathbf{w} \cdot \phi(s_t) \mid \pi_E, M \right] \quad (20)$$

$$- E \left[\sum_{t=0}^H \gamma^t \mathbf{w} \cdot \phi(s_t) \mid \pi_E, M_Z \right] - E \left[\sum_{t=H+1}^{\infty} \gamma^t \mathbf{w} \cdot \phi(s_t) \mid \pi_E, M_Z \right] \quad (21)$$

$$\leq \eta \frac{1-\gamma^H}{1-\gamma} + \frac{\gamma^{H+1}}{1-\gamma} \quad (22)$$

$$\leq \frac{\eta}{1-\gamma} + \epsilon_H \quad (23)$$

A parallel argument can be made in case $\mathbf{w} \cdot \mu(\pi_E, M) \leq \mathbf{w} \cdot \mu(\pi_E, M_Z)$. \square

Since we will not know M_Z exactly, we will need to estimate it. The following lemma, due to Abbeel and Ng [5] (Lemma 14), says that if two MDP/R's M and \widehat{M} do not differ much, then the value of the same policy in M and \widehat{M} is not very different.

Lemma 5. Let $M = (\mathcal{S}, \mathcal{A}, \gamma, \theta, \phi)$ and $\widehat{M} = (\mathcal{S}, \mathcal{A}, \gamma, \widehat{\theta}, \widehat{\phi})$ be two MDP/R's that differ only in their transition functions. Suppose θ and $\widehat{\theta}$ satisfy

$$\forall s \in \mathcal{S}, a \in \mathcal{A} \quad \|\theta(s, a, \cdot), \widehat{\theta}(s, a, \cdot)\|_1 \leq \epsilon. \quad (24)$$

Then for all $\psi \in \Psi$ and $\mathbf{w} \in \mathbb{S}^k$, we have

$$\left| \mathbf{w} \cdot \mu(\psi, M) - \mathbf{w} \cdot \mu(\psi, \widehat{M}) \right| \leq \frac{2\epsilon}{(1-\gamma)^2}. \quad (25)$$

The following lemma, due to Abbeel and Ng [5] (Lemma 17), bounds the number of trajectories needed from π_E to make θ and $\hat{\theta}$ similar on a subset $Z \subseteq \mathcal{S} \times \mathcal{A}$ that is (η, H) -visited by π_E .

Lemma 6. *Let $M = (\mathcal{S}, \mathcal{A}, \gamma, \theta, \phi)$. Let $Z \subseteq \mathcal{S} \times \mathcal{A}$ be (ϵ, H) -visited by π_E in M . Let $\hat{\theta}$ be the MLE for θ formed by observing m independent trajectories from π_E . Also, let $K(s, a)$ denote the actual number of times (s, a) is visited in the m trajectories. Then for*

$$\forall (s, a) \in Z, K(s, a) \geq \frac{|\mathcal{S}|^2}{4\epsilon^2} \ln \frac{|\mathcal{S}|^3 |\mathcal{A}|}{\epsilon} \quad (26)$$

$$\forall (s, a) \in Z, \|\theta(s, a, \cdot), \hat{\theta}(s, a, \cdot)\|_1 \leq \epsilon \quad (27)$$

to hold with probability $1 - \delta$, it suffices that

$$m \geq \frac{|\mathcal{S}|^3 |\mathcal{A}|}{8\epsilon^3} \ln \frac{|\mathcal{S}|^3 |\mathcal{A}|}{\delta\epsilon} + |\mathcal{S}| |\mathcal{A}| \ln \frac{2|\mathcal{S}| |\mathcal{A}|}{\delta}. \quad (28)$$

3.1 Putting it all together

Here is the algorithm:

1. Collect $m \geq \frac{|\mathcal{S}|^3 |\mathcal{A}|}{8\epsilon^3} \ln \frac{|\mathcal{S}|^3 |\mathcal{A}|}{\delta\epsilon} + |\mathcal{S}| |\mathcal{A}| \ln \frac{2|\mathcal{S}| |\mathcal{A}|}{\delta}$ sample trajectories from the expert.
2. Define the following:
 - (a) Let Z be the set of all state-action pairs (s, a) such that $K(s, a) \geq \frac{|\mathcal{S}|^2}{4\epsilon^2} \ln \frac{|\mathcal{S}|^3 |\mathcal{A}|}{\epsilon}$.
 - (b) Let $\hat{\theta}$ be the MLE for θ .
 - (c) Let $M = (\mathcal{S}, \mathcal{A}, \gamma, \theta, \phi)$ and $\hat{M} = (\mathcal{S}, \mathcal{A}, \gamma, \hat{\theta}, \phi)$.
3. Submit \hat{M}_Z and $\hat{\mu}_E$ to the MWAL algorithm, which returns $\bar{\psi}$.

Lemma 3 shows that $V(\bar{\psi}, M)$ is more than $V(\bar{\psi}, M_Z)$. Lemma 5 says that $V(\bar{\psi}, M_Z)$ is close to $V(\bar{\psi}, \hat{M}_Z)$. Since \hat{M}_Z is the MDP that we gave to the MWAL algorithm, Theorem 2 says that $V(\bar{\psi}, \hat{M}_Z)$ is more than $V(\pi_E, \hat{M}_Z)$. Lemma 5 says that $V(\pi_E, \hat{M}_Z)$ is close to $V(\pi_E, M_Z)$. Lemma 4 says that $V(\pi_E, M_Z)$ is close to $V(\pi_E, M)$.

References

- [1] P. Abbeel, A. Ng (2004). Apprenticeship Learning via Inverse Reinforcement Learning. *ICML* **21**
- [2] Y. Freund, R. E. Schapire (1996). Game Theory, On-line Prediction and Boosting. *COLT* **9**
- [3] Y. Freund, R. E. Schapire (1999). Adaptive Game Playing Using Multiplicative Weights. *Games and Economic Behavior* **29**, 79–103.
- [4] M. Kearns, S. Singh (2002). Near-Optimal Reinforcement Learning in Polynomial Time. *Machine Learning* **49**, 209–232.
- [5] P. Abbeel, A. Ng (2005). Exploration and Apprenticeship Learning in Reinforcement Learning. *ICML* **22** (Long version; available at <http://www.cs.stanford.edu/~pabbeel/>)