Online Inverse Reinforcement Learning under Occlusion

Supplementary File

Saurabh Arora, Prashant Doshi
THINC Lab, Dept. of Computer Science
University of Georgia, Athens, GA
{sa08751,pdoshi}@uga.edu

Bikramjit Banerjee
School of Computing Sciences & Computer Engineering
University of Southern Mississippi, Hattiesburg, MS
Bikramjit.Banerjee@usm.edu

ACM Reference Format:
Saurabh Arora, Prashant Doshi and Bikramjit Banerjee. 2019. Online Inverse Reinforcement Learning under Occlusion. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

PRELIMINARIES FOR PROOFS

$\mathbb{X}$ is the space of all possible trajectories. The expected value of any feature $\phi_k \in [0,1], k \in \{1, 2 \dots K\}$ for trajectory X is given by function $f_i : \mathbb{X} \rightarrow \mathbb{R}$ defined as $f_k(X) = \sum_{\langle s,a\rangle_t \in X} \gamma^t \phi_k(\langle s,a\rangle_t)$. Although a trajectory in a non-terminating MDP can be infinitely long, we derive range of f_k first for bounded-length trajectories and extend it later by applying infinity limit. Let T_{max} be the maximum length of any trajectory, $0 \leq |X| \leq T_{max}$.

Then,

$$\sum_{t=0}^{0} \gamma^t 0 \leq \sum_{\langle s,a\rangle_t \in X} \gamma^t \phi_k(\langle s,a\rangle_t) \leq \sum_{t=0}^{T_{max}} \gamma^t$$
$$0 \leq f_k(X) \leq (1-\gamma^{T_{max}})/(1-\gamma)$$

Applying limit $T_{max} \rightarrow \infty$ gives us

$$0 \leq f_k(X) \leq \frac{1}{1-\gamma} \tag{1}$$

Extending the definition to all k features, we introduce function $f : \mathbb{X} \rightarrow \mathbb{R}^k$ as $f(X) = \sum_{\langle s,a\rangle_t \in X} \gamma^t \phi(\langle s,a\rangle_t)$.

Note that learned feature expectations can be expressed in terms of f_k as

$$E_{\mathbb{X}}[\phi_k] \triangleq \sum\nolimits_{X \in \mathbb{X}} Pr(X)\ f_k(X),\ k = 1 \dots K$$

The sessions for latent and full-observation MAXENTIRL updates estimated feature expectations of expert as follows.

$$\hat{\phi}^{Z|_Y,1:i}_{\boldsymbol{\theta}^i,k} \triangleq \frac{1}{|\boldsymbol{\mathcal{Y}}_{1:i}|} \sum_{Y \in \boldsymbol{\mathcal{Y}}_{1:i}} \sum_{Z \in \mathbb{Z}} Pr(Z|Y;\boldsymbol{\theta}) \sum_{\langle s,a\rangle_t \in Y \cup Z} \gamma^t \phi_k(\langle s,a\rangle_t)$$
$$= \frac{|\boldsymbol{\mathcal{Y}}_{1:i-1}|}{|\boldsymbol{\mathcal{Y}}_{1:i-1}| + |\boldsymbol{\mathcal{Y}}_i|}\ \hat{\phi}^{Z|_Y,1:i-1}_{\boldsymbol{\theta}^{i-1},k} + \frac{|\boldsymbol{\mathcal{Y}}_i|}{|\boldsymbol{\mathcal{Y}}_{1:i-1}| + |\boldsymbol{\mathcal{Y}}_i|}\ \hat{\phi}^{Z|_Y,i}_{\boldsymbol{\theta}^i,k} \tag{2}$$

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

$$\hat{\phi}^{1:i}_k \triangleq \frac{1}{|\boldsymbol{\mathcal{Y}}_{1:i}|} \sum_{Y \in \boldsymbol{\mathcal{Y}}_{1:i}} \sum_{Z \in \mathbb{Z}} Pr(Z|Y;\boldsymbol{\theta}) \sum_{\langle s,a\rangle_t \in Y \cup Z} \gamma^t \phi_k(\langle s,a\rangle_t)$$
$$= \frac{|\boldsymbol{\mathcal{Y}}_{1:i-1}|}{|\boldsymbol{\mathcal{Y}}_{1:i-1}| + |\boldsymbol{\mathcal{Y}}_i|}\ \hat{\phi}^{1:i-1}_k + \frac{|\boldsymbol{\mathcal{Y}}_i|}{|\boldsymbol{\mathcal{Y}}_{1:i-1}| + |\boldsymbol{\mathcal{Y}}_i|}\ \hat{\phi}^i_k \tag{3}$$

From definitions of feature-expectations and Equation 1, $E_{\mathbb{X}}[\phi_k]$, $\hat{\phi}^{1:i}_k, \hat{\phi}^{Z|_Y,1:i}_{\boldsymbol{\theta}^i,k} \in \left[0, \frac{1}{(1-\gamma)}\right]$.

PROOFS OF PERFORMANCE GUARANTEES

Proof of Theorem 1. We use the notation:

$$E_{\mathbb{X}}[\phi_k] \triangleq \sum_{X \in \mathbb{X}} Pr(X) \sum_{\langle s,a\rangle \in X} \phi_k(s,a),\ k = 1 \dots K$$

By allowing a relaxation in the constraints of maximum entropy estimation problem, [2] derived sample complexity bounds for the problem.

$$\max_{\Delta} (-\textstyle\sum_{X \in \mathbb{X}} Pr(X)\ log\ Pr(X))$$
$$\textbf{subject to } \textstyle\sum_{X \in \mathbb{X}} Pr(X) = 1$$
$$\left|E_{\mathbb{X}}[\phi_k] - \hat{\phi}^{1:i}_k\right| \leq \beta^{full}_k\ \forall k \in \{1 \dots K\} \tag{4}$$

Here $\beta^{full} \in \mathbb{R}^K$ is a vector of upper bounds on the differences between $E_{\mathbb{X}}[\phi_k]$ and $\hat{\phi}^{1:i}_k$.

Following proofs by Dudik et al., relaxed constraints maximum entropy IRL problem is same as $\min_{\boldsymbol{\theta}}(-\sum_{X \in \mathcal{X}_{1:i}} \tilde{Pr}(X)\ \log Pr(X|\boldsymbol{\theta}) + \sum_k \beta^{full}_k |\theta_k|) = \min_{\boldsymbol{\theta}}(-LL(\boldsymbol{\theta}|\mathcal{X}_i, |\mathcal{X}_{i-1}|, \hat{\phi}^{1:i-1}, \boldsymbol{\theta}^{i-1}) + \sum_k \beta^{full}_k |\theta_k|) = \min_{\boldsymbol{\theta}} NLL_{\beta^{full}}(\boldsymbol{\theta}|\mathcal{X}_i, |\mathcal{X}_{i-1}|, \hat{\phi}^{1:i-1}, \boldsymbol{\theta}^{i-1})$ (say).

The proof here is partially inspired from Corollary 1 in [2]. Let $\beta^{full}_k = \beta^{full}_c = \varepsilon/(1-\gamma)$ for all $k \in \{1 \dots K\}$, where β^{full}_c is constant because ε is fixed input. For normalized exponentiated gradient descent used here for computing maximum, $\sum_1^K |\theta_k| = 1$. Then, $NLL_{\beta^{full}}(\boldsymbol{\theta}|\mathcal{X}_i, |\mathcal{X}_{i-1}|, \hat{\phi}^{1:i-1}, \boldsymbol{\theta}^{i-1}) = (-LL(\boldsymbol{\theta}|\mathcal{X}_i,\ |\mathcal{X}_{i-1}|, \hat{\phi}^{1:i-1}, \boldsymbol{\theta}^{i-1}) + \beta^{full}_c \sum_1^k |\theta_k|) = (-LL(\boldsymbol{\theta}|\mathcal{X}_i, |\mathcal{X}_{i-1}|,\ \hat{\phi}^{1:i-1}, \boldsymbol{\theta}^{i-1}) + \beta^{full}_c)$. Assume that $\boldsymbol{\theta}^i$ minimizes $NLL_{\beta^{full}}(\boldsymbol{\theta}|\ \mathcal{X}_i, |\mathcal{X}_{i-1}|, \hat{\phi}^{1:i-1}, \boldsymbol{\theta}^{i-1})$, a solution maximizing $LL(\boldsymbol{\theta}|\mathcal{X}_i, |\mathcal{X}_{i-1}|,\ \hat{\phi}^{1:i-1}, \boldsymbol{\theta}^{i-1})$.

Since $E_{\mathbb{X}}[\phi_k] \in \left[0, \frac{1}{(1-\gamma)}\right]$, we get $(1-\gamma)E_{\mathbb{X}}[\phi_k] \in [0,1]$. We multiply the relaxed constraint with $(1-\gamma)$ and define the negation of constraint as following event: $A : \left|(1-\gamma)E_{\mathbb{X}}[\phi_k] - (1-\gamma)\hat{\phi}^{1:i}_k\right| > (1-\gamma)\beta^{full}_c = \varepsilon$ for some $k \in \{1 \dots K\}$. A can be decomposed into following feature specific events

$$A_k : (1-\gamma)\left|E_{\mathbb{X}}[\phi_k] - \hat{\phi}^{1:i}_k\right| > \varepsilon,$$

where $k \in \{1, 2 \ldots K\}$. We divide this constraint on absolute value further in two signed events:

$$(A_k)_1 : (1-\gamma)E_{\mathbb{X}}[\phi_k] - (1-\gamma)\hat{\phi}_k^{1:i} > \varepsilon$$

$$(A_k)_2 : -(1-\gamma)E_{\mathbb{X}}[\phi_k] + (1-\gamma)\hat{\phi}_k^{1:i} > \varepsilon$$

Then event A is same as logical disjunction $(A_1)_1 \vee (A_1)_2 \vee (A_2)_1 \ldots$.

Applying Hoeffding's inequality, the upper bound of probability of each signed event is given by: $P((A_k)_1) \leq \exp(-2\varepsilon^2 \ |\mathcal{X}_{1:i}|) = \frac{\delta}{2K}(say), P((A_k)_2) \leq \frac{\delta}{2K}$. Applying aforesaid bounds to events for each of the K features, we get $2K$ events with exactly same upper bound $\frac{\delta}{2K}$ on their respective probabilities. We use Fretchet's inequality to derive an upper bound for the disjunction:

$P(A) = P((A_1)_1 \vee (A_1)_2 \vee (A_2)_1 \ldots) \leq \min(1, P((A_1)_1) + P((A_1)_2) + P((A_2)_1) \ldots)$

As each of the probabilities in RHS are bounded from above by $\frac{\delta}{2K}$, their sum is bounded as:

$P(A) \leq \min(1, \sum_1^{2K} \frac{\delta}{2K}) = \min(1, \delta)$

Reverting to the negation of A, the probability that $\left|(1-\gamma)E_{\mathbb{X}}[\phi_k] - (1-\gamma)\hat{\phi}_k^{1:i}\right| \leq \varepsilon \ \forall k \in \{1 \ldots K\}$ is at least $1 - \min(1, \delta) = \max(0, 1-\delta)$.

To keep reward value bounded, IRL assumes $||\boldsymbol{\theta}^*||_1 \leq 1$ for all $\boldsymbol{\theta}^*$. Using the assumption and Theorem 1 in [2], we get error bound:

For every $\boldsymbol{\theta}^* \in [0,1]^K$, $NLL_{\beta full}(\boldsymbol{\theta}^i|\mathcal{X}_i, |\mathcal{X}_{i-1}|, \hat{\phi}^{1:i-1}, \boldsymbol{\theta}^{i-1})$ $-NLL_{\beta full}(\boldsymbol{\theta}^*|\mathcal{X}_i, |\mathcal{X}_{i-1}|, \hat{\phi}^{1:i-1}, \boldsymbol{\theta}^{i-1}) \leq 2\sum_1^K \beta_c^{full} = 2K \ \beta_c^{full} = \frac{2K\varepsilon}{(1-\gamma)}$ with probability at least $\max(0, 1-\delta)$, where $\delta = \ 2K \exp\big(-2\varepsilon^2|\mathcal{X}_{1:i}|\big)$.

We modify the bound in the form of positive log-likelihood of expert's policy, by using relation $NLL_{\beta full}(\boldsymbol{\theta}^*|\mathcal{X}_{1:i}) = (-LL(\boldsymbol{\theta}^*|\mathcal{X}_{1:i}) + \sum_1^K \beta_k^{full}|\theta_k|)$ and $\boldsymbol{\theta}^* = \boldsymbol{\theta}_E$.

Then, with $\mathcal{X}_{1:i}$ as input, with probability at least $\max(0, 1-\delta)$,

$$NLL_{\beta full}(\boldsymbol{\theta}^i|\mathcal{X}_i, |\mathcal{X}_{i-1}|, \hat{\phi}^{1:i-1}, \boldsymbol{\theta}^{i-1}) - NLL_{\beta full}(\boldsymbol{\theta}_E|\mathcal{X}_{1:i})$$
$$= LL(\boldsymbol{\theta}_E|\mathcal{X}_{1:i}) - LL(\boldsymbol{\theta}^i|\mathcal{X}_i, |\mathcal{X}_{i-1}|, \hat{\phi}^{1:i-1}, \boldsymbol{\theta}^{i-1}) \leq \frac{2K\varepsilon}{(1-\gamma)}$$

Insight: Likelihood-loss $LL(\boldsymbol{\theta}_E|\mathcal{X}_{1:i}) - LL(\boldsymbol{\theta}^i|\mathcal{X}_i, |\mathcal{X}_{i-1}|, \hat{\phi}^{1:i-1}, \boldsymbol{\theta}^{i-1})$ for MAXENTIRL gets smaller and the learned weights $\boldsymbol{\theta}^i$ is getting closer to the best weights possible (expert's weights θ_E) with more training trajectories or higher $|\mathcal{X}_{1:i}|$. The confidence of convergence $1-\delta$ increases with more training (higher $|\mathcal{X}_{1:i}|$), more room for error (higher ε) and less features (lower K).

□

PROOF OF LEMMA 1. Log-likelihood of demonstrated behavior can be split as

$$LL(\boldsymbol{\theta}^i|\boldsymbol{\mathcal{Y}}_{1:i}) = LL(\boldsymbol{\theta}^i|\boldsymbol{\mathcal{Y}}_i, |\boldsymbol{\mathcal{Y}}_{i-1}|, \hat{\phi}_{\boldsymbol{\theta}^{i-1}}^{Z|_Y,1:i-1}, \boldsymbol{\theta}^{i-1})$$
$$= \sum_{Y \in \boldsymbol{\mathcal{Y}}_{1:i}} \tilde{Pr}(Y) \ log \ Pr(Y; \boldsymbol{\theta})$$
$$= \sum_{Y \in \boldsymbol{\mathcal{Y}}_{1:i}} \tilde{Pr}(Y) \sum_{Z \in \mathbb{Z}} Pr(Z|Y; \boldsymbol{\theta}^i) \ log \ Pr(Y, Z; \boldsymbol{\theta}) +$$
$$(- \sum_{Y \in \boldsymbol{\mathcal{Y}}_{1:i}} \tilde{Pr}(Y) \sum_{Z \in \mathbb{Z}} Pr(Z|Y; \boldsymbol{\theta}^i) \ log \ Pr(Z|Y; \boldsymbol{\theta}))$$
$$= Q(\boldsymbol{\mathcal{Y}}_i, |\boldsymbol{\mathcal{Y}}_{i-1}|, \hat{\phi}_{\boldsymbol{\theta}^{i-1}}^{Z|_Y,1:i-1}, \boldsymbol{\theta}, \boldsymbol{\theta}^{i-1}) + C(\boldsymbol{\mathcal{Y}}_{1:i}, \boldsymbol{\theta}, \boldsymbol{\theta}^i)$$

Here $\tilde{Pr}$ is distribution of trajectories in observed training data ($\sum_{X \in \mathcal{X}} \tilde{Pr}(X)[\cdot]$ and $\frac{1}{|\mathcal{X}|} \sum_{X \in \mathcal{X}} [\cdot]$ can be used interchangeably). EM method maximizes the log-likelihood by maximizing only Q value over $\boldsymbol{\theta}$; and $\boldsymbol{\theta} = \boldsymbol{\theta}^i$ maximizes

$Q(\boldsymbol{\mathcal{Y}}_i, |\boldsymbol{\mathcal{Y}}_{i-1}|, \hat{\phi}_{\boldsymbol{\theta}^{i-1}}^{Z|_Y,1:i-1}, \boldsymbol{\theta}, \boldsymbol{\theta}^{i-1})$ ([3]). After all the EM iterations for current session i, the final Q value is $Q(\boldsymbol{\mathcal{Y}}_i, |\boldsymbol{\mathcal{Y}}_{i-1}|, \hat{\phi}_{\boldsymbol{\theta}^{i-1}}^{Z|_Y,1:i-1}, \boldsymbol{\theta}^i, \boldsymbol{\theta}^i)$. Therefore, the difference in the likelihoods achieved by weights learned in consecutive sessions can be expressed as a difference in Q values. Note that LME IRL learns reward weights by inferring the maximum entropy distribution $Pr(X; \boldsymbol{\theta}) = \frac{\exp(\sum_k \theta_k f_k(X))}{\Omega_{\boldsymbol{\theta}}^X}$, where $\Omega_{\boldsymbol{\theta}}^X = \sum_{X \in \mathbb{X}} \exp(\sum_k \theta_k f_k(X))$ and $X = (Y, Z)$. Expand Q value as $Q(\boldsymbol{\mathcal{Y}}_i, |\boldsymbol{\mathcal{Y}}_{i-1}|, \hat{\phi}_{\boldsymbol{\theta}^{i-1}}^{Z|_Y,1:i-1}, \boldsymbol{\theta}^i, \boldsymbol{\theta}^i) = \sum_{Y \in \boldsymbol{\mathcal{Y}}_{1:i}} \tilde{Pr}(Y) \sum_{Z \in \mathbb{Z}} Pr(Z|Y; \boldsymbol{\theta}^i) \ \log \left(\frac{\exp(\sum_k \theta_k^i f_k((Y,Z)))}{\Omega_{\boldsymbol{\theta}^i}^{(Y,Z)}} \right) = \sum_k \theta_k^i \cdot \sum_{Y \in \boldsymbol{\mathcal{Y}}_{1:i}} \tilde{Pr}(Y) \sum_{Z \in \mathbb{Z}} Pr(Z|Y; \boldsymbol{\theta}^i) f_k((Y,Z)) - \log \ \Omega_{\boldsymbol{\theta}^i}^{(Y,Z)} = \sum_k \theta_k^i \cdot \hat{\phi}_{\boldsymbol{\theta}^i,k}^{Z|_Y,1:i} - \log \ \Omega_{\boldsymbol{\theta}^i}^{(Y,Z)}$.

Therefore the improvement in log likelihood over session i is

$$LL(\boldsymbol{\theta}^i|\boldsymbol{\mathcal{Y}}_i, |\boldsymbol{\mathcal{Y}}_{i-1}|, \hat{\phi}_{\boldsymbol{\theta}^{i-1}}^{Z|_Y,1:i-1}, \boldsymbol{\theta}^{i-1}) - LL(\boldsymbol{\theta}^{i-1}|\boldsymbol{\mathcal{Y}}_{i-1}, |\boldsymbol{\mathcal{Y}}_{i-2}|, \hat{\phi}_{\boldsymbol{\theta}^{i-2}}^{Z|_Y,1:i-2}, \boldsymbol{\theta}^{i-2})$$
$$= Q(\boldsymbol{\mathcal{Y}}_i, |\boldsymbol{\mathcal{Y}}_{i-1}|, \hat{\phi}_{\boldsymbol{\theta}^{i-1}}^{Z|_Y,1:i-1}, \boldsymbol{\theta}^i, \boldsymbol{\theta}^i) - Q(\boldsymbol{\mathcal{Y}}_{i-1}, |\boldsymbol{\mathcal{Y}}_{i-2}|, \hat{\phi}_{\boldsymbol{\theta}^{i-2}}^{Z|_Y,1:i-2}, \boldsymbol{\theta}^{i-1}, \boldsymbol{\theta}^{i-1})$$
$$= \sum_k \theta_k^i \hat{\phi}_{\boldsymbol{\theta}^i,k}^{Z|_Y,1:i} - \log \ \Omega_{\boldsymbol{\theta}^i}^{(Y,Z)} - \sum_k \theta_k^{i-1} \hat{\phi}_{\boldsymbol{\theta}^{i-1},k}^{Z|_Y,1:i-1} + \log \ \Omega_{\boldsymbol{\theta}^{i-1}}^{(Y,Z)}$$
$$= \log \frac{\Omega_{\boldsymbol{\theta}^{i-1}}^{(Y,Z)}}{\Omega_{\boldsymbol{\theta}^i}^{(Y,Z)}} + \sum_k \left(\theta_k^i \frac{|\boldsymbol{\mathcal{Y}}_{1:i-1}|}{|\boldsymbol{\mathcal{Y}}_i| + |\boldsymbol{\mathcal{Y}}_{1:i-1}|} - \theta_k^{i-1} \right) \hat{\phi}_{\boldsymbol{\theta}^{i-1},k}^{Z|_Y,1:i-1} + \sum_k \left(\theta_k^i \frac{1}{|\boldsymbol{\mathcal{Y}}_i| + |\boldsymbol{\mathcal{Y}}_{1:i-1}|} \hat{\phi}_{\boldsymbol{\theta}^i,k}^{Z|_Y,i} \right)$$

(substitute $\hat{\phi}_{\boldsymbol{\theta}^i,k}^{Z|_Y,1:i}$ using Eq. 2 and simplifying)

The final expression is minimized only for $\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1}$ when $|\boldsymbol{\mathcal{Y}}_{1:i-1}| \gg |\boldsymbol{\mathcal{Y}}_i|$, i.e., when a significant amount of training data has been accumulated. The expression is also concave in parameter $\boldsymbol{\theta}^i$. Therefore, $LL(\boldsymbol{\theta}^i|\boldsymbol{\mathcal{Y}}_i, |\boldsymbol{\mathcal{Y}}_{i-1}|, \hat{\phi}_{\boldsymbol{\theta}^{i-1}}^{Z|_Y,1:i-1}, \boldsymbol{\theta}^{i-1}) - LL(\boldsymbol{\theta}^{i-1}|\boldsymbol{\mathcal{Y}}_{i-1}, |\boldsymbol{\mathcal{Y}}_{i-2}|, \hat{\phi}_{\boldsymbol{\theta}^{i-2}}^{Z|_Y,1:i-2}, \boldsymbol{\theta}^{i-2}) \geq 0$ for consecutive sessions thereafter.

Hence, the LME I2RL is proved to converge over sequence of sessions, yielding a feasible log-linear solution to latent-MAXENT and corresponding weights solving IRL.

□

PROOF OF LEMMA 2. We define the event

$A_k : (1-\gamma)\left|E_{\mathbb{X}}[\phi_k] - \hat{\phi}_k^{1:i}\right| > \varepsilon, k \in \{1, 2 \dots K\}$.

Applying Hoeffding's inequality for A_k, we get

$P(A_k) \leq 2\exp(-2\varepsilon^2|\mathcal{X}_{1:i}|) \leq \frac{\delta}{K}$ for any $k \in \{1, 2 \dots K\}$, and for the same ε, δ as in Theorem 1. Similarly, for partial observation, given ε_s as the bound on the error in sampling based approximation of $\hat{\phi}_l^{1:i}$ as $\hat{\phi}_{\boldsymbol{\theta}^i,l}^{Z|_Y,1:i}$, and n_s samples, let us define the event

$B_l : (1-\gamma)\left|\hat{\phi}_l^{1:i} - \hat{\phi}_{\boldsymbol{\theta}^i,l}^{Z|_Y,1:i}\right| > \varepsilon_s, l \in \{1, 2 \dots K\}$.

Similar to procedure for $P(A_k)$, applying Hoeffding bound gives us $P(B_l) < \frac{\delta_s}{K}, \delta_s = 2K\exp(-2(\varepsilon_s)^2 n_s)$.

Applying Fretchets inequality over both sets A and B of events gives us:

$P\left((\cup_k A_k) \vee (\cup_l B_l)\right) < \min(1, \sum_{k=1}^{K}\frac{\delta}{K} + \sum_{l=1}^{K}\frac{\delta_s}{K}) = \min(1, \delta + \delta_s)$.

That is, $P\left(\exists k, l, s.t. A_k \vee B_l\right) < \min(1, \delta + \delta_s)$. Taking complement, $P\left(\forall k, l, \overline{A}_k \wedge \overline{B}_l\right) \geq \max(0, 1 - \delta - \delta_s)$. But $\forall k, l, \overline{A}_k \wedge \overline{B}_l$ implies that $\forall k$:

$(1-\gamma)(\left|E_{\mathbb{X}}[\phi_k] - \hat{\phi}_k^{1:i}\right| + \left|\hat{\phi}_k^{1:i} - \hat{\phi}_{\boldsymbol{\theta}^i,k}^{Z|_Y,1:i}\right|) \leq \varepsilon + \varepsilon_s$

Calling $(\varepsilon + \varepsilon_s) = 2\varepsilon_l$, and $(\delta + \delta_s) = \delta_l$ we get

$P\big(\forall k, (1-\gamma)(\left|E_{\mathbb{X}}[\phi_k] - \hat{\phi}_k^{1:i}\right| + \left|\hat{\phi}_k^{1:i} - \hat{\phi}_{\boldsymbol{\theta}^i,k}^{Z|_Y,1:i}\right|) \leq 2\varepsilon_l\big) \geq \max(0, 1 - \delta_l)$.

Using inequality $\left|E_{\mathbb{X}}[\phi_k] - \hat{\phi}_{\boldsymbol{\theta}^i,k}^{Z|_Y,1:i}\right| \leq \left|E_{\mathbb{X}}[\phi_k] - \hat{\phi}_k^{1:i}\right| + \left|\hat{\phi}_k^{1:i} - \hat{\phi}_{\boldsymbol{\theta}^i,k}^{Z|_Y,1:i}\right|$, we get:

$P\left(\forall k, (1-\gamma)(\left|E_{\mathbb{X}}[\phi_k] - \hat{\phi}_{\boldsymbol{\theta}^i,k}^{Z|_Y,1:i}\right|) \leq 2\varepsilon_l\right) \geq \max(0, 1 - \delta_l)$.

□

PROOF OF THEOREM 2. Latent maximum entropy IRL problem is equivalent to $\max_{\boldsymbol{\theta}} \sum_{Y \in \boldsymbol{\mathcal{Y}}_{1:i}} \tilde{Pr}(Y) \log Pr(Y|\boldsymbol{\theta})$ (Section 3.3, [1]) or $\max_{\boldsymbol{\theta}} LL(\boldsymbol{\theta}^i|\boldsymbol{\mathcal{Y}}_i, |\boldsymbol{\mathcal{Y}}_{i-1}|, \hat{\phi}_{\boldsymbol{\theta}^{i-1}}^{Z|_Y,1:i-1}, \boldsymbol{\theta}^{i-1})$.

Relaxed constraint latent maximum entropy IRL is:

$$\begin{aligned} &\max_{\Delta} \left(-\textstyle\sum_{X \in \mathbb{X}} Pr(X)\ log\ Pr(X)\right) \\ &\textbf{subject to } \textstyle\sum_{X \in \mathbb{X}} Pr(X) = 1 \\ &\left|E_{\mathbb{X}}[\phi_k] - \hat{\phi}_{\boldsymbol{\theta}^i,k}^{Z|_Y,1:i}\right| \leq \beta_k\ \forall k \in \{1 \dots K\} \end{aligned} \tag{5}$$

Here $\beta \in \mathbb{R}^K$ is an estimate of vector of upper bounds on the differences between $E_{\mathbb{X}}[\phi_k]$ and $\hat{\phi}_{\boldsymbol{\theta}^i,k}^{Z|_Y,1:i}$.

The form of relaxed latent maximum entropy problem and the likelihood for that problem is no different than those for relaxed maximum entropy. Starting from results in Lemma 2, assuming $\beta_k = \beta_c = 2\varepsilon_l/(1-\gamma)$ for all $k \in \{1 \dots K\}$ and using steps similar to the proof of Theorem 1, we get

$LL(\boldsymbol{\theta}_E|\boldsymbol{\mathcal{Y}}_{1:i}) - LL(\boldsymbol{\theta}^i|\boldsymbol{\mathcal{Y}}_i, |\boldsymbol{\mathcal{Y}}_{i-1}|, \hat{\phi}_{\boldsymbol{\theta}^{i-1}}^{Z|_Y,1:i-1}, \boldsymbol{\theta}^{i-1}) \leq \frac{4K\varepsilon_l}{(1-\gamma)}$ with probbility at least $\max(0, 1 - \delta_l)$.

□

REFERENCES

[1] Kenneth Bogert, Jonathan Feng-Shun Lin, Prashant Doshi, and Dana Kulic. 2016. Expectation-Maximization for Inverse Reinforcement Learning with Hidden Data. In *2016 International Conference on Autonomous Agents and Multiagent Systems*. 1034–1042.

[2] Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. 2004. Performance Guarantees for Regularized Maximum Entropy Density Estimation. In *Learning Theory*, John Shawe-Taylor and Yoram Singer (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 472–486.

[3] Shaojun Wang and Dale Schuurmans Yunxin Zhao. 2012. The Latent Maximum Entropy Principle. *ACM Transactions on Knowledge Discovery from Data* 6, 8 (2012).