Appendix: Online Inverse Reinforcement Learning with Learned Observation Model

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

Bikramjit Banerjee
School of CSCE
Univ. Southern Mississippi, Hattiesburg, MS
United States of America
Bikramjit.Banerjee@usm.edu

Contents

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

Appendix A

Open-Source Code Links

The MDP definitions for our two experimental domains are available at: `https://github.com/s-arora-1987/sorting_patrol_MDP_irl`.

RIMEO implementation is available at: `https://github.com/s-arora-1987/irld`.

Finally, the Gazebo simulation with Sawyer is available at: `https://github.com/thinclab/sawyer_irl_project`.

Algorithm 1 RIMEO

WS (window size) $\leftarrow 5$; max-restarts $\leftarrow 5$; $i \leftarrow 1$; $\Xi_{d,1:i-1} \leftarrow \emptyset$; $\hat{\phi}^{1:i-1}_{\boldsymbol{\theta}^{i-1},k} \leftarrow 0$; $[\boldsymbol{\theta}^0]_k \sim$ uniform$(0,1)$;$P^*_{1:i-1}(\psi) \sim$ uniform$(0,1)$
while $std\_dev\_z > \rho$ **do**
  $P^*_{1:i} \leftarrow P^*_{1:i-1}$
  Compute $\hat{O}_o$ using scores for $\Xi_{d,i}$ from Eq. 8.
  repeat
    Compute $\mathcal{L}', \nabla\mathcal{L}'$ using $P^*_{1:i}(\psi)$ and $\Xi_{d,i}$.
    $P^*_{1:i}(\psi) \leftarrow$ update-step-LBFGS$(\mathcal{L}', \nabla\mathcal{L}')$
  until $||\nabla\mathcal{L}'||_1 \approx 0$
  Update learned observation model using $P^*_{1:i}$ in Eq. 7.
  repeat
    compute $\hat{\phi}^i_{\boldsymbol{\theta}^i}$ and $\hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,k}$ using Eqs. 10, 11.
    $|\Xi_{d,1:i}| \leftarrow |\Xi_{d,1:i-1}| + |\Xi_{d,i}|$
    $\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}^{i-1}, t \leftarrow 1$
    repeat
      Compute $\pi^*_{E,(t-1)}$ using $\boldsymbol{\theta}_{(t-1)}$ and $E_\Xi[\phi_k]$ using trajectories sampled from $\pi^*_{E,(t-1)}$.
      $z_{(t-1)} \leftarrow \hat{\phi}^{1:i}_{\boldsymbol{\theta}^i} - E_\Xi[\phi]$ {gradient}
      $\boldsymbol{\theta}_{t,k} \leftarrow \frac{\boldsymbol{\theta}_{(t-1),k} \exp(-\eta z_{(t-1),k})}{\sum_{k=1}^{K} \boldsymbol{\theta}_{(t-1),k} \exp(-\eta z_{(t-1),k})}$
      $t \leftarrow t + 1$
    until $|z_t| \leq \varepsilon_r/(1-\gamma)$
    $j \leftarrow j + 1$
  until $j >$ max-restarts
  Compute $\hat{\pi}_i$ using learned reward $\boldsymbol{\theta}^i \leftarrow \boldsymbol{\theta}_t$.
  $i \leftarrow i + 1$ {next session}
  $z_i \leftarrow z_t$; mov-window-z $\leftarrow [z_{i-WS}, \ldots, z_i]$
  $std\_dev\_z \leftarrow$ std-dev(mov-window-z)

Appendix B

Proof of Lemma 1

LEMMA 1 (MONOTONICITY). *The demonstration likelihood increases monotonically with each new session,* $LL(\boldsymbol{\theta}^i|\Xi_{d,i}, \alpha_{1:i-1}, \boldsymbol{\theta}^{i-1}) - LL(\boldsymbol{\theta}^{i-1}|\Xi_{d,i-1}, \alpha_{1:i-2}, \boldsymbol{\theta}^{i-2}) \geqslant 0$, *when* $|\Xi_{d,1:i-1}| \gg |\Xi_{d,i}|$.

Proof: Log-likelihood of demonstrated behavior can be split as

$$
\begin{aligned}
&LL(\boldsymbol{\theta}^i|\Xi_{d,i}, \alpha_{1:i-1}, \boldsymbol{\theta}^{i-1})\\
&= \sum_{\xi' \in \Xi_{d,1:i}} \tilde{P}(\xi') \log P(\xi'; \boldsymbol{\theta})\\
&= \sum_{\xi' \in \Xi_{d,1:i}} \tilde{P}(\xi') \sum_{\xi \in \Xi} P(\xi|\xi'; \boldsymbol{\theta}^i) \log P(\xi, \xi'; \boldsymbol{\theta}) + (- \sum_{\xi' \in \Xi_{d,1:i}} \tilde{P}(\xi) \sum_{\xi \in \Xi} P(\xi|\xi'; \boldsymbol{\theta}^i) \log P(\xi|\xi'; \boldsymbol{\theta}))\\
&= Q(\Xi_{d,1:i}, \boldsymbol{\theta}^i) + C(\Xi_{d,1:i}, \boldsymbol{\theta}^i)
\end{aligned}
$$

Here $\tilde{P}$ is distribution of trajectories in observed training data ($\sum\limits_{\xi' \in \Xi_{d,1:i}} \tilde{P}(\xi')[\cdot]$ and $\frac{1}{|\Xi_{d,1:i}|} \sum\limits_{\xi' \in \Xi_{d,1:i}} [\cdot]$ can be used interchangeably). The EM method maximizes the log-likelihood by maximizing only Q value over $\boldsymbol{\theta}$; and $\boldsymbol{\theta} = \boldsymbol{\theta}^i$ maximizes $Q(\Xi_{d,1:i}, \boldsymbol{\theta}^i)$ ([1]). After all the EM iterations for current session i, the final Q value is $Q(\Xi_{d,1:i}, \boldsymbol{\theta}^i)$. Therefore, the difference in the likelihoods achieved by weights learned in consecutive sessions can be expressed as a difference in Q values. Note that Robust IRL learns reward weights by inferring the maximum entropy distribution $P(\xi, \xi'; \boldsymbol{\theta}) = \frac{\exp(\sum_k \theta_k f_k(\xi))}{\Omega^{\Xi}_{\boldsymbol{\theta}}}$ (Equation 15 in [2]), where $\Omega^{\Xi}_{\boldsymbol{\theta}} = \sum_{\xi \in \Xi} \exp(\sum_k \theta_k f_k(\xi))$. Expand Q value as $Q(\Xi_{d,1:i}, \boldsymbol{\theta}^i) = \sum\limits_{\xi' \in \Xi_{d,1:i}} \tilde{P}(\xi') \sum\limits_{\xi \in \xi} P(\xi|\xi'; \boldsymbol{\theta}^i) \log \left(\frac{\exp(\sum_k \theta^i_k f_k(\xi))}{\Omega^{\Xi}_{\boldsymbol{\theta}^i}} \right) = \sum_k \theta^i_k \cdot \sum\limits_{\xi' \in \Xi_{d,1:i}} \tilde{P}(\xi') \sum\limits_{\xi \in \Xi} P(\xi|\, \xi'; \boldsymbol{\theta}^i) f_k(\xi) - \log \Omega^{\Xi}_{\boldsymbol{\theta}^i} = \sum_k \theta^i_k \cdot \hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,k} - \log \Omega^{\Xi}_{\boldsymbol{\theta}^i}$.

Therefore the improvement in log likelihood over session i is

$$
\begin{aligned}
&LL(\boldsymbol{\theta}^i|\Xi_{d,i}, \alpha_{1:i-1}, \boldsymbol{\theta}^{i-1}) - LL(\boldsymbol{\theta}^{i-1}|\Xi_{d,i-1}, \alpha_{1:i-2}, \boldsymbol{\theta}^{i-2})\\
&= Q(\Xi_{d,1:i}, \boldsymbol{\theta}^i) - Q(\Xi_{d,1:i-1}, \boldsymbol{\theta}^{i-1})\\
&= \sum_k \theta^i_k \hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,k} - \log \Omega^{\Xi}_{\boldsymbol{\theta}^i} - \sum_k \theta^{i-1}_k \hat{\phi}^{1:i-1}_{\boldsymbol{\theta}^{i-1},k} + \log \Omega^{\Xi}_{\boldsymbol{\theta}^{i-1}}\\
&= \log \frac{\Omega^{\Xi}_{\boldsymbol{\theta}^{i-1}}}{\Omega^{\Xi}_{\boldsymbol{\theta}^i}} + \sum_k \left(\theta^i_k \frac{|\Xi_{d,1:i-1}|}{|\Xi_{d,i}| + |\Xi_{d,1:i-1}|} - \theta^{i-1}_k \right) \hat{\phi}^{1:i-1}_{\boldsymbol{\theta}^{i-1},k} + \sum_k \left(\theta^i_k \frac{1}{|\Xi_{d,i}| + |\Xi_{d,1:i-1}|} \hat{\phi}^i_{\boldsymbol{\theta}^i,k} \right)
\end{aligned}
$$

(substitute $\hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,k}$ using Eq. 11 from main paper and simplifying)

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

Proof of Lemma 2

LEMMA 2 (CONSTRAINT BOUND). *Under the assumptions stated in Sec. 5.2 (main paper), the following holds with probability at least* $\max(0, 1 - \delta_r)$*:*

$$\left| (1-\gamma)\big(E_{\Xi}[\phi_k] - \hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,k}\big) \right|_1 \leqslant \varepsilon_r, k \in \{1, 2 \ldots K\}$$

where L is the maximum length of any trajectory, $\delta_r = \delta + \delta_s + \delta_o$ *and* $\varepsilon_r = \varepsilon + \varepsilon_s + L|\Psi|\varepsilon_o$, *and* ε, δ *are as defined in Theorem 1 in [3].*

Proof: Suppose the true (unknown) observation model $\forall o, g$ is $O^*_{o,g}$. Solving the NLP with the true observation model gives the true $P(\psi)$, since the constraint below is satisfied.

$$\prod_{\psi^{o,g}=1} P(\psi) \prod_{\psi^{o,g}=0} (1 - P(\psi)) = O^*_{o,g} \tag{1}$$

Using these true $P(\psi)$ instead of $P^*(\psi)$, we can generate a version of Eq. 11 (main paper):

$$\phi^{1:i}_{\boldsymbol{\theta},k} = \frac{1}{|\Xi_{d,1:i}|} \sum_{\xi' \in \Xi_{d,1:i}} \sum_{\xi \in \Xi} \eta P(\xi'|\xi) P(\xi; \boldsymbol{\theta}) f_k(\xi)$$

From the accumulated sessions, we get estimates of $O^*_{o,g}$, call it $\hat{O}_{o,g}$ (Eq. 8 in the main paper). We assume that this estimate satisfies Hoeffding bounds for *the observed state-action pairs*, viz., $P(|O^*_{o,g} - \hat{O}_{o,g}| \leq \epsilon_o) \geq 1 - \frac{\delta_o}{K}$, where $\delta_o = 2K|\Psi| \exp(-2\epsilon_o^2 n_o)$, n_o being the number of samples used to construct $\hat{O}_{o,g}$. The key issue is that this estimate may not be available yet for the $\langle s, a \rangle_o$ pairs that were not observed. Regardless, we assume that all features in Ψ are observed in the very first session. Hence, after solving the NLP, we obtain $\hat{O}_{o,g}$ for *all* o, g, using the $P^*(\psi)$ from observed $\langle s, a \rangle_o$s and

$$\hat{O}_{o,g} = \prod_{\psi^{o,g}=1} P^*(\psi) \prod_{\psi^{o,g}=0} (1 - P^*(\psi)) \tag{2}$$

Under the assumptions above, with probability $\geq 1 - \delta_o$, $\max_{\langle s,a \rangle_o} |O^*_{o,g} - \hat{O}_{o,g}| \leq \epsilon_o$, but only for the observed $\langle s, a \rangle_o$. Since $\max_{any\ \psi} |P(\psi) - P^*(\psi)| \leq 1$, in turn this yields $\max_{any \langle s,a \rangle_o} |O^*_{o,g} - \hat{O}_{o,g}| \leq |\Psi|\epsilon_o$. Consequently, if the length of trajectories is bounded by L, then with probability $\geq 1 - \frac{\delta_o}{K}$ we have $\forall k$

$$\begin{aligned}
|\phi^{1:i}_{\boldsymbol{\theta}^i,k} - \hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,k}| &= \frac{1}{|\Xi_{d,1:i}|} \sum_{\xi' \in \Xi_{d,1:i}} \sum_{\xi \in \Xi} f_k(\xi) \eta P(\xi; \boldsymbol{\theta}) |(P(\xi'|\xi) - P^*(\xi'|\xi))| \\
&= \frac{1}{|\Xi_{d,1:i}|} \sum_{\xi' \in \Xi_{d,1:i}} \sum_{\xi \in \Xi} f_k(\xi) \eta P(\xi; \boldsymbol{\theta}) |(\prod_{o,g} O^*_{o,g} - \prod_{o,g} \hat{O}_{o,g})| \\
&\leq \frac{1}{|\Xi_{d,1:i}|} \sum_{\xi' \in \Xi_{d,1:i}} \sum_{\xi \in \Xi} f_k(\xi) \eta P(\xi; \boldsymbol{\theta}) L \max_{any\ o} |O^*_{o,g} - \hat{O}_{o,g}| \\
&\leq L|\Psi|\epsilon_o/(1-\gamma)
\end{aligned}$$

The rest of the proof follows similar steps as in [3]. We define the events A_k, B_l, C_j as:

$A_k : (1-\gamma)\big|E_{\Xi}[\phi_k] - \hat{\phi}^{1:i}_k\big| > \varepsilon, k \in \{1, 2 \ldots K\}$.

Applying Hoeffding's inequality for A_k, we get $P(A_k) \leq 2\exp(-2\varepsilon^2 |\Xi_{d,1:i}|) \leq \frac{\delta}{K}$ for any $k \in \{1, 2 \ldots K\}$, and for the same ε, δ as in Theorem 1. Similarly, for noisy observation, given ε_s as the bound on the error in sampling based approximation of $\hat{\phi}^{1:i}_l$ as $\phi^{1:i}_{\boldsymbol{\theta}^i,l}$, and n_s samples, let us define the event

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

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

$C_j : (1-\gamma)\left|\phi^{1:i}_{\boldsymbol{\theta}^i,j} - \hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,j}\right| > L|\Psi|\varepsilon_o, j \in \{1, 2 \ldots K\}$. Then following the argument above, $P(C_j) < \frac{\delta_o}{K}$.

Applying Fretchets inequality over the sets A, B, and C of events gives us:

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

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

$(1-\gamma)(\left|E_{\Xi}[\phi_k] - \hat{\phi}^{1:i}_k\right| + \left|\hat{\phi}^{1:i}_k - \phi^{1:i}_{\boldsymbol{\theta}^i,k}\right| + \left|\phi^{1:i}_{\boldsymbol{\theta}^i,k} - \hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,k}\right|) \leq \varepsilon + \varepsilon_s + L|\Psi|\varepsilon_o$.

Hence $P\big(\forall k, (1-\gamma)(\left|E_{\Xi}[\phi_k] - \hat{\phi}^{1:i}_k\right| + \left|\hat{\phi}^{1:i}_k - \phi^{1:i}_{\boldsymbol{\theta}^i,k}\right| + \left|\phi^{1:i}_{\boldsymbol{\theta}^i,k} - \hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,k}\right|) \leq \varepsilon + \varepsilon_s + L|\Psi|\varepsilon_o\big) \geq \max(0, 1 - \delta - \delta_s - \delta_o)$.

Using $\left|E_\Xi[\phi_k] - \hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,k}\right| \le \left|E_\Xi[\phi_k] - \hat{\phi}^{1:i}_k\right| + \left|\hat{\phi}^{1:i}_k - \phi^{1:i}_{\boldsymbol{\theta}^i,k}\right| + \left|\phi^{1:i}_{\boldsymbol{\theta}^i,k} - \hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,k}\right|$, $\delta_r = \delta + \delta_s + \delta_o$, and $\varepsilon_r = \varepsilon + \varepsilon_s + L|\Psi|\varepsilon_o$, we get:

$P\left(\forall k, (1-\gamma)(\left|E_\Xi[\phi_k] - \hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,k}\right|) \le \varepsilon_r\right) \ge \max(0, 1-\delta_r).$ □

Proof of Theorem 1

THEOREM 1 (CONFIDENCE). *Let ε_r, δ_r be as defined in Lemma 2, and $\boldsymbol{\theta}^i$ be the solution of session i for RI2RL-MEOM. Then*

$$LL(\boldsymbol{\theta}_E|\Xi_{d,1:i}) - LL(\boldsymbol{\theta}^i|\Xi_{d,i}, \alpha_{1:i-1}, \boldsymbol{\theta}^{i-1}) \le \frac{2K\varepsilon_r}{(1-\gamma)},$$

with confidence at least $\max(0, 1-\delta_r)$, where $\boldsymbol{\theta}_E$ are the true weights of the expert.

Proof: Each session of RI2RL-MEOM solves a maximum entropy estimation problem for Robust IRL. By allowing a relaxation in the constraints for a session, we get

$$\begin{array}{l} \max\limits_{\Delta} \left(-\sum_{\xi' \in \Xi_{d,1:i}, \xi \in \Xi} P(\xi', \xi) \log P(\xi', \xi)\right) \\ \textbf{subject to} \quad \sum_{\xi' \in \Xi_{d,1:i}, \xi \in \Xi} P(\xi', \xi) = 1 \\ \left|E_\Xi[\phi_k] - \hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,k}\right| \le \beta_k \qquad \forall k \end{array} \tag{3}$$

where

$$E_\Xi[\phi_k] \triangleq \sum_{\xi \in \Xi, \xi' \in \Xi_{d,1:i}} P(\xi, \xi')\ f_k(\xi),\ k = 1 \ldots K \tag{4}$$

Here $\beta \in \mathbb{R}^K$ is a vector of upper bounds on the differences between feature expectations. Following the proofs by Dudik et al. [4], the above relaxed constraints problem is the same as $\min_{\boldsymbol{\theta}}(-\sum_{\xi \in \Xi_{d,1:i}} \tilde{P}(\xi) \quad \log\ P(\xi|\boldsymbol{\theta}) + \sum_k \beta_k|\theta_k|) = \min_{\boldsymbol{\theta}}(-LL(\boldsymbol{\theta}|\ \Xi_{d,i}, \alpha_{1:i-1},\ \boldsymbol{\theta}^{i-1}) + \sum_k \beta_k|\theta_k|) = \min_{\boldsymbol{\theta}} NLL_\beta(\boldsymbol{\theta}|\Xi_{d,i}, \alpha_{1:i-1}\ , \boldsymbol{\theta}^{i-1})$ (say). Here NLL = negative log likelihood.

The proof here is partially inspired from Corollary 1 in [4]. Let $\beta_k = \beta_c = \varepsilon/(1-\gamma)$ for all $k \in \{1 \ldots K\}$, where β_c is a constant because ε is a fixed input. For normalized exponentiated gradient descent used in reward-learning part of RI2RL session, $\sum_1^K |\theta_k| = 1$. Then, $NLL_\beta(\boldsymbol{\theta}|\Xi_{d,i}, \alpha_{1:i-1}, \boldsymbol{\theta}^{i-1}) = (-LL(\boldsymbol{\theta}|\Xi_{d,i}, \alpha_{1:i-1}, \boldsymbol{\theta}^{i-1}) + \beta_c \sum_1^k |\theta_k|) = (-LL(\boldsymbol{\theta}|\Xi_{d,i}, \alpha_{1:i-1}, \boldsymbol{\theta}^{i-1}) + \beta_c)$. Assume that $\boldsymbol{\theta}^i$ minimizes $NLL_\beta(\boldsymbol{\theta}|\ \Xi_{d,i}, \alpha_{1:i-1}, \boldsymbol{\theta}^{i-1})$, a solution maximizing $LL(\boldsymbol{\theta}|\Xi_{d,i}, \alpha_{1:i-1}, \boldsymbol{\theta}^{i-1})$.

Since $E_\Xi[\phi_k] \in \left[0, \frac{1}{(1-\gamma)}\right]$, we get $(1-\gamma)E_\Xi[\phi_k] \in [0,1]$. Using the result from the previous Lemma, the probability that $\left|(1-\gamma)E_\Xi[\phi_k] - (1-\gamma)\hat{\phi}^{1:i}_{\boldsymbol{\theta}^i,k}\right| \le \varepsilon_r\ \forall k \in \{1 \ldots K\}$ is at least $\max(0, 1-\delta_r)$. To keep the reward value bounded, IRL assumes $||\boldsymbol{\theta}^*||_1 \le 1$ for all $\boldsymbol{\theta}^*$. Using the assumption and Theorem 1 in [4], we get the following error bound:

For every $\boldsymbol{\theta}^* \in [0,1]^K$, $NLL_\beta(\boldsymbol{\theta}^i|\Xi_{d,i}, \alpha_{1:i-1}, \boldsymbol{\theta}^{i-1}) - NLL_\beta(\boldsymbol{\theta}^*|\Xi_{d,i}, \alpha_{1:i-1}, \boldsymbol{\theta}^{i-1}) \le 2\sum_1^K \beta_c = 2K\ \beta_c = \frac{2K\varepsilon_r}{(1-\gamma)}$, with probability at least $\max(0, 1-\delta_r)$.

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

Then, with $\Xi_{d,1:i}$ as input, with probability at least $\max(0, 1-\delta_r)$,

$$\begin{aligned} &NLL_\beta(\boldsymbol{\theta}^i|\Xi_{d,i}, \alpha_{1:i-1}, \boldsymbol{\theta}^{i-1}) - NLL_\beta(\boldsymbol{\theta}_E|\Xi_{d,1:i}) \\ &= LL(\boldsymbol{\theta}_E|\Xi_{d,1:i}) - LL(\boldsymbol{\theta}^i|\Xi_{d,i}, \alpha_{1:i-1}, \boldsymbol{\theta}^{i-1}) \le \frac{2K\varepsilon_r}{(1-\gamma)}. \end{aligned}$$

□

Appendix C (Features of Onion Sorting)

Reward Features

The 11 reward features $\phi_k(s, a)$ are:

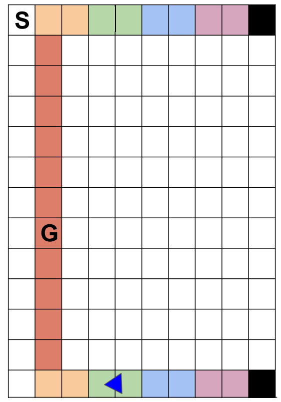
- *CreateList*(s,a): Roll all onions and create a list of predictions (blemished/unblemished/unknown);
- *ClaimNewOnion*(s,a): considers a new onion on table;
- *PickUnknown*(s,a): is 1 when onion with unknown prediction is picked;
- *AvoidNoOp*(s,a): the action a changes the state;
- *InspectNewOnion*(s,a): is 1 when an onion is inspected for the first time and a prediction is made for it;
- *GoodOnTable*(s,a): considered onion is unblemished and is placed on the table;
- *BlemishedNotOnTable*(s,a): onion is blemished and is not placed on the table;
- *GoodNotInBin*(s,a): onion is unblemished and is not placed in the bin;
- *BlemishedInBin*(s,a): onion is blemished and is placed in the bin;
- *PickBlemished*(s,a): onion with prediction blemished is picked;
- *EmptyList*(s,a): finish sorting bad onions out of the conveyor.

Observation Features

The 8 observation features, $\psi_j, j = 1, \ldots, 8$, are listed below. Each indicator $\psi_j^{o,g}$ takes the value 1 iff the predicate value is the same for both $\langle s, a \rangle_g$ and $\langle s, a \rangle_o$.

- *BlemishedOnion*: considered onion is blemished;
- *MoveWithHand*: onion moves with the hand;
- *StartFromConv*: onion was on the table before action;
- *LeavingAtEye*: onion leaves atEye location;
- *OnionToBin*: onion moves to the bin;
- *HandToBin*: hand moves to the bin;
- *OnionToTable*: onion moves to the table;
- *HandToTable*: hand moves to the table;

Appendix D (Features of Perimeter Patrol)



Reward Features

The 6 reward features $\phi_k(s, a)$, in the context of the above figure, are:

- *HasMoved*(s, a): true iff a in s makes the patroller change its grid cell;
- *Turn1*(s, a): true iff a in s makes the patroller turn (left or right) in the orange part of the hallway;
- *Turn2*(s, a): true iff a in s makes the patroller turn in the yellow part of hallway;
- *Turn3*(s, a): true iff a in s makes the patroller turn in the green part of hallway;
- *Turn4*(s, a): true iff a in s makes the patroller turn in the blue part of hallway;
- *Turn5*(s, a): true iff a in s makes the patroller turn in the magenta part of hallway.

A weight vector $\boldsymbol{\theta}_E$ for these features such as $\langle .57, 0, 0, 0, .43, 0 \rangle$ makes the patroller constantly execute a cyclic trajectory.

Observation Features

The observation feature set Ψ contains the following 4 binary predicates:

- *MoveForward:* patroller is moving forward;
- *TurnLeft:* patroller is turning left;
- *y is 0:* patroller location has $y = 0$;
- *TurnRight:* patroller is turning right;

Average of pairwise feature correlation from the patroller's demonstration is -0.14 (p-value 0.06), indicating that the features are reasonably independent.

References

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

[2] S. Shahryari and P. Doshi. Inverse Reinforcement Learning Under Noisy Observations. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, Richland, SC, 2017. International Foundation for Autonomous Agents and Multiagent Systems.

[3] S. Arora, P. Doshi, and B. Banerjee. I2RL: Online inverse reinforcement learning under occlusion. *Autonomous Agents and Multi-Agent Systems*, 35(1):4, Nov 2020. ISSN 1573-7454.

[4] M. Dudík, S. J. Phillips, and R. E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In J. Shawe-Taylor and Y. Singer, editors, *Learning Theory*, pages 472–486, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-27819-1.