Deep Adaptive Multi-Intention Inverse Reinforcement Learning - Supplementary Materials

Ariyan Bighashdel(✉), Panagiotis Meletis, Pavol Jancura, and Gijs Dubbelman

Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands
{a.bighashdel, p.c.meletis, p.jancura, g.dubbelman}@tue.nl

Abstract. Supplementary material for our ECMLPKDD 2021 paper titled "Deep Adaptive Multi-intention Inverse Reinforcement Learning"

Keywords: Inverse reinforcement learning · Multiple intentions · Deep learning.

1 Full Derivation of prior probabilities of intention vectors

We assume that we have $M-1$ demonstrated trajectories with a set of known latent intention vectors $\boldsymbol{H}^{-m} = \{\boldsymbol{\eta}^1, \boldsymbol{\eta}^2, ..., \boldsymbol{\eta}^{m-1}, \boldsymbol{\eta}^{m+1}, ..., \boldsymbol{\eta}^M\}$ with K intentions. Then, we have a new demonstrated trajectory $\boldsymbol{\tau}^m$ and the task is to obtain the latent intention vector $\boldsymbol{\eta}^m$, which can be a new intention $K+1$, and update the reward parameters Ψ. We are willing to consider growing/infinite number of intentions.

In the case of K intentions, we define a Categorical prior distribution over $\boldsymbol{H} = \{\boldsymbol{H}^{-m}, \boldsymbol{\eta}^m\}$:

$$\begin{aligned} p(\boldsymbol{H}|\boldsymbol{\phi}) &= \prod_{m=1}^{M} \mathrm{Cat}(\boldsymbol{\phi}) \\ &= \prod_{k=1}^{K} \phi_k^{M_k} \end{aligned} \tag{1}$$

where M_k is the number of trajectories with intention k and $\boldsymbol{\phi}$ is the vector of mixing coefficients $\boldsymbol{\phi} = \{\phi_1, \phi_2, ...\phi_K\}$ with prior distribution of:

$$\begin{aligned} p(\boldsymbol{\phi}) &= \mathrm{Dir}(\alpha/K) \\ &= \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \pi_k^{\alpha/K-1} \end{aligned} \tag{2}$$

where α is the concentration parameter. The main problematic variable as $K \to \infty$ are the mixing coefficients. We marginalize out $\boldsymbol{\phi}$:

$$\begin{aligned} p(\boldsymbol{H}) &= \int p(\boldsymbol{H}|\boldsymbol{\phi})p(\boldsymbol{\phi}) \\ &= \frac{\Gamma(\alpha)}{\Gamma(M+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(M_k + \alpha/K)}{\Gamma(\alpha/K)} \end{aligned} \tag{3}$$

Given that:

$$p(\boldsymbol{H}) = p(\boldsymbol{\eta}^m|\boldsymbol{H}^{-m})p(\boldsymbol{H}^{-m}) \tag{4}$$

we can define the conditional prior over $\boldsymbol{\eta}^m = \{\eta_1^m, \eta_2^m, ..., \eta_K^m\}$ as:

$$p(\eta_k^m = 1|\boldsymbol{H}^{-m}) = \frac{M_k^{-m} + \alpha/K}{M - 1 + \alpha} \tag{5}$$

where M_k^{-m} is the number of trajectories with intention k excluding $\boldsymbol{\tau}^m$. By letting $K \to \infty$, we reach:

$$p(\eta_k^m = 1|\boldsymbol{H}^{-m}) = \frac{M_k^{-m}}{M - 1 + \alpha} \tag{6}$$

where $p(\eta_k^m = 1|\boldsymbol{H}^{-m})$ is the prior probability of assigning the trajectory $\boldsymbol{\tau}^m$ to intention $k \in \{1, 2, ..., K\}$. Since:

$$\sum_{k=1}^{K} p(\eta_k^m = 1|\boldsymbol{H}^{-m}) = \frac{M-1}{M-1+\alpha} \neq 1 \tag{7}$$

we define $p(\eta_{K+1}^m = 1|\boldsymbol{H}^{-m})$ as the prior probability of assigning the trajectory $\boldsymbol{\tau}^m$ to intention $k+1$:

$$\begin{aligned} p(\eta_{K+1}^m = 1|\boldsymbol{H}^{-m}) &= 1 - \frac{M-1}{M-1+\alpha} \\ &= \frac{\alpha}{M-1+\alpha} \end{aligned} \tag{8}$$

Equations (6) and (8) are known as Chinese Restaurant Process [3].

2 Full Derivation of E-step and M-step

Given the predictive distribution for m^{th} trajectory:

$$p(\boldsymbol{\tau}^m|\boldsymbol{H}^{-m}, \Psi) = \sum_{k=1}^{K+1} p(\boldsymbol{\tau}^m|\eta_k^m = 1, \Psi)p(\eta_k^m = 1|\boldsymbol{H}^{-m}) \tag{9}$$

the following optimization problem can be defined $\forall m \in \{1, 2, ..., M\}$ by employing the exchangeability property [2]:

$$\max_{\Psi} L^m(\Psi) = \log \sum_{k=1}^{K+1} p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m}) \tag{10}$$

The parameters Ψ can be estimated via Expectation Maximization (EM) [1]. Differentiating the log-likelihood function $L(\Psi)$ with respect to $\psi \in \Psi$ yields:

$$\begin{aligned}
\nabla_\psi L^m &= \frac{\nabla_\psi \sum_{k=1}^{K+1} p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})}{\sum_{\hat{k}} p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})} \\
&= \sum_{k=1}^{K+1} \frac{\nabla_\psi p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})}{\sum_{\hat{k}} p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})}
\end{aligned} \tag{11}$$

A standard trick in setting up the EM procedure is to introduce the posterior distribution over the latent intention vector $\boldsymbol{\eta}^m$ [1]:

$$\begin{aligned}
\gamma_k^m = p(\eta_k^m = 1 | \boldsymbol{\tau}^m, \boldsymbol{H}^{-m}, \Psi) &= \frac{p(\boldsymbol{\tau}^m, \eta_k^m = 1 | \boldsymbol{H}^{-m}, \Psi)}{\sum_{\hat{k}=1}^{K+1} p(\boldsymbol{\tau}^m, \eta_{\hat{k}}^m = 1 | \boldsymbol{H}^{-m}, \Psi)} \\
&= \frac{p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})}{\sum_{\hat{k}=1}^{K+1} p(\boldsymbol{\tau}^m | \eta_{\hat{k}}^m = 1, \Psi) p(\eta_{\hat{k}}^m = 1 | \boldsymbol{H}^{-m})}
\end{aligned} \tag{12}$$

Now the term under summation in (11) can be written as::

$$\begin{aligned}
&\frac{\nabla_\psi p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})}{\sum_{\hat{k}=1}^{K+1} p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})} \\
&= \frac{p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})}{\sum_{\hat{k}=1}^{K+1} p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})} \frac{\nabla_\psi p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})}{p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})} \\
&= \gamma_k^m \frac{\nabla_\psi p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})}{p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})} \\
&= \gamma_k^m \nabla_\psi \log p(\boldsymbol{\tau}^m | \eta_k^m = 1, \Psi) p(\eta_k^m = 1 | \boldsymbol{H}^{-m})
\end{aligned} \tag{13}$$

Performing the differentiation of the second term in (13) yields:

$$
\begin{aligned}
&\nabla_\psi \log p(\boldsymbol{\tau}^m|\eta_k^m = 1, \Psi)p(\eta_k^m = 1|\boldsymbol{H}^{-m}) \\
&\quad = \nabla_\psi \log p(\boldsymbol{\tau}^m|\eta_k^m = 1, \Psi) + \cancelto{0}{\nabla_\psi \log p(\eta_k^m = 1|\boldsymbol{H}^{-m})} \\
&\quad = \nabla_\psi \log(\frac{\exp(R_k(\boldsymbol{\tau}^m, \psi_k))}{Z(k)}) \\
&\quad = \nabla_\psi (R_k(\boldsymbol{\tau}^m, \psi_k) - \log Z(k)) \\
&\quad = \nabla_\psi (R_k(\boldsymbol{\tau}^m, \psi_k) - \log \sum_{\tau} \exp(R_k(\boldsymbol{\tau}, \psi_k))) \\
&\quad = \frac{dR_k(\boldsymbol{\tau}^m, \psi_k)}{d\psi} - \frac{\sum_\tau \exp(R_k(\boldsymbol{\tau}, \psi_k)))\frac{dR_k(\boldsymbol{\tau}, \psi_k)}{d\psi}}{\sum_\tau \exp(R_k(\boldsymbol{\tau}, \psi_k)))} \\
&\quad = \frac{dR_k(\boldsymbol{\tau}^m, \psi_k)}{d\psi} - \sum_\tau p(\boldsymbol{\tau}|\eta_k = 1, \Psi)\frac{dR_k(\boldsymbol{\tau}, \psi_k)}{d\psi} \\
&\quad = (\boldsymbol{\mu}(\boldsymbol{\tau}^m) - \mathbb{E}_{p(\boldsymbol{\tau}|\eta_k=1,\Psi)}[\boldsymbol{\mu}(\boldsymbol{\tau})])^\intercal \frac{d\boldsymbol{R}_{\Psi_k}(\boldsymbol{\tau})}{d\psi}
\end{aligned}
\tag{14}
$$

Therefore (11) results in:

$$
\nabla_\psi L = \sum_{k=1}^{K+1} \gamma_k^m (\boldsymbol{\mu}(\boldsymbol{\tau}^m) - \mathbb{E}_{p(\boldsymbol{\tau}|\eta_k=1,\Psi)}[\boldsymbol{\mu}(\boldsymbol{\tau})])^\intercal \frac{d\boldsymbol{R}_{\Psi_k}(\boldsymbol{\tau})}{d\psi} \tag{15}
$$

which is knwon as the M-step. The posterior distribution over the latent intention vector $\boldsymbol{\eta}^m$ can be obtained as:

$$
\begin{aligned}
\gamma_k^m &= \frac{p(\boldsymbol{\tau}^m|\eta_k^m = 1, \Psi)p(\eta_k^m = 1|\boldsymbol{H}^{-m})}{\sum_{\hat{k}=1}^{K+1} p(\boldsymbol{\tau}^m|\eta_{\hat{k}}^m = 1, \Psi)p(\eta_{\hat{k}}^m = 1|\boldsymbol{H}^{-m})} \\
&= \frac{b_0(s_0)\prod_{t=0}^{T-1} T(s_{t+1}|s_t, a_t)\pi_k(a_t|s_t)p(\eta_k^m = 1|\boldsymbol{H}^{-m})}{\sum_{\hat{k}=1}^{K+1} b_0(s_0)\prod_{t=0}^{T-1} T(s_{t+1}|s_t, a_t)\pi_{\hat{k}}(a_t|s_t)p(\eta_{\hat{k}}^m = 1|\boldsymbol{H}^{-m})} \\
&= \frac{\prod_{t=0}^{T-1} \pi_k(a_t|s_t)p(\eta_k^m = 1|\boldsymbol{H}^{-m})}{\sum_{\hat{k}=1}^{K+1} \prod_{t=0}^{T-1} \pi_{\hat{k}}(a_t|s_t)p(\eta_{\hat{k}}^m = 1|\boldsymbol{H}^{-m})}
\end{aligned}
\tag{16}
$$

Using (6) and (8) yields $\forall k \in \{1, 2, ..., K\}$:

$$
\gamma_k^m = \frac{M_k^{-m}\prod_{t=0}^{T-1} \pi_k(a_t|s_t)}{\alpha \prod_{t=0}^{T-1} \pi_{K+1}(a_t|s_t) + \sum_{\hat{k}=1}^{K} M_k^{-m} \prod_{t=0}^{T-1} \pi_{\hat{k}}(a_t|s_t)} \tag{17}
$$

and for $K + 1$:

$$
\gamma_k^m = \frac{\alpha \prod_{t=0}^{T-1} \pi_k(a_t|s_t)}{\alpha \prod_{t=0}^{T-1} \pi_{K+1}(a_t|s_t) + \sum_{\hat{k}=1}^{K} M_k^{-m} \prod_{t=0}^{T-1} \pi_{\hat{k}}(a_t|s_t)} \tag{18}
$$

Which are known as the E-step.

3 Full Derivation of likelihood ratio

The likelihood ratio for the m^{th} trajectory is obtained as:

$$\begin{aligned}\frac{p(\boldsymbol{\tau}^m|\eta_{k^*}^{*m}=1,\Psi)}{p(\boldsymbol{\tau}^m|\eta_k^m=1,\Psi)} &= \frac{b_0(s_0)\prod_{t=1}^{T_\tau} T(s_{t+1}|s_t,a_t)\pi_{k^*}(a_t^m|s_t^m)}{b_0(s_0)\prod_{t=1}^{T_\tau} T(s_{t+1}|s_t,a_t)\pi_k(a_t^m|s_t^m)} \\ &= \frac{\prod_{t=1}^{T_\tau}\pi_{k^*}(a_t^m|s_t^m)}{\prod_{t=1}^{T_\tau}\pi_k(a_t^m|s_t^m)}\end{aligned} \tag{19}$$

with $k \in \{1,2,...,K\}$ and $k^* \in \{1,2,...,K,K+1\}$.

References

1. Bishop, C.M.: Pattern recognition and machine learning. springer (2006)
2. Gershman, S.J., Blei, D.M.: A tutorial on Bayesian nonparametric models. Journal of Mathematical Psychology **56**(1), 1–12 (2012)
3. Li, Y., Schofield, E., Gönen, M.: A tutorial on Dirichlet process mixture modeling. Journal of Mathematical Psychology **91**, 128–144 (2019)