Marginal MAP Estimation for Inverse RL under Occlusion with Observer Noise (Supplementary material)

Pranath Sengadu Suresh

1THINC Lab, Department of Computer Science, University of Georgia, Athens, GA 30606, USA.

1 EXTENDED DERIVATION OF MMAP-BIRL REWARD GRADIENTS:

Following the notations provided in the main paper, the likelihood of the visible portions of the trajectories are written as the marginal of the complete trajectory $X$ by summing out the corresponding hidden portion $Z$:

$$Pr(Y|R_\theta) = \prod_{Y \in Y} Pr(Y|R_\theta)$$

$$= \prod_{Y \in Y} \sum_{Z \in Z} Pr(Y,Z|R_\theta) = \prod_{Y \in Y} \sum_{Z \in Z} Pr(X|R_\theta).$$

Here, the parameters $\theta$ are the maximization variables and the occluded portion $Z$ of a trajectory comprises the summation variables of the marginal MAP inference. Using the above likelihood function, the MMAP-BIRL problem is more specifically formulated as:

$$R_\theta^* = \arg \max_{\theta \in \Theta} \prod_{Y \in Y} \sum_{Z \in Z} Pr(Y,Z|R_\theta) Pr(R_\theta).$$

Let $Z$ be the collection of the observations in the occluded time steps of $X$, and $Y = X/Z$. Then,

$$R_\theta^* = \arg \max_{\theta \in \Theta} \prod_{Y \in Y} \sum_{Z \in Z} Pr(o_1^t,o_2^t,o_3^t,\ldots,o_T^t|R_\theta)$$

$$\times Pr(R_\theta).$$

The learner’s observation $o_t^t$ is a noisy perception of the expert’s state and action at time step $t$, and the observations are conditionally independent of each other given the expert’s state and action. Therefore, we introduce the state-action pairs in the likelihood function above.

$$Pr(o_1^t,o_2^t,o_3^t,\ldots,o_T^t|R_\theta) = \sum_{s_1^T,a_1^T,s_2^T,a_2^T,\ldots,a_T^T} Pr(o_1^t,o_2^t,o_3^t,\ldots,o_T^t|s_1^T,a_1^T,s_2^T,a_2^T,\ldots,a_T^T,R_\theta).$$

For convenience, let $\tau$ denote the underlying trajectory of state-action pairs, $\tau = (s^T,a^T)$. Then, we may reformulate the MMAP-BIRL problem as:

$$R_\theta^* = \arg \max_{R_\theta} \prod_{Y \in Y} \sum_{Z \in Z} \sum_{\tau \in (|S||A|)^T} Pr(o_1^t,o_2^t,o_3^t,\ldots,o_T^t,\tau|R_\theta) Pr(R_\theta).$$

Now the log-posterior can be represented as:

$$L_\theta = L_{\theta}^{lh} + L_{\theta}^{pr}.$$

The log forms of the prior and the likelihood function are represented as

$$L_{\theta}^{pr} = \log Pr(R_\theta)$$

$$L_{\theta}^{lh} = \sum_{Y \in Y} \log \sum_{Z \in Z} \sum_{\tau \in (|S||A|)^T} Pr(o_1^t,o_2^t,o_3^t,\ldots,o_T^t,\tau|R_\theta).$$

Consequently, the partial differential of (1) becomes:

$$\frac{\partial L_\theta}{\partial \theta} = \frac{\partial L_{\theta}^{lh}}{\partial \theta} + \frac{\partial L_{\theta}^{pr}}{\partial \theta}.$$

1.1 DERIVATIVE OF LOG-PRIOR

If we choose the prior $Pr(\theta; \mu_\theta, \sigma^2_\theta)$ to be Gaussian, then the distribution is given as:

$$Pr(\theta; \mu_\theta, \sigma^2_\theta) = \frac{1}{\sqrt{2\pi \sigma^2_\theta}} e^{-(\theta - \mu_\theta)^2/2\sigma^2_\theta}.$$ 

where the mean $\mu_\theta$ and standard deviation $\sigma^2_\theta$ may differ between the feature weights. Then, log prior becomes:

$$L_{\theta}^{pr} = \log \left( \frac{1}{\sqrt{2\pi \sigma^2_\theta}} e^{-(\theta - \mu_\theta)^2/2\sigma^2_\theta} \right)$$

$$= \log \left( \frac{1}{\sqrt{2\pi \sigma^2_\theta}} \right) + \log \left( e^{-(\theta - \mu_\theta)^2/2\sigma^2_\theta} \right)$$

$$= -\log (\sqrt{2\pi \sigma^2_\theta}) + \log \left( \frac{-(\theta - \mu_\theta)^2}{2\sigma^2_\theta} \right).$$

Therefore, partial differential of $L_{\theta}^{pr}$ becomes:

$$\frac{\partial L_{\theta}^{pr}}{\partial \theta} = \frac{-(\theta - \mu_\theta)}{\sigma^2_\theta}.$$
1.2 DERIVATIVE OF LOG-LIKELIHOOD

As explained in the paper, the log-likelihood can be fully written as:

\[
L_{th}^h = \sum_{Y \in Y} \log \sum_{Z \in \mathcal{Z} \tau \in ([S]|A)^T} Pr(s^1) \pi(a^1|s^1; \theta) \times \prod_{t=1}^{T-1} O_i(s^t, a^t, o^t) T(s^t, a^t, s^{t+1}) \pi(a^{t+1}|s^{t+1}; \theta) \times O_i(s^T, a^T, o^T).
\]

Now, for convenience, let’s represent everything within log in (3) as:

\[
h_\theta = \sum_{Z \in \mathcal{Z} \tau \in ([S]|A)^T} Pr(s^1) \pi(a^1|s^1; \theta) \times \prod_{t=1}^{T-1} O_i(s^t, a^t, o^t) T(s^t, a^t, s^{t+1}) \pi(a^{t+1}|s^{t+1}; \theta) \times O_i(s^T, a^T, o^T).
\]

Log-likelihood now becomes:

\[
L_{th}^h = \sum_{Y \in Y} \log h_\theta = \frac{\partial L_{th}^h}{\partial \theta} = \sum_{Y \in Y} \frac{1}{h_\theta} \frac{\partial h_\theta}{\partial \theta}
\]

\[
\frac{\partial h_\theta}{\partial \theta} = \sum_{Z \in \mathcal{Z} \tau \in ([S]|A)^T} Pr(s^1) \pi(a^1|s^1; \theta) \times \prod_{t=1}^{T-1} O_i(s^t, a^t, o^t) T(s^t, a^t, s^{t+1}) \frac{\partial}{\partial \theta} \left( \prod_{t=1}^{T-1} \pi(a^{t+1}|s^{t+1}; \theta) \right) \times O_i(s^T, a^T, o^T).
\]

Now let’s say for convenience \(P_\theta^a\) holds \(\prod_{t=1}^{T-1} \pi(a^{t+1}|s^{t+1}; \theta)\) term from the above equation:

\[
P_\theta^a = \prod_{t=1}^{T-1} \pi(a^{t+1}|s^{t+1}; \theta)
\]

\[
= \pi(a^2|s^2; \theta) \times \pi(a^3|s^3; \theta) \times \pi(a^4|s^4; \theta) \times \pi(a^{T-1}|s^{T-1}; \theta)
\]

\[
\frac{\partial P_\theta^a}{\partial \theta} = \left( \frac{\partial \pi(a^3|s^3; \theta)}{\partial \theta} \times \pi(a^4|s^4; \theta) \times \pi(a^{T-1}|s^{T-1}; \theta) \right) + \left( \pi(a^2|s^2; \theta) \times \frac{\partial \pi(a^4|s^4; \theta)}{\partial \theta} \times \pi(a^{T-1}|s^{T-1}; \theta) \right) + \left( \pi(a^2|s^2; \theta) \times \pi(a^3|s^3; \theta) \times \frac{\partial \pi(a^{T-1}|s^{T-1}; \theta)}{\partial \theta} \right) + \ldots
\]

\[
= \left( \sum_{t=1}^{T-1} \frac{\partial \pi(a^{t+1}|s^{t+1}; \theta)}{\partial \theta} \right) \prod_{t \neq t}^{T-1} \pi(a^t|s^t; \theta)
\]

Partial derivative of the policy \(\pi(a^{t+1}|s^{t+1}; \theta)\) is given as,

\[
\frac{\partial \pi(a^{t+1}|s^{t+1}; \theta)}{\partial \theta} = \pi(a^{t+1}|s^{t+1}; \theta) \left( \frac{\partial Q^*(s^{t+1}, a^{t+1}; \theta)}{\partial \theta} - \sum_{a' \in A} \pi(a'|s^{t+1}; \theta) \frac{\partial Q^*(s^{t+1}, a'; \theta)}{\partial \theta} \right)
\]