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

Prasanth Sengadu Suresh[1] **Prashant Doshi**[1]

[1]THINC 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:

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

Here, the parameters $\boldsymbol{\theta}$ 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:

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

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

$$R^*_{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \prod_{Y\in\mathcal{Y}}\sum_{Z\in\mathcal{Z}} Pr(o_l^1, o_l^2, o_l^3, \ldots, o_l^{\mathcal{T}}|R_{\boldsymbol{\theta}}) \times Pr(R_{\boldsymbol{\theta}}).$$

The learner's observation o_l^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.

$$Pr(o_l^1, o_l^2, o_l^3, \ldots, o_l^{\mathcal{T}}|R_{\boldsymbol{\theta}}) = \sum_{s^1,a^1,s^2,a^2,\ldots,s^{\mathcal{T}},a^{\mathcal{T}}} Pr(o_l^1, o_l^2, o_l^3, \ldots, o_l^{\mathcal{T}}, s^1, a^1, s^2, a^2, \ldots, s^{\mathcal{T}}, a^{\mathcal{T}}|R_{\boldsymbol{\theta}}).$$

For convenience, let τ denote the underlying trajectory of state-action pairs, $\tau = (s^1, a^1, s^2, a^2 \ldots, s^{\mathcal{T}}, a^{\mathcal{T}})$. Then, we may reformulate the MMAP-BIRL problem as:

$$R^*_{\boldsymbol{\theta}} = \arg\max_{R_{\boldsymbol{\theta}}} \prod_{Y\in\mathcal{Y}}\sum_{Z\in\mathcal{Z}}\sum_{\tau\in(|S||A|)^{\mathcal{T}}} Pr(o_l^1, o_l^2, o_l^3, \ldots, o_l^{\mathcal{T}}, \tau|R_{\boldsymbol{\theta}})\; Pr(R_{\boldsymbol{\theta}}).$$

Now the log-posterior can be represented as:

$$L_{\boldsymbol{\theta}} = L^{lh}_{\boldsymbol{\theta}} + L^{pr}_{\boldsymbol{\theta}}. \quad (1)$$

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

$$L^{pr}_{\boldsymbol{\theta}} = \log Pr(R_{\boldsymbol{\theta}}) \text{ and } L^{lh}_{\boldsymbol{\theta}} = \sum_{Y\in\mathcal{Y}} \log \sum_{Z\in\mathcal{Z}}\sum_{\tau\in(|S||A|)^{\mathcal{T}}} Pr(o_l^1, o_l^2, o_l^3, \ldots, o_l^{\mathcal{T}}, \tau|R_{\boldsymbol{\theta}}).$$

Consequently, the partial differential of (1) becomes:

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

1.1 DERIVATIVE OF LOG-PRIOR

If we choose the prior $Pr(\boldsymbol{\theta};\mu_{\boldsymbol{\theta}},\sigma_{\boldsymbol{\theta}})$ to be Gaussian, then the distribution is given as:

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

where the mean $\mu_{\boldsymbol{\theta}}$ and standard deviation $\sigma_{\boldsymbol{\theta}}$ may differ between the feature weights. Then, log prior becomes:

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

Therefore, partial differential of $L^{pr}_{\boldsymbol{\theta}}$ becomes:

$$\frac{\partial L^{pr}_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}} = \left(\frac{-(\boldsymbol{\theta}-\mu_{\boldsymbol{\theta}})}{\sigma^2_{\boldsymbol{\theta}}}\right). \quad (2)$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

1.2 DERIVATIVE OF LOG-LIKELIHOOD

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

$$L^{lh}_{\boldsymbol{\theta}} = \sum_{Y\in\mathcal{Y}} \log \sum_{Z\in\mathcal{Z}} \sum_{\tau\in(|S||A|)^{\mathcal{T}}} Pr(s^1)\,\pi(a^1|s^1;\boldsymbol{\theta}) \left(\prod_{t=1}^{\mathcal{T}-1} O_l(s^t,a^t,o^t_l)\,T(s^t,a^t,s^{t+1})\,\pi(a^{t+1}|s^{t+1};\boldsymbol{\theta})\right) \times\ O_l(s^{\mathcal{T}},a^{\mathcal{T}},o^{\mathcal{T}}_l). \tag{3}$$

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

$$h_\theta = \sum_{Z\in\mathcal{Z}} \sum_{\tau\in(|S||A|)^{\mathcal{T}}} Pr(s^1)\,\pi(a^1|s^1;\boldsymbol{\theta})\times \left(\prod_{t=1}^{\mathcal{T}-1} O_l(s^t,a^t,o^t_l)\,T(s^t,a^t,s^{t+1})\,\pi(a^{t+1}|s^{t+1};\boldsymbol{\theta})\right) \times O_l(s^{\mathcal{T}},a^{\mathcal{T}},o^{\mathcal{T}}_l). \tag{4}$$

Log-likelihood now becomes:

$$L^{lh}_{\boldsymbol{\theta}} = \sum_{Y\in\mathcal{Y}} \log h_\theta \implies \frac{\partial L^{lh}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} = \sum_{Y\in\mathcal{Y}} \frac{1}{h_\theta}\frac{\partial h_\theta}{\partial \theta}.$$

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

Now let's say for convenience P^π_θ holds $\prod_{t=1}^{\mathcal{T}-1}\pi(a^{t+1}|s^{t+1};\boldsymbol{\theta})$ term from the above equation:

$$\begin{aligned}
P^\pi_\theta &= \prod_{t=1}^{\mathcal{T}-1}\pi(a^{t+1}|s^{t+1};\boldsymbol{\theta})\\
&= \pi(a^2|s^2;\boldsymbol{\theta})\times\pi(a^3|s^3;\boldsymbol{\theta})\times\pi(a^4|s^4;\boldsymbol{\theta})...\pi(a^{\mathcal{T}-1}|s^{\mathcal{T}-1};\boldsymbol{\theta})\\
\frac{\partial P^\pi_\theta}{\partial\theta} &= \Big(\pi(a^3|s^3;\boldsymbol{\theta})\times\pi(a^4|s^4;\boldsymbol{\theta})...\pi(a^{\mathcal{T}-1}|s^{\mathcal{T}-1};\boldsymbol{\theta})\Big)\frac{\partial\pi(a^2|s^2;\boldsymbol{\theta})}{\partial\theta}+\\
&\Big(\pi(a^2|s^2;\boldsymbol{\theta})\times\pi(a^4|s^4;\boldsymbol{\theta})...\pi(a^{\mathcal{T}-1}|s^{\mathcal{T}-1};\boldsymbol{\theta})\Big)\frac{\partial\pi(a^3|s^3;\boldsymbol{\theta})}{\partial\theta}+\\
&\Big(\pi(a^2|s^2;\boldsymbol{\theta})\times\pi(a^3|s^3;\boldsymbol{\theta})...\pi(a^{\mathcal{T}-1}|s^{\mathcal{T}-1};\boldsymbol{\theta})\Big)\frac{\partial\pi(a^4|s^4;\boldsymbol{\theta})}{\partial\theta}+....\\
&\Big(\pi(a^2|s^2;\boldsymbol{\theta})\times\pi(a^3|s^3;\boldsymbol{\theta})...\pi(a^{\mathcal{T}-2}|s^{\mathcal{T}-2};\boldsymbol{\theta})\Big)\frac{\partial\pi(a^{\mathcal{T}-1}|s^{\mathcal{T}-1};\boldsymbol{\theta})}{\partial\theta}\\
&= \left(\sum_{t=1}^{\mathcal{T}-1}\frac{\partial\pi(a^{t+1}|s^{t+1};\boldsymbol{\theta})}{\partial\theta}\prod_{k\neq t}^{\mathcal{T}-1}\pi(a^k|s^k;\boldsymbol{\theta})\right)
\end{aligned} \tag{5}$$

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

$$\frac{\partial\pi(a^{t+1}|s^{t+1};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \pi(a^{t+1}|s^{t+1};\boldsymbol{\theta})\Big(\frac{\beta\,\partial Q^*(s^{t+1},a^{t+1};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \sum_{a'\in A}\pi(a'|s^{t+1};\boldsymbol{\theta})\frac{\beta\,\partial Q^*(s^{t+1},a';\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big)$$

where the partial derviative of the Q-function can be obtained as:

$$\frac{\partial Q^*(s^{t+1},a^{t+1};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \frac{\partial R_\theta(s^{t+1},a^{t+1})}{\partial\boldsymbol{\theta}} + \gamma\sum_{s'\in S}T(s^{t+1},a^{t+1},s')\sum_{a'\in A}\pi(a'|s^{t+1};\boldsymbol{\theta})\frac{\partial Q^*(s',a';\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}).$$