Supplementary Material: MAP Inference for Bayesian Inverse Reinforcement Learning

Jaedeug Choi and Kee-Eung Kim
Department of Computer Science
Korea Advanced Institute of Science and Technology
Daejeon 305-701, Korea
jdchoi@ai.kaist.ac.kr, kekim@cs.kaist.ac.kr

Corollary 1 *Given an MDP\R $\langle S, A, T, \gamma, \alpha \rangle$, policy π is optimal if and only if reward function $\boldsymbol{R}$ satisfies*

$$\left[\boldsymbol{I} - (\boldsymbol{I}^A - \gamma \boldsymbol{T})(\boldsymbol{I} - \gamma \boldsymbol{T}^\pi)^{-1} \boldsymbol{E}^\pi\right] \boldsymbol{R} \leq \boldsymbol{0}, \tag{1}$$

where $\boldsymbol{E}^\pi$ is an $|S| \times |S||A|$ matrix with the $(s, (s', a'))$ element being 1 if $s = s'$ and $\pi(s') = a'$, and $\boldsymbol{I}^A$ is an $|S||A| \times |S|$ matrix constructed by stacking the $|S| \times |S|$ identity matrix $|A|$ times.

Proof

$$\begin{aligned}
&\text{Policy } \pi \text{ is optimal} \\
&\Leftrightarrow \boldsymbol{Q}_a^\pi(\boldsymbol{R}) \leq \boldsymbol{V}^\pi(\boldsymbol{R}) \\
&\Leftrightarrow \boldsymbol{R}^a + \gamma \boldsymbol{T}^a \boldsymbol{V}^\pi(\boldsymbol{R}) \leq \boldsymbol{R}^\pi + \gamma \boldsymbol{T}^\pi \boldsymbol{V}^\pi(\boldsymbol{R}) \\
&\Leftrightarrow \boldsymbol{R}^a + \gamma \boldsymbol{T}^a (\boldsymbol{I} - \gamma \boldsymbol{T}^\pi)^{-1} \boldsymbol{R}^\pi \leq \boldsymbol{R}^\pi + \gamma \boldsymbol{T}^\pi (\boldsymbol{I} - \gamma \boldsymbol{T}^\pi)^{-1} \boldsymbol{R}^\pi \\
&\Leftrightarrow \boldsymbol{R}^a - (\boldsymbol{I} - \gamma \boldsymbol{T}^a)(\boldsymbol{I} - \gamma \boldsymbol{T}^\pi)^{-1} \boldsymbol{R}^\pi \leq \boldsymbol{R}^\pi - (\boldsymbol{I} - \gamma \boldsymbol{T}^\pi)(\boldsymbol{I} - \gamma \boldsymbol{T}^\pi)^{-1} \boldsymbol{R}^\pi \\
&\Leftrightarrow \boldsymbol{R}^a - (\boldsymbol{I} - \gamma \boldsymbol{T}^a)(\boldsymbol{I} - \gamma \boldsymbol{T}^\pi)^{-1} \boldsymbol{E}^\pi \boldsymbol{R} \leq \boldsymbol{0}
\end{aligned} \tag{2}$$

The third equivalence holds by $\boldsymbol{V}^\pi(\boldsymbol{R}) = (\boldsymbol{I} - \gamma \boldsymbol{T}^\pi)^{-1} \boldsymbol{R}^\pi$. The fifth equivalence holds because the right-hand side is $\boldsymbol{0}$ and $\boldsymbol{R}^\pi = \boldsymbol{E}^\pi \boldsymbol{R}$. Stacking up Equation (2) for all $a \in A$, we obtain Equation (1). ■

Theorem 1 *IRL algorithms listed in Table 1 are equivalent to computing the MAP estimates with the prior and the likelihood using $f(\mathcal{X}; \boldsymbol{R})$ defined as follows:*

- $f_V(\mathcal{X}; \boldsymbol{R}) = \hat{V}^E(\boldsymbol{R}) - V^*(\boldsymbol{R})$
- $f_J(\mathcal{X}; \boldsymbol{R}) = -\sum_{s,a} \hat{\mu}_E(s) \left(J(s, a; \boldsymbol{R}) - \hat{\pi}_E(s, a)\right)^2$
- $f_G(\mathcal{X}; \boldsymbol{R}) = \min_i \left[V_i^{\pi^*(\boldsymbol{R})} - \hat{V}_i^E\right]$
- $f_E(\mathcal{X}; \boldsymbol{R}) = \log \mathcal{P}_{\textit{MaxEnt}}(\mathcal{X} | \boldsymbol{T}, \boldsymbol{R})$

where $\pi^*(\boldsymbol{R})$ is an optimal policy induced by the reward function $\boldsymbol{R}$, $J(s, a; \boldsymbol{R})$ is a smooth mapping from reward function $\boldsymbol{R}$ to a greedy policy such as the soft-max function, and $\mathcal{P}_{\textit{MaxEnt}}$ is the distribution on the behaviour data (trajectory or path) satisfying the principle of maximum entropy.

We prove Theorem 1 by the following lemmas.

Lemma 1 *The reward function sought by Ng and Russell's IRL algorithm from sampled trajectories [2] is equivalent to the MAP estimate with the uniform prior and the likelihood using $f_V(\mathcal{X}; \boldsymbol{R}) = \hat{V}^E(\boldsymbol{R}) - V^*(\boldsymbol{R})$.*

Table 1: IRL algorithms and their equivalent $f(\mathcal{X}; \boldsymbol{R})$ and prior for the Bayesian formulation. $q \in \{1, 2\}$ is for representing L_1 or L_2 slack penalties.

Previous algorithm	$f(\mathcal{X}; \boldsymbol{R})$	Prior
Ng and Russell's IRL from sampled trajectories [2]	f_V	Uniform
MMP without the loss function [3]	$(f_V)^q$	Gaussian
MWAL [4]	f_G	Uniform
Policy matching [1]	f_J	Uniform
MaxEnt [5]	f_E	Uniform

Proof This IRL algorithm seeks the reward function defined by

$$\boldsymbol{R}_{\text{N\&R}} = \operatorname*{argmax}_{\boldsymbol{R}} \left[\hat{V}^E(\boldsymbol{R}) - V^*(\boldsymbol{R}) \right].$$

The MAP estimate with the uniform prior and the likelihood using f_V is computed as

$$\begin{aligned}
\boldsymbol{R}_{\text{MAP}} &= \operatorname*{argmax}_{\boldsymbol{R}} P(\boldsymbol{R}|\mathcal{X}) = \operatorname*{argmax}_{\boldsymbol{R}} \log P(\boldsymbol{R}|\mathcal{X}) \\
&= \operatorname*{argmax}_{\boldsymbol{R}} \left[\log P(\mathcal{X}|\boldsymbol{R}) + \log P(\boldsymbol{R}) \right] = \operatorname*{argmax}_{\boldsymbol{R}} f_V(\mathcal{X}; \boldsymbol{R}) \\
&= \operatorname*{argmax}_{\boldsymbol{R}} \left[\hat{V}^E(\boldsymbol{R}) - V^*(\boldsymbol{R}) \right].
\end{aligned}$$

The MAP estimate is thus equivalent to $\boldsymbol{R}_{\text{N\&R}}$. ■

Lemma 2 *The reward function sought by the MMP algorithm [3] without the loss function is equivalent to the MAP estimate with a Gaussian prior and the likelihood using $(f_V)^q$ where $q \in \{1, 2\}$.*

Proof Without the loss function, the MMP algorithm seeks the reward function defined by

$$\boldsymbol{R}_{\text{MMP}} = \operatorname*{argmin}_{\boldsymbol{R}} \left[\left(V^*(\boldsymbol{R}) - \hat{V}^E(\boldsymbol{R}) \right)^q + \frac{\lambda}{2} \parallel \boldsymbol{R} \parallel_2^2 \right]$$

where $q \in \{1, 2\}$ denotes L_1 or L_2 slack penalties. The MAP estimate with a Gaussian prior $\mathcal{N}(0, \sigma^2)$ and the likelihood using $(f_V)^q$ is computed as

$$\begin{aligned}
\boldsymbol{R}_{\text{MAP}} &= \operatorname*{argmax}_{\boldsymbol{R}} P(\boldsymbol{R}|\mathcal{X}) = \operatorname*{argmax}_{\boldsymbol{R}} \left[\log P(\mathcal{X}|\boldsymbol{R}) + \log P(\boldsymbol{R}) \right] \\
&= \operatorname*{argmax}_{\boldsymbol{R}} \left[\beta \left(f_V(\mathcal{X}; \boldsymbol{R}) \right)^q - \frac{1}{2\sigma^2} \sum_{s,a} \boldsymbol{R}(s,a)^2 \right] \\
&= \operatorname*{argmax}_{\boldsymbol{R}} \left[\left(f_V(\mathcal{X}; \boldsymbol{R}) \right)^q - \frac{1}{2\beta\sigma^2} \parallel \boldsymbol{R} \parallel_2^2 \right] \\
&= \operatorname*{argmin}_{\boldsymbol{R}} \left[\left(V^*(\boldsymbol{R}) - \hat{V}^E(\boldsymbol{R}) \right)^q + \frac{1}{2\beta\sigma^2} \parallel \boldsymbol{R} \parallel_2^2 \right].
\end{aligned}$$

If we set $\lambda = 1/(\beta\sigma^2)$, the MAP estimate is equivalent to R_{MMP}. ■

Lemma 3 *When the reward function is linearly parameterized using the weight vector $\boldsymbol{w} \geq \mathbf{0}$ such that $\sum_i w_i = 1$, the policy sought by the MWAL algorithm [4] is equivalent to an optimal policy on the reward function which is the MAP estimate with the uniform prior and the likelihood using $f_G(\mathcal{X}; \boldsymbol{R}) = \min_i [V_i^{\pi^*(\boldsymbol{R})} - \hat{V}_i^E]$ where $\pi^*(\boldsymbol{R})$ is an optimal policy induced by the reward function $\boldsymbol{R}$.*

Proof The MWAL algorithm seeks the policy π_{MWAL} defined by

$$\pi_{\text{MWAL}} = \operatorname*{argmax}_{\pi} \min_i \left[V_i^{\pi} - \hat{V}_i^E \right],$$

with an implicitly computed reward function $\boldsymbol{R}_{\text{MWAL}}$ that induces π_{MWAL} as an optimal policy. Hence, we can rewrite $\pi_{\text{MWAL}} = \pi^*(\boldsymbol{R}_{\text{MWAL}})$ where

$$\boldsymbol{R}_{\text{MWAL}} = \operatorname*{argmax}_{\boldsymbol{R}} \min_i \left[V_i^{\pi^*(\boldsymbol{R})} - \hat{V}_i^E \right].$$

The MAP estimate of the reward function with the uniform prior and the likelihood using f_G is computed as

$$\boldsymbol{R}_{\text{MAP}} = \operatorname*{argmax}_{\boldsymbol{R}} P(\boldsymbol{R}|\mathcal{X}) = \operatorname*{argmax}_{\boldsymbol{R}} f_G(\mathcal{X}; \boldsymbol{R}) = \operatorname*{argmax}_{\boldsymbol{R}} \min_i \left[V_i^{\pi^*(\boldsymbol{R})} - \hat{V}_i^E \right].$$

Hence, the optimal policy induced by $\boldsymbol{R}_{\text{MAP}}$ is equivalent to π_{MWAL} since $\boldsymbol{R}_{\text{MAP}} = \boldsymbol{R}_{\text{MWAL}}$. ■

Lemma 4 *The policy sought by the policy matching algorithm [1] is equivalent to an optimal policy on the reward function which is the MAP estimate with the uniform prior and the likelihood using $f_J(\mathcal{X}; \boldsymbol{R}) = -\sum_{s,a} \hat{\mu}_E(s)(J(s,a;\boldsymbol{R}) - \hat{\pi}_E(s,a))^2$, where $J(s,a;\boldsymbol{R})$ is a smooth mapping from reward function $\boldsymbol{R}$ to a greedy policy, such as the soft-max function.*

Proof The policy matching algorithm seeks the policy $\pi_{\text{PM}} = J(\boldsymbol{R}_{\text{PM}})$ such that

$$\boldsymbol{R}_{\text{PM}} = \operatorname*{argmin}_{\boldsymbol{R}} \sum_{s,a} \hat{\mu}_E(s)(J(s,a;\boldsymbol{R}) - \hat{\pi}_E(s,a))^2.$$

The MAP estimate of the reward function with the uniform prior and the likelihood using f_J is computed as

$$\boldsymbol{R}_{\text{MAP}} = \operatorname*{argmax}_{\boldsymbol{R}} P(\boldsymbol{R}|\mathcal{X}) = \operatorname*{argmax}_{\boldsymbol{R}} f_J(\mathcal{X}; \boldsymbol{R}) = \operatorname*{argmin}_{\boldsymbol{R}} \sum_{s,a} \hat{\mu}_E(s)(J(s,a;\boldsymbol{R}) - \hat{\pi}_E(s,a))^2.$$

Hence, $\boldsymbol{R}_{\text{MAP}} = \boldsymbol{R}_{\text{PM}}$ and the optimal policy induced by $\boldsymbol{R}_{\text{MAP}}$ is equivalent to π_{PM}. ■

Lemma 5 *The reward function sought by the MaxEnt algorithm [5] is equivalent to the MAP estimate with the uniform prior and the likelihood using $f_E(\mathcal{X}; \boldsymbol{R}) = \log \mathcal{P}_{\textit{MaxEnt}}(\mathcal{X}|\boldsymbol{T}, \boldsymbol{R})$ where $\mathcal{P}_{\textit{MaxEnt}}$ is the distribution for the behavior data (trajectory or path) satisfying the principle of maximum entropy.*

Proof The MaxEnt algorithm seeks the reward function defined by

$$\boldsymbol{R}_{\text{MaxEnt}} = \operatorname*{argmax}_{\boldsymbol{R}} \log \mathcal{P}_{\text{MaxEnt}}(\mathcal{X}|\boldsymbol{T}, \boldsymbol{R})$$

where

$$\begin{aligned}\mathcal{P}_{\text{MaxEnt}}(\mathcal{X}|\boldsymbol{T}, \boldsymbol{R}) &= \prod_{m=1}^{M} \mathcal{P}_{\text{MaxEnt}}(\mathcal{X}_m|\boldsymbol{T}, \boldsymbol{R}) \\ &= \prod_{m=1}^{M} \frac{1}{Z} \exp\left(\sum_{h=1}^{H} \gamma^{h-1} \boldsymbol{R}(s_h^m, a_h^m) \right) \prod_{h=1}^{H-1} \boldsymbol{T}(s_h^m, a_h^m, s_{h+1}^m).\end{aligned}$$

The MAP estimate with the uniform prior and the likelihood using f_E is computed as

$$\boldsymbol{R}_{\text{MAP}} = \operatorname*{argmax}_{\boldsymbol{R}} P(\boldsymbol{R}|\mathcal{X}) = \operatorname*{argmax}_{\boldsymbol{R}} f_E(\mathcal{X}; \boldsymbol{R}) = \operatorname*{argmax}_{\boldsymbol{R}} \log \mathcal{P}_{\text{MaxEnt}}(\mathcal{X}|\boldsymbol{T}, \boldsymbol{R}).$$

The MAP estimate is thus equivalent to $\boldsymbol{R}_{\text{MaxEnt}}$. ■

Theorem 2 *$\boldsymbol{V}^*(\boldsymbol{R})$ and $\boldsymbol{Q}^*(\boldsymbol{R})$ are convex.*

Proof Let $C(\pi)$ be the reward optimality region w.r.t. π. $\boldsymbol{V}^*(\boldsymbol{R}) = \boldsymbol{V}^\pi(\boldsymbol{R}) = (\boldsymbol{I} - \gamma \boldsymbol{T}^\pi)^{-1}\boldsymbol{E}^\pi \boldsymbol{R}$ for any $\boldsymbol{R} \in C(\pi)$, $\boldsymbol{V}^*(\boldsymbol{R})$ is linear w.r.t. $\boldsymbol{R}$. For each and every $\boldsymbol{R}_1$, $\boldsymbol{R}_2$, and $0 \leq \mu \leq 1$,

$$\begin{aligned}\boldsymbol{V}^*(\mu \boldsymbol{R}_1 + (1-\mu)\boldsymbol{R}_2) &= \boldsymbol{H}^\pi(\mu \boldsymbol{R}_1 + (1-\mu)\boldsymbol{R}_2) = \mu \boldsymbol{H}^\pi \boldsymbol{R}_1 + (1-\mu)\boldsymbol{H}^\pi \boldsymbol{R}_2 \\ &= \mu \boldsymbol{V}^\pi(\boldsymbol{R}_1) + (1-\mu)\boldsymbol{V}^\pi(\boldsymbol{R}_2) \leq \mu \boldsymbol{V}^*(\boldsymbol{R}_1) + (1-\mu)\boldsymbol{V}^*(\boldsymbol{R}_2)\end{aligned}$$

where π is an optimal policy for $\mu \boldsymbol{R}_1 + (1-\mu)\boldsymbol{R}_2$ and $\boldsymbol{H}^\pi = (\boldsymbol{I} - \gamma \boldsymbol{T}^\pi)^{-1}\boldsymbol{E}^\pi$. Thus, $\boldsymbol{V}^*(\boldsymbol{R})$ is convex. In the same manner, we can also show that $\boldsymbol{Q}^*(\boldsymbol{R})$ is convex using the definition $\boldsymbol{Q}^\pi(\boldsymbol{R}) = \boldsymbol{R} + \gamma \boldsymbol{T}\boldsymbol{E}^\pi \boldsymbol{Q}^\pi(\boldsymbol{R})$. ■

Theorem 3 *$\boldsymbol{V}^*(\boldsymbol{R})$ and $\boldsymbol{Q}^*(\boldsymbol{R})$ are differentiable almost everywhere.*

Proof Let $C(\pi)$ be the reward optimality region w.r.t. π. Since $\boldsymbol{V}^*(\boldsymbol{R}) = \boldsymbol{V}^\pi(\boldsymbol{R}) = (\boldsymbol{I} - \gamma \boldsymbol{T}^\pi)^{-1}\boldsymbol{E}^\pi \boldsymbol{R}$ is linear for any $\boldsymbol{R} \in C(\pi)$, $\boldsymbol{V}^*(\boldsymbol{R})$ is differentiable and $\nabla_{\boldsymbol{R}} \boldsymbol{V}^*(\boldsymbol{R}) = (\boldsymbol{I} - \gamma \boldsymbol{T}^\pi)^{-1}\boldsymbol{E}^\pi$ when $\boldsymbol{R}$ is strictly inside the region. On the boundary, $\nabla_{\boldsymbol{R}} \boldsymbol{V}^\pi(\boldsymbol{R})$ is a subgradient of $\boldsymbol{V}^*(\boldsymbol{R})$ since the function is convex from Theorem 2 and thus $\nabla_{\boldsymbol{R}} \boldsymbol{V}^\pi(\boldsymbol{R})(\boldsymbol{R} - \boldsymbol{R}') \leq \boldsymbol{V}^*(\boldsymbol{R}) - \boldsymbol{V}^*(\boldsymbol{R}')$ for any $\boldsymbol{R}'$. In the same manner, we can also show that $\boldsymbol{Q}^*(\boldsymbol{R})$ is differentiable with $\nabla_{\boldsymbol{R}} \boldsymbol{Q}^*(\boldsymbol{R}) = (\boldsymbol{I} - \gamma \boldsymbol{T}\boldsymbol{E}^\pi)^{-1}$ strictly inside reward optimality regions and $\nabla_{\boldsymbol{R}} \boldsymbol{Q}^\pi(\boldsymbol{R})$ is a subgradient on the boundaries. ■

References

[1] G. Neu and C. Szepesvári. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Proceedings of UAI*, 2007.

[2] A. Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of ICML*, 2000.

[3] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *Proceedings of ICML*, 2006.

[4] U. Syed and R. E. Schapire. A game-theoretic approach to apprenticeship learning. In *Proceedings of NIPS*, 2008.

[5] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of AAAI*, 2008.