Inverse Reinforcement Learning with Simultaneous Estimation of Rewards and Dynamics - Supplementary Material

Michael Herman[⋆†] **Tobias Gindele**[⋆] **Jörg Wagner**[⋆] **Felix Schmitt**[⋆] **Wolfram Burgard**[†]

[⋆]Robert Bosch GmbH
D-70442 Stuttgart, Germany

[†]University of Freiburg
D-79110 Freiburg, Germany

Abstract

This document contains supplementary material to the paper *Inverse Reinforcement Learning with Simultaneous Estimation of Rewards and Dynamics* with more detailed derivations, additional proofs to lemmata and theorems as well as larger illustrations and plots of the evaluation task.

1 Partial Derivative of the Policy

$$
\begin{aligned}
&\frac{\partial}{\partial \theta_i} \log \pi_{\boldsymbol{\theta}}(s,a) \\
&= \frac{\partial}{\partial \theta_i} \left(Q_{\boldsymbol{\theta}}(s,a) - V_{\boldsymbol{\theta}}(s) \right) \\
&= \frac{\partial}{\partial \theta_i} \left(Q_{\boldsymbol{\theta}}(s,a) - \log \sum_{a' \in A} \exp\left(Q_{\boldsymbol{\theta}}(s,a') \right) \right) \\
&= \frac{\partial}{\partial \theta_i} Q_{\boldsymbol{\theta}}(s,a) - \frac{\partial}{\partial \theta_i} \log \sum_{a' \in A} \exp\left(Q_{\boldsymbol{\theta}}(s,a') \right) \\
&= \frac{\partial}{\partial \theta_i} Q_{\boldsymbol{\theta}}(s,a) \\
&\quad - \frac{\sum_{a' \in A} \left[\exp\left(Q_{\boldsymbol{\theta}}(s,a') \right) \frac{\partial}{\partial \theta_i} Q_{\boldsymbol{\theta}}(s,a') \right]}{\sum_{a' \in A} \exp\left(Q_{\boldsymbol{\theta}}(s,a') \right)} \\
&= \frac{\partial}{\partial \theta_i} Q_{\boldsymbol{\theta}}(s,a) - \mathbb{E}_{\pi_{\boldsymbol{\theta}}(s,a')} \left[\frac{\partial}{\partial \theta_i} Q_{\boldsymbol{\theta}}(s,a') \right].
\end{aligned}
$$

2 Partial Derivative of the Soft Q-Function with Respect to the Individual Parameter Types

The partial derivative can be further simplified if it is taken with respect to the three individual parameter types: the feature weights, the agent's dynamics parameters, and the parameters of the true environment's dynamics:

$$
\begin{aligned}
&\forall \theta_i \in \boldsymbol{\theta}_R : \frac{\partial}{\partial \theta_i} Q_{\boldsymbol{\theta}}(s,a) \\
&\quad = f_i(s,a) \\
&\qquad + \gamma \sum_{s' \in S} \left\{ P_{\boldsymbol{\theta}_{T_A}}(s'|s,a)\, \mathbb{E}_{\pi_{\boldsymbol{\theta}}(s',a')} \left[\frac{\partial}{\partial \theta_i} Q_{\boldsymbol{\theta}}(s',a') \right] \right\} \\
&\forall \theta_i \in \boldsymbol{\theta}_{T_A} : \frac{\partial}{\partial \theta_i} Q_{\boldsymbol{\theta}}(s,a) \\
&\quad = \gamma \sum_{s' \in S} \left[\left(\frac{\partial}{\partial \theta_i} P_{\boldsymbol{\theta}_{T_A}}(s'|s,a) \right) V_{\boldsymbol{\theta}}(s') \right] \\
&\qquad + \gamma \sum_{s' \in S} \left[P_{\boldsymbol{\theta}_{T_A}}(s'|s,a)\, \mathbb{E}_{\pi_{\boldsymbol{\theta}}(s',a')} \left[\frac{\partial}{\partial \theta_i} Q_{\boldsymbol{\theta}}(s',a') \right] \right] \\
&\forall \theta_i \in \boldsymbol{\theta}_T : \frac{\partial}{\partial \theta_i} Q_{\boldsymbol{\theta}}(s,a) = 0
\end{aligned}
$$

3 Proof: Soft Q-iteration is a Contraction Mapping

It has to be shown that the soft Q-iteration is a fixed point iteration with only one fixed point, since this is a requirement of our algorithm. Bloem et al. have shown in Bloem and Bambos (2014) that the soft value iteration operator is a contraction mapping. It has to be proven that the same holds for the soft Q-iteration operator $T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q})$. Therefore, we adjust their proof to be valid for the Q-iteration.

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51. Copyright 2016 by the authors.

The soft Q-iteration operator is defined as

$$T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q})(s,a) = \boldsymbol{\theta}_R^{\intercal}\boldsymbol{f}(s,a) + \gamma \sum_{s' \in S} \left[P_{\boldsymbol{\theta}_{T_A}}(s'|s,a) \operatorname*{softmax}_{a' \in A}\left(Q\left(s',a'\right)\right)\right]$$

with the function

$$\operatorname*{softmax}_{x_i \in \boldsymbol{x}}(x_i) = \log\left(\sum_{i=1}^{N} \exp\left(x_i\right)\right).$$

We will begin with deriving proofs for necessary auxiliary definitions and lemmata. In order to argue about the monotonicity of multidimensional functions, a partial order on $\mathbb{R}^{A \times B}$ is introduced. Then, a property of the softmax function is derived and afterwards the monotonicity of the operator $T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q})(s,a) : \mathbb{R}^{|S|\times|A|} \rightarrow \mathbb{R}^{|S|\times|A|}$ with respect to the introduced partial order is proven.

Definition 3.1. *For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{A\times B}$ with $A, B \in \mathbb{N}^+$, the partial order $\preceq$ is defined as $\boldsymbol{x} \preceq \boldsymbol{y} \Leftrightarrow \forall a \in A, b \in B : x_{a,b} \leq y_{a,b}$.*

Lemma 3.2. *The* softmax *function has the property that for any $\boldsymbol{x} \in \mathbb{R}^N$ and $d \in \mathbb{R}$ it holds that $\operatorname*{softmax}_{x_i \in \boldsymbol{x}}(x_i + d) = \operatorname*{softmax}_{x_i \in \boldsymbol{x}}(x_i) + d$.*

Proof. The property can be easily shown by extracting the variable d from the softmax formulation:

$$\begin{aligned}\operatorname*{softmax}_{x_i \in \boldsymbol{x}}(x_i + d) &= \log\left(\sum_{i=1}^{N} \exp\left(x_i + d\right)\right) \\ &= \log\left(\exp\left(d\right)\sum_{i=1}^{N} \exp\left(x_i\right)\right) \\ &= \log\left(\sum_{i=1}^{N} \exp\left(x_i\right)\right) + \log\left(\exp\left(d\right)\right) \\ &= \log\left(\sum_{i=1}^{N} \exp\left(x_i\right)\right) + d \\ &= \operatorname*{softmax}_{x_i \in \boldsymbol{x}}(x_i) + d\end{aligned}$$

□

Lemma 3.3. *The soft Q-iteration operator $T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q})(s,a)$ is monotone, satisfying $\forall \boldsymbol{Q}_m, \boldsymbol{Q}_n \in \mathbb{R}^{|S|\times|A|} : \boldsymbol{Q}_m \preceq \boldsymbol{Q}_n \rightarrow T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q}_m) \preceq T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q}_n)$.*

Proof. The partial derivative of the $T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q})(s,a)$ with respect to a single value $Q(s_i, a_i)$ is

$$\frac{\partial}{\partial Q(s_i,a_i)} T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q})(s,a) = \gamma P_{\boldsymbol{\theta}_{T_A}}\left(s_i|s,a\right) \frac{\exp\left(Q\left(s_i,a_i\right)\right)}{\sum_{a_j \in A}\exp\left(Q\left(s_i,a_j\right)\right)}.$$

From the definition of the MDP it follows that $\gamma \in [0,1)$ and the probability distribution $P_{\boldsymbol{\theta}_{T_A}}\left(s_i|s,a\right) \in [0,1]$. As $\forall x_i \in \mathbb{R} : \exp(x_i) \in (0,+\infty)$, all terms of the partial derivative $\frac{\partial}{\partial Q(s_i,a_i)} T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q})(s,a)$ are positive or zero, which finishes the proof that $\frac{\partial}{\partial Q(s_i,a_i)} T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q})(s,a) \geq 0$. □

Based on Lemma 3.2 and 3.3 it is possible to derive the proof that the soft Q-iteration is a contraction mapping. We transfer the proof of Bloem et al. Bloem and Bambos (2014) for the value iteration and adjust it, such that it applies for the Q-iteration.

Theorem 3.4. *The soft Q-iteration operator $T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q})(s,a)$ is a contraction mapping with only one fixed point. Therefore, it is Lipschitz continuous $||T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q}_m) - T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q}_n)||_\infty \leq L||\boldsymbol{Q}_m - \boldsymbol{Q}_n||_\infty$ for all $\boldsymbol{Q}_m, \boldsymbol{Q}_n \in \mathbb{R}^{|S|\times|A|}$ with a Lipschitz constant $L \in [0,1)$.*

Proof. Consider $\boldsymbol{Q}_m, \boldsymbol{Q}_n \in \mathbb{R}^{|S|\times|A|}$. There exists a distance d under the supremum norm, for which $\exists d \in \mathbb{R}_0^+ : ||\boldsymbol{Q}_m - \boldsymbol{Q}_n||_\infty = d$ holds and therefore

$$-d\mathbf{1} \preceq \boldsymbol{Q}_m - \boldsymbol{Q}_n \preceq d\mathbf{1}$$

with $\mathbf{1} = (1)_{k,l}$, where $1 \leq k \leq |S|, 1 \leq l \leq |A|$. Since d bounds the components of the vector difference $\boldsymbol{Q}_m - \boldsymbol{Q}_n$, it can be derived that $\boldsymbol{Q}_m \preceq \boldsymbol{Q}_n + d\mathbf{1}$ and $\boldsymbol{Q}_n \preceq \boldsymbol{Q}_m + d\mathbf{1}$. For both cases, the monotonicity condition of Lemma 3.3 is satisfied, which allows for the following inequality: $T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q}_m) \preceq T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q}_n + d\mathbf{1})$. By applying Lemma 3.2, it follows that $\forall s \in S, a \in A$

$$\begin{aligned}&T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q}_m)(s,a) \\ &\quad\leq T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q}_n + d\mathbf{1})(s,a) \\ &\quad= \boldsymbol{\theta}_R^{\intercal}\boldsymbol{f}(s,a) \\ &\qquad + \gamma \sum_{s' \in S} \left[P_{\boldsymbol{\theta}_{T_A}}(s'|s,a) \operatorname*{softmax}_{a' \in A}\left(Q\left(s',a'\right) + d\right)\right] \\ &\quad= \boldsymbol{\theta}_R^{\intercal}\boldsymbol{f}(s,a) \\ &\qquad + \gamma \sum_{s' \in S} \left[P_{\boldsymbol{\theta}_{T_A}}(s'|s,a) \operatorname*{softmax}_{a' \in A}\left(Q\left(s',a'\right)\right) + d\right] \\ &\quad= \boldsymbol{\theta}_R^{\intercal}\boldsymbol{f}(s,a) \\ &\qquad + \gamma \sum_{s' \in S} \left[P_{\boldsymbol{\theta}_{T_A}}(s'|s,a) \operatorname*{softmax}_{a' \in A}\left(Q\left(s',a'\right)\right)\right] + \gamma d \\ &\quad= T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q}_n)(s,a) + \gamma d\end{aligned}$$

In vector notation this results in $T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q}_m) \preceq T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q}_n) + \gamma d\mathbf{1}$. As from the symmetric definition of $-d\mathbf{1} \preceq \boldsymbol{Q}_m - \boldsymbol{Q}_n \preceq d\mathbf{1}$, it has been derived that $\boldsymbol{Q}_n \preceq \boldsymbol{Q}_m + d$, it consequently follows that $T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q}_n) \preceq T_{\boldsymbol{\theta}}^{soft}(\boldsymbol{Q}_m) + \gamma d\mathbf{1}$. To finish the proof, it has to be shown that the soft Q-iteration operator is

Lipschitz continuous with $L \in [0,1)$. This can be done by combining the related inequations of the operator:

$$
\begin{aligned}
-\gamma d\mathbf{1} \preceq \quad & T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q}_m) - T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q}_n) \quad \preceq \gamma d\mathbf{1}\\
& ||T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q}_m) - T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q}_n)||_\infty \leq \gamma d\\
& ||T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q}_m) - T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q}_n)||_\infty \leq \gamma ||\boldsymbol{Q}_m - \boldsymbol{Q}_n||_\infty
\end{aligned}
$$

This proves that the soft Q-iteration operator $T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q})$ is Lipschitz continuous with a Lipschitz constant $L = \gamma$ and $\gamma \in [0,1)$, resulting in a contraction mapping. As this holds for the whole input space of $\mathbb{R}^{|S|\times|A|}$, two points would always contract, so there cannot exist two fixed points. □

4 Proof: The Converged Soft Q-Function is Differentiable

Theorem 4.1. *The converged soft Q-function is differentiable with respect to* $\boldsymbol{\theta}$.

Proof. Since we provide an iterative formula for the gradient of the converged soft Q-function $\tilde{Q}(s,a)$, we need to revisit the soft Q-iteration operator $T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q}) : \mathbb{R}^{S\times A} \mapsto \mathbb{R}^{S\times A}$, element-wise defined as

$$
\begin{aligned}
& T^{soft}_{\boldsymbol{\theta}}(Q(s',a'))[s,a] = \boldsymbol{\theta}_R^\intercal \boldsymbol{f}(s,a)\\
& \qquad + \gamma \sum_{s'\in S} \Big[P_{\boldsymbol{\theta}_{T_A}}(s'|s,a) \log(\sum_{a'\in A} \exp(Q(s',a'))\Big].
\end{aligned}
$$

It is ensured that repeatedly applying $T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q})$ to an initial $\boldsymbol{Q}_0$ converges to a fixed-point $\tilde{\boldsymbol{Q}}$ given $\gamma \in [0,1)$, as the soft Q-operator converges [see Section 3]. $T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q})$ is differentiable with respect to both $\boldsymbol{Q}$ and $\boldsymbol{\theta}$ as being the composition of differentiable functions. This requires the transition model $P_{\boldsymbol{\theta}_{T_A}}$ to be differentiable with respect to $\boldsymbol{\theta}$, too. We now apply the implicit function theorem Krantz and Parks (2002) to compute the derivative $\frac{\partial}{\partial \boldsymbol{\theta}} \tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}$ given by the equation

$$
T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q})(s,a) - Q(s,a) = 0.
$$

The theorem states that if the Jacobian $\frac{\partial}{\partial \boldsymbol{Q}}[T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q}) - \boldsymbol{Q}]$ is invertible at $\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}$, the derivative $\frac{\partial}{\partial \boldsymbol{\theta}} \tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}$ exists and is given by

$$
\frac{\partial}{\partial \boldsymbol{\theta}} \tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}} = \big(\frac{\partial}{\partial \boldsymbol{Q}}[T^{soft}_{\boldsymbol{\theta}}(.) - .]\big)^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} T^{soft}_{\boldsymbol{\theta}}(.)\,(\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}).
$$

Since the partial derivative of the operator $T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q})$ has already been derived in Section 3, the Jacobian of $T^{soft}_{\boldsymbol{\theta}}(\boldsymbol{Q}) - \boldsymbol{Q}$ is

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{Q}}[T^{soft}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}) - \tilde{\boldsymbol{Q}}_{\theta}]([s,a],[s',a'])\\
& \qquad = -\delta(a'=a, s'=s) + \gamma P_{\boldsymbol{\theta}_{T_A}}(s'|s,a)\\
& \qquad \quad \cdot \frac{1}{\sum_{a''\in A} \exp(Q(s',a''))} \exp(Q(s',a'))\\
& \qquad = -\delta(a'=a, s'=s) + \gamma \underbrace{\pi(a'|s') P_{\boldsymbol{\theta}_{T_A}}(s'|a,s)}_{M_{[s,a],[s',a']}}.
\end{aligned}
$$

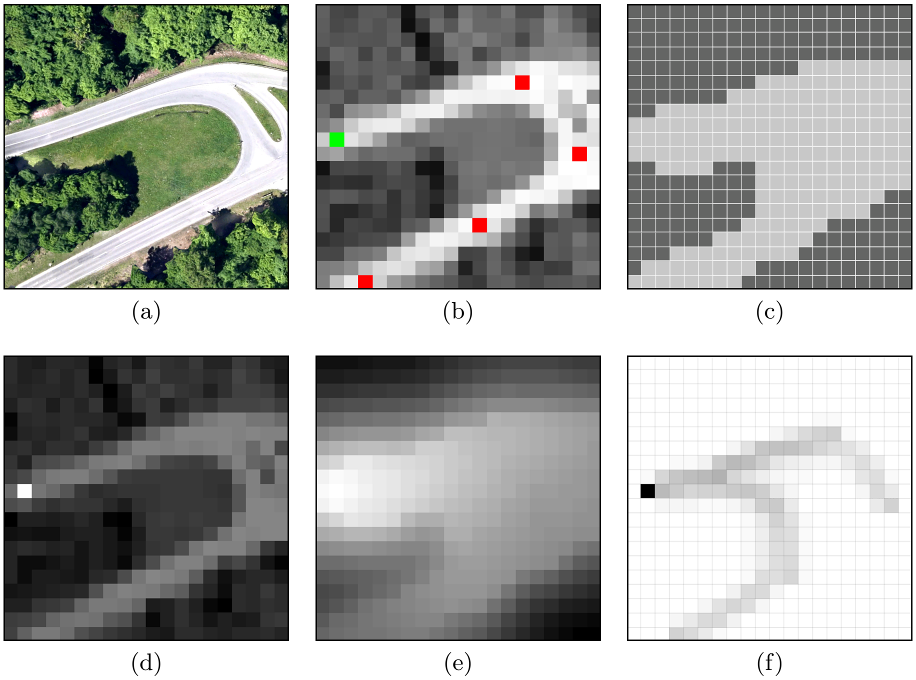
It holds $1 > \gamma \geq ||\gamma \boldsymbol{M}||_\infty$ in the $L\infty$ induced matrix-norm defined as $||A||_\infty := \max_x \frac{|Ax|_\infty}{|x|_\infty} = \max_i \sum_j |A_{i,j}|$, as

$$
\begin{aligned}
& \max_{[s,a]} \sum_{[s',s']} |\gamma M_{[s,a],[s',a']}|\\
& \qquad\qquad = \max_{[s,a]} \sum_{[s',a']} |\gamma \pi(a'|s') P_\theta(s'|a,s)| = \gamma.
\end{aligned}
$$

Hence, $(\gamma \boldsymbol{M} - I)^{-1}$ exists and is given by the *Neuman* operator-series $-\sum_{i=0}^{\infty} (\gamma \boldsymbol{M})^i$. Since the Jacobian is invertible it is proven that the partial derivative of the converged soft Q-function $\frac{\partial}{\partial \boldsymbol{\theta}} \tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}$ with respect to the parameters $\boldsymbol{\theta}$ exists. □

5 Grid World Terrain Motion Task

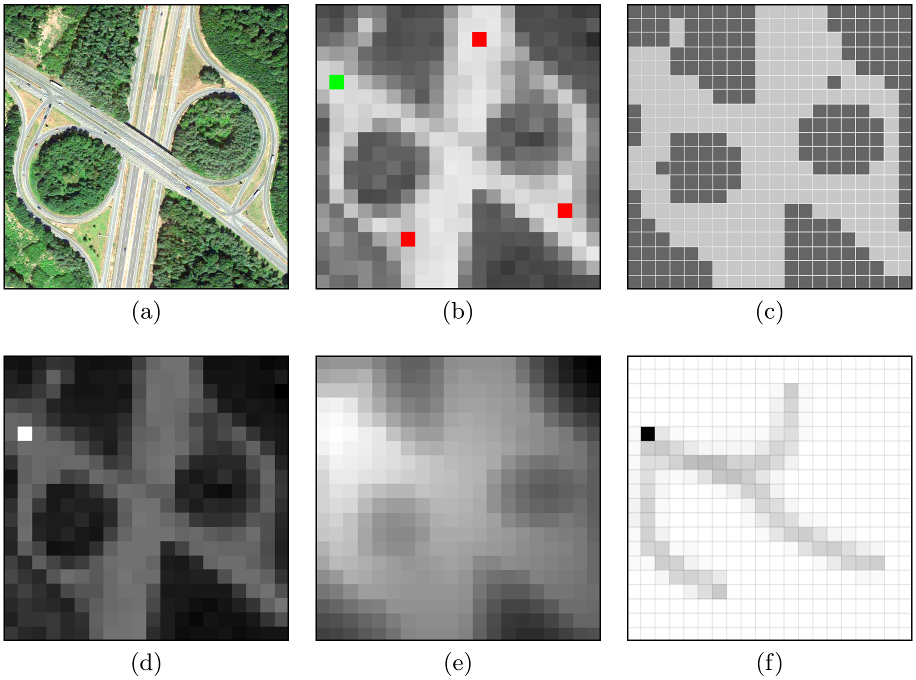
This section provides larger illustrations and results for the evaluation task. Figure 1 and 2 illustrate the environment of the training and transfer task. The results are summarized in Figure 3, where (a), (b), (c) are results on the training task and (d) presents the performance on the transfer task. We used Welch's t-test Welch (1947) to verify that the differences of the mean log likelihood of the demonstrations under the trained models in Figure 3 (a) and (d) are statistically significant ($p < 0.05$). In the training task, the performance of SERD against the other approaches is statistically significant for sample set sizes that are larger than 3, while in the transfer task statistical significance is given at least for demonstration set sizes larger than 12.



(a) (b) (c)

(d) (e) (f)

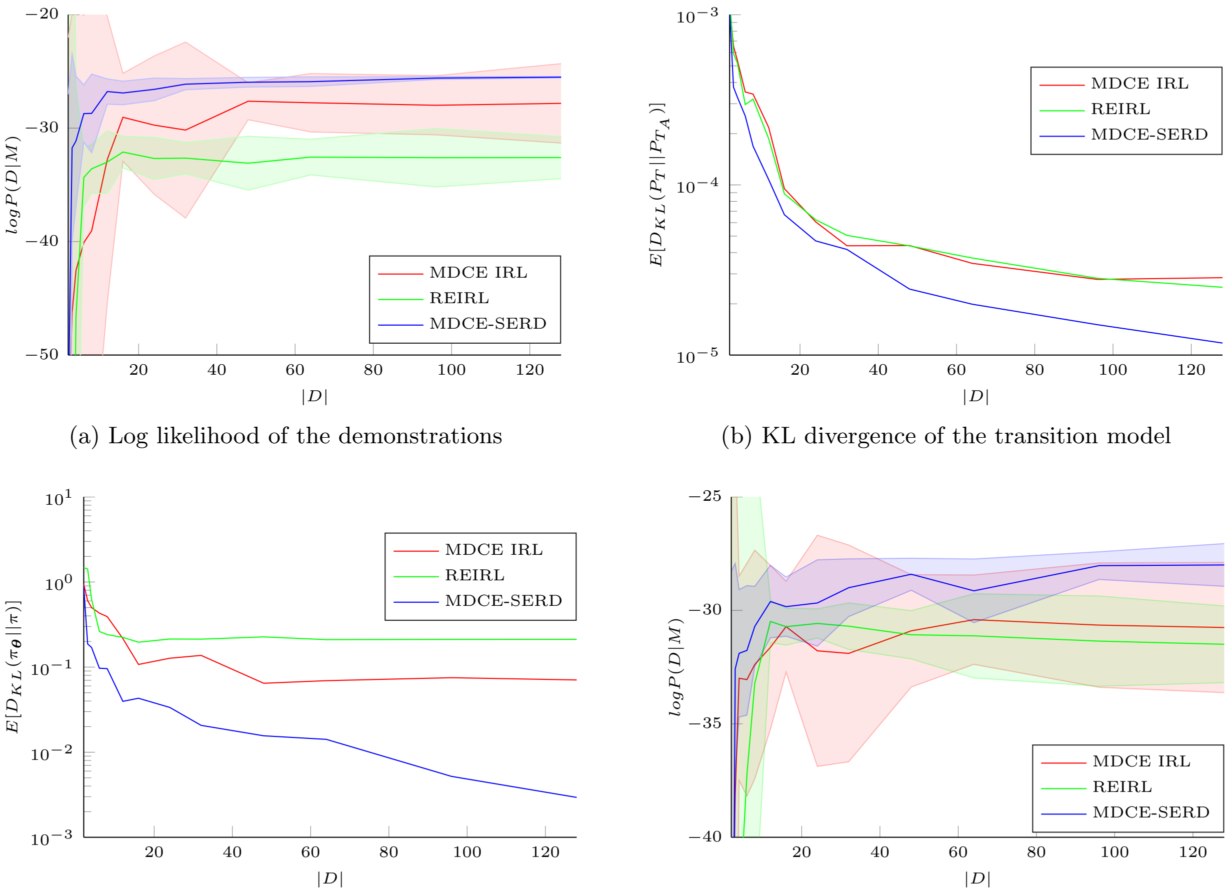
Figure 1: The training and test task. (a) Environment, Map data: Google. (b) Discretized state space. The goal state is indicated in green and start states in red. (c) Forest states are indicated in a dark-gray color and open terrain in light gray. Furthermore, plot (d) shows the reward, (e) the resulting value function, and (f) the expected state frequency.



(a) (b) (c)

(d) (e) (f)

Figure 2: The transfer task. (a) Environment, Map data: Google. (b) Discretized state space. The goal state is indicated in green and start states in red. (c) Forest states are indicated in a dark-gray color and open terrain in light gray. Furthermore, plot (d) shows the reward, (e) the resulting value function, and (f) the expected state frequency.



(a) Log likelihood of the demonstrations

(b) KL divergence of the transition model

(c) KL divergence of the policy

(d) Log likelihood of the demonstrations (transfer)

Figure 3: (a) Average log likelihood of demonstrations drawn from the true model under the estimated model. (b) Average Kullback-Leibler divergence between the estimated dynamics and the true ones. (c) Average Kullback-Leibler divergence between the trained stochastic policy and the true one. (d) Average log likelihood of demonstrations drawn from the true model under the estimated model in the transfer task environment.

References

Michael Bloem and Nicholas Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. In *53rd IEEE Conference on Decision and Control, CDC 2014, Los Angeles, CA, USA, December 15-17, 2014*, pages 4911–4916, 2014.

Steven G. Krantz and Harold R. Parks. *The Implicit Function Theorem: History, Theory, and Applications*. The Implicit Function Theorem: History, Theory, and Applications. Birkhäuser, 2002.

B L Welch. The Generalization of ‘Student’s’ Problem when Several Different Population Variances are Involved. *Biometrika*, 34(1/2):28–35, 1947.