ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

# Simultaneous estimation of rewards and dynamics from noisy expert demonstrations

Michael Herman[1,2], Tobias Gindele[1], Jörg Wagner[1],
Felix Schmitt[1], and Wolfram Burgard[2]

1- Robert Bosch GmbH - 70442 Stuttgart - Germany

2- University of Freiburg - Department of Computer Science
79110 Freiburg - Germany

**Abstract**. Inverse Reinforcement Learning (IRL) describes the problem of learning an unknown reward function of a Markov Decision Process (MDP) from demonstrations of an expert. Current approaches typically require the system dynamics to be known or additional demonstrations of state transitions to be available to solve the inverse problem accurately. If these assumptions are not satisfied, heuristics can be used to compensate the lack of a model of the system dynamics. However, heuristics can add bias to the solution. To overcome this, we present a gradient-based approach, which simultaneously estimates rewards, dynamics, and the parameterizable stochastic policy of an expert from demonstrations, while the stochastic policy is a function of optimal Q-values.

## 1 Introduction

The growing number of autonomous systems requires efficient methods to adjust the system to new environments and tasks. Learning from demonstration offers methods to parameterize a desired behavior and can be split into two subfields: Behavioral Cloning and Inverse Reinforcement Learning (IRL). Behavioral Cloning estimates a policy from demonstrations and therefore mimics the expert directly. Especially, if the environment or its dynamics change, pretrained policies can be inappropriate. Therefore, IRL [1] has been introduced, which describes the problem of recovering a reward function from demonstrations, as the reward function encodes the expert's goal. Approaches have been proposed which solve the IRL problem under various assumptions, e.g. [2, 3, 4, 5]. The cited approaches require the true system dynamics to be known. Inaccurate transition models can bias the reward estimate. Since the system dynamics are often unknown, model-free IRL algorithms have been proposed, such as in [6, 7, 8]. Typically, those approaches require access to additional observations of transitions. If these cannot be obtained, the approaches tend to suffer from wrong generalizations due to heuristics. Often, experts are unable to produce optimal demonstrations. As a consequence, IRL approaches are necessary that deal with stochastic behavior. In [9, 10, 11], stochastic policies of maximum (causal) entropy are trained under the constraint of matching feature expectations. This causes the stochastic policy to be a Boltzmann distribution over soft Q-values. However, if the expert's stochastic policy follows a different type of distribution, these approaches can be inappropriate.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Our contribution is to generalize IRL to the case of unknown dynamics and unknown stochastic policies. We propose an approach that simultaneously optimizes rewards, dynamics, and the expert's stochastic policy by maximizing the a posteriori probability of the demonstrations. Even though many transitions have never been observed, they influenced the expert's policy and can therefore to some degree be inferred from demonstrations. The expert's stochastic policy is modeled as a parametric function of optimal Q-values, which assumes that the expert is able to correctly estimate the value of different actions, but is unable to choose them appropriately. We provide a gradient-based solution and evaluate our approach on a synthetic gridworld satellite navigation task.

## 2 Fundamentals

An MDP is a tuple $M = \{S, A, P(s'|s,a), \gamma, P(s_0), R\}$, where $S$ is the state space with states $s \in S$, $A$ is the action space with actions $a \in A$, $P(s'|s,a)$ is the probability of a transition to $s'$ when action $a$ is applied in state $s$, $\gamma \in [0,1)$ is a discount factor, $P(s_0)$ is a start state probability distribution, and $R : S \times A \to \mathbb{R}$ is a reward function which assigns a real-valued reward for picking action $a$ in state $s$. Often, this reward is expressed as a linear function $R(s,a) = \boldsymbol{\theta}^T \boldsymbol{f}(s,a)$ of state- and action-dependent features $\boldsymbol{f} : S \times A \to \mathbb{R}^d$ with feature weights $\boldsymbol{\theta}$. The goal of an MDP is to find an optimal policy $\pi^*(s) \in A$, which specifies state-dependent actions $a$, such that its execution maximizes the expected, discounted, cumulated reward $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \,|s_0 = s, \pi\right]$. The optimal value function can be computed by value iteration, which repeatedly applies Eq. (1) to an arbitrary initial Q-function. After convergence, the optimal policy chooses the actions with the largest Q-value: $\pi^*(s) = \text{argmax}_{a'} Q(s,a')$.

$$Q(s,a) = R(s,a) + \gamma \sum_{s' \in S} \left[ P(s'|s,a) \max_{a'} Q(s',a') \right] \tag{1}$$

## 3 Simultaneous Estimation of Rewards, Dynamics, and Stochastic Policy (SERD-SP)

We propose an approach, called Simultaneous Estimation of Rewards, Dynamics, and Stochastic Policy (SERD-SP), to account for problems, where neither the rewards, the dynamics, nor the expert's stochastic policy $\pi(s,a) = P(a|s)$ is known. Since the expert's estimate of the transition model may differ from the real one, we introduce independent models. Additionally, we assume that there exists a parameterizable stochastic mapping $\pi = g(Q)$ from optimal Q-values to the stochastic policy of the expert. Then, the problem can be formalized as:

**Determine:**

- Expert's reward function $R(s,a)$
- Expert's estimate of the dynamics $P_A(s'|s,a)$
- Real dynamics $P(s'|s,a)$
- Stochastic policy mapping $g(Q)$

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

**Given:**

- MDP $M \setminus \{R, P(s'|s,a), P_A(s'|s,a)\}$ without rewards and dynamics
- Demonstrations $D = \{\tau_1, \tau_2, \ldots, \tau_N\}$ with trajectories $\tau = \{(s_0^\tau, a_0^\tau), (s_1^\tau, a_1^\tau), \ldots, (s_{T_\tau}^\tau, a_{T_\tau}^\tau)\}$ of an expert acting in $M$ based on a policy that depends on $R(s,a)$, $P_A(s'|s,a)$, and $g(Q)$

A set of parameters of the rewards, dynamics, and the stochastic policy is introduced, which should be estimated from the given demonstrations $D$:

| | |
|---|---|
| $\boldsymbol{\theta}_R$ | Feature weights of the reward function $R(s,a)$ |
| $\boldsymbol{\theta}_{T_A}$ | Parameters of the expert's transition model $P_{\boldsymbol{\theta}_{T_A}}$ |
| $\boldsymbol{\theta}_T$ | Parameters of the real transition model $P_{\boldsymbol{\theta}_T}$ |
| $\boldsymbol{\theta}_P$ | Parameters of the expert's stochastic policy mapping $g(Q)$ |

We propose to maximize the a posteriori probability of the demonstrations with respect to the parameters $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_R^\intercal \quad \boldsymbol{\theta}_{T_A}^\intercal \quad \boldsymbol{\theta}_T^\intercal \quad \boldsymbol{\theta}_P^\intercal\right)^\intercal$. Assuming independent trajectories, the likelihood of the demonstrations in $D$ can be expressed as

$$P(D|M,\boldsymbol{\theta}) = \prod_{\tau \in D} P(s_0^\tau) \prod_{t=0}^{T_\tau - 1} \left[\pi_{\boldsymbol{\theta}}(s_t^\tau, a_t^\tau) P_{\boldsymbol{\theta}_T}\left(s_{t+1}^\tau | s_t^\tau, a_t^\tau\right)\right]. \tag{2}$$

It should be noted that the policy $\pi_{\boldsymbol{\theta}}(s,a)$ depends on the parameters $\boldsymbol{\theta}_R$, $\boldsymbol{\theta}_{T_A}$, and $\boldsymbol{\theta}_P$. In contrast, the transition model $P_{\boldsymbol{\theta}_T}(s'|s,a)$ only depends on $\boldsymbol{\theta}_T$. Then, the maximum a posteriori estimator of the parameters can be formulated:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log P(D|M,\boldsymbol{\theta}) + \log P(\boldsymbol{\theta}). \tag{3}$$

We propose a gradient-based method to optimize the parameters according to Eq. (3) with $L_{\boldsymbol{\theta}}(D) = \log P(D|M,\boldsymbol{\theta}) + \log P(\boldsymbol{\theta})$:

$$\begin{aligned}\frac{\partial}{\partial \theta_i} L_{\boldsymbol{\theta}}(D) = \sum_{\tau \in D} \sum_{t=0}^{T_\tau - 1} \left[\frac{\partial}{\partial \theta_i} \log \pi_{\boldsymbol{\theta}}(s_t^\tau, a_t^\tau) + \frac{\partial}{\partial \theta_i} \log P_{\boldsymbol{\theta}_T}\left(s_{t+1}^\tau | s_t^\tau, a_t^\tau\right)\right] \\ + \frac{\partial}{\partial \theta_i} \log P(\boldsymbol{\theta}).\end{aligned} \tag{4}$$

Since the system dynamics and the prior are problem-dependent, the following derivations will focus on the partial derivative $\frac{\partial}{\partial \theta_i} \log \pi_{\boldsymbol{\theta}}(s_t^\tau, a_t^\tau)$. This requires the stochastic policy mapping $\pi = g(Q)$ of the expert to be specified. We will exemplarily derive the gradient for a Boltzmann policy with *temperature* $\theta_P$:

$$\pi_{\boldsymbol{\theta}}(s,a) = g(Q)(s,a) = \frac{\exp(\frac{1}{\theta_P} Q_{\boldsymbol{\theta}}(s,a))}{\sum_{a' \in A} \exp\left(\frac{1}{\theta_P} Q_{\boldsymbol{\theta}}(s,a')\right)}. \tag{5}$$

Then, the partial derivative of the log policy $\frac{\partial}{\partial \theta_i} \log \pi_{\boldsymbol{\theta}}(s_t^\tau, a_t^\tau)$ results in:

$$\frac{\partial}{\partial \theta_i} \log \pi_{\boldsymbol{\theta}}(s,a) = \begin{cases} \frac{1}{\theta_P}\left[\frac{\partial}{\partial \theta_i} Q_{\boldsymbol{\theta}}(s,a) - \mathbb{E}_{\pi_{\boldsymbol{\theta}}(s,a')}\left[\frac{\partial}{\partial \theta_i} Q_{\boldsymbol{\theta}}(s,a')\right]\right] & \text{if } \theta_i \neq \theta_P \\ \frac{1}{\theta_P{}^2}\left[\mathbb{E}_{\pi_{\boldsymbol{\theta}}(s,a')}\left[Q_{\boldsymbol{\theta}}(s,a')\right] - Q_{\boldsymbol{\theta}}(s,a)\right] & \text{if } \theta_i = \theta_P \end{cases}$$

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

The gradient of the policy depends on the gradient of the state-action value function $\frac{\partial}{\partial\theta_i}Q_{\boldsymbol{\theta}}(s,a)$. Since we assume that the expert chooses actions based on an optimal, greedy value function, the derivative of the Q-function from Eq. (1) has to be computed. This can result in a sub-derivative, as the max-function is not differentiable. Nevertheless, for the sake of simplicity, we call it Q-gradient.

$$\begin{aligned}\frac{\partial}{\partial\theta_i}Q_{\boldsymbol{\theta}}(s,a) =& \frac{\partial}{\partial\theta_i}\boldsymbol{\theta}_R^{\mathsf{T}}\boldsymbol{f}(s,a) + \gamma\sum_{s'\in S}\left[\left(\frac{\partial}{\partial\theta_i}P_{\boldsymbol{\theta}_{T_A}}\left(s'|s,a\right)\right)V_{\boldsymbol{\theta}}\left(s'\right)\right] \\ &+ \gamma\sum_{s'\in S}\left\{P_{\boldsymbol{\theta}_{T_A}}\left(s'|s,a\right)\frac{\partial}{\partial\theta_i}Q_{\boldsymbol{\theta}}(s',\pi^*_{\boldsymbol{\theta}}\left(s'\right))\right\}\end{aligned} \tag{6}$$

Eq. (6) shares similarities with the approach from Neu and Szepesvári [3]. It is a linear equation system and can be computed directly. However, since it is a fixed point equation, repeatedly applying Eq. (6) to an arbitrary Q-gradient will converge to the true one. Especially in large state and action spaces, this Q-gradient iteration can require less computations than directly solving the linear equation system. Algorithm 1 summarizes the proposed algorithm.

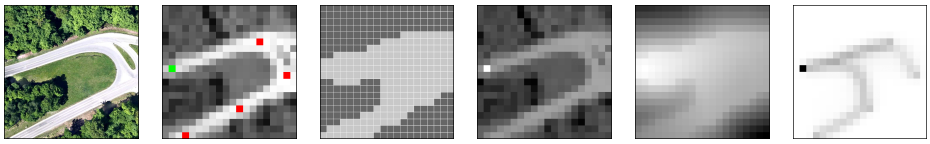
**Algorithm 1** SERD algorithm

**Require:** MDP $M\setminus\{R, P_T, P_{T_A}, g(Q)\}$, Demonstrations $D$, initial $\boldsymbol{\theta}_0$, step size $\alpha:\mathbb{N}_+\rightarrow\mathbb{R}_+$, $t\leftarrow 0$
**while** not sufficiently converged **do**
  $\boldsymbol{Q}_{\boldsymbol{\theta}} \leftarrow$ QIteration($M$, $\boldsymbol{\theta}_t$) ▷ Eq. (1)
  $\boldsymbol{\pi}_{\boldsymbol{\theta}} \leftarrow$ DerivePolicy($M$, $\boldsymbol{Q}_{\boldsymbol{\theta}}$) ▷ Eq. (5)
  $d\boldsymbol{Q}_{\boldsymbol{\theta}} \leftarrow$ ComputeQGradient($M$, $\boldsymbol{Q}_{\boldsymbol{\theta}}$, $\boldsymbol{\pi}_{\boldsymbol{\theta}}$, $\boldsymbol{\theta}_t$) ▷ Eq. (6)
  $d\boldsymbol{L}_{\boldsymbol{\theta}}\left(D\right) \leftarrow$ ComputeGradient($M$, $D$, $d\boldsymbol{Q}_{\boldsymbol{\theta}}$) ▷ Eq. (4)
  $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \alpha(t)d\boldsymbol{L}_{\boldsymbol{\theta}}\left(D\right)$
  $t \leftarrow t+1$
**end while**

## 4 Evaluation

We evaluate the proposed approach in a satellite gridworld navigation task, which is illustrated in Fig. 1. The motion dynamics are stochastic and differ in the forest and on open terrain. The action space allows the agent to choose from five different actions: moving in one of four directions (north, east, south, or west) or remaining in the state, respectively. Possible successor states are the four neighbouring ones or the current one. On the open terrain (depicted in light gray in Fig. 1 (c)), the agent has a probability of 0.8 to successfully execute the desired motion and 0.1 to fall either to the right or to the left. In the forest (depicted in dark gray in Fig. 1 (c)), successful motions only occur with a probability of 0.3. The remaining successor states have a probability of 0.175. Staying in a state is always successful in both forest and open terrain. Due to this definition of the motion dynamics, the agent has to trade off between short cuts through the forest, which are less likely to be successful, or longer paths on

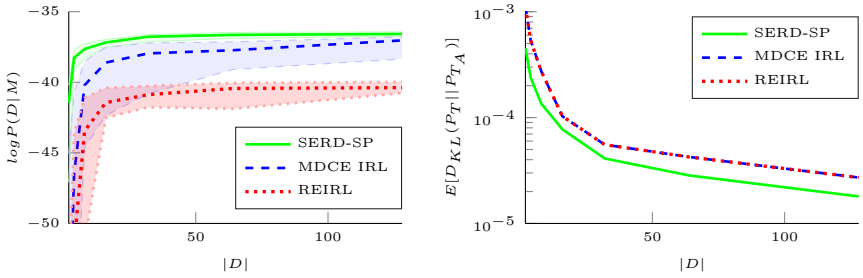
ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.



(a) (b) (c) (d) (e) (f)

Fig. 1: (a) Environment, Map data: Google. (b) Discretized state space (Goal: green. Initial states: red.). (c) Forest states are indicated in dark-gray and open terrain in light gray. (d) Reward (e) Value function (f) Expected state frequency.

open terrain. The reward is a function of two features, which are weighted by $\boldsymbol{\theta}_R = (6, 6)^\intercal$. The first feature encodes the normalized gray scale value $[0, 1]$ of the image, while the second one is a goal indicator $\{0, 1\}$. The discount is 0.99 and the *temperature* of the Boltzmann policy was set to $\theta_P = 2$. We compute the optimal Q-function and sample trajectories from the resulting stochastic policy to obtain expert demonstrations. We assume that the expert has knowledge about the true transition model. Therefore, the parameters of the transition model $\boldsymbol{\theta}_{T_A}$ and $\boldsymbol{\theta}_T$ are identical. The system dynamics are modeled as energies of Boltzmann distributions. Since there exist 4 motion actions in each, forest and open terrain, as well as one staying action, 9 models are trained with 5 possible outcomes, resulting in 45 parameters. An m-estimator with a uniform prior is used to estimate the dynamics from demonstrations before applying SERD-SP or alternative IRL approaches. The feature weights are initialized randomly ($\forall i : \theta_i \in [-10, 10]$). We use Gaussian priors for the feature weights and the policy parameter. The prior of the dynamics is favoring high entropies. We optimize all parameters for various sizes of demonstration sets with SERD-SP and compare it to the result of Maximum Discounted Causal Entropy IRL [11] (MDCE IRL), and Relative Entropy IRL [6] (REIRL). The additional samples, which are needed by REIRL, are sampled from the m-estimated transition model. Fig. 2 summarizes the results. The median log likelihood of demonstrations from the true model on the learned ones in Fig. 2 (a) shows that SERD-SP outperforms




(a) Log likelihood of the demonstrations (b) KL divergence of the transition model

Fig. 2: (a) Median with quartiles of the 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.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

the other algorithms, while being sample efficient. This result is understandable, as the comparative approaches model different types of stochastic policies. In addition, Fig. 2 (b) illustrates that SERD-SP is further optimizing the initially m-estimated dynamics, which results in more accurate models.

## 5 Conclusion

In this paper, we presented a gradient-based solution for a simultaneous estimation of rewards, dynamics, as well as the expert's stochastic policy. We assume that the expert is able to compute an optimal Q-function, but executes suboptimal actions. This stochasticity is modeled by a parameterizable function of optimal Q-values. The evaluation shows improved performance against traditional IRL methods with more accurate policies and dynamics. Future work could elaborate on different types of stochastic policies and on the case that the agent's estimate of the dynamics differs from the true one.

## References

[1] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[2] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, New York, NY, USA, 2004. ACM.

[3] Gergely Neu and Csaba Szepesvári. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *UAI 2007, Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, Vancouver, BC, Canada, July 19-22, 2007*, pages 295–302, 2007.

[4] Deepak Ramachandran and Eyal Amir. Bayesian Inverse Reinforcement Learning. *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, 51:2586–2591, 2007.

[5] Constantin A. Rothkopf and Christos Dimitrakakis. Preference elicitation and inverse reinforcement learning. In *ECML/PKDD (3)*, volume 6913 of *Lecture Notes in Computer Science*, pages 34–48. Springer, 2011.

[6] Abdeslam Boularias, Jens Kober, and Jan Peters. Relative entropy inverse reinforcement learning. In *Proceedings of Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, 2011.

[7] Edouard Klein, Matthieu Geist, Bilal Piot, and Olivier Pietquin. Inverse Reinforcement Learning through Structured Classification. In *Advances in Neural Information Processing Systems (NIPS 2012)*, Lake Tahoe (NV, USA), December 2012.

[8] Edouard Klein, Bilal Piot, Matthieu Geist, and Olivier Pietquin. A cascaded supervised learning approach to inverse reinforcement learning. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2013)*, Prague (Czech Republic), September 2013.

[9] Brian D. Ziebart, Andrew Maas, J. Andrew (Drew) Bagnell, and Anind Dey. Maximum entropy inverse reinforcement learning. In *Proceeding of AAAI 2008*, July 2008.

[10] Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. Modeling interaction via the principle of maximum causal entropy. In *Proc. of the International Conference on Machine Learning*, pages 1255–1262, 2010.

[11] 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.