Generalized Inverse Reinforcement Learning with Linearly Solvable MDP

Masahiro Kohjima(✉), Tatsushi Matsubayashi, and Hiroshi Sawada

NTT Service Evolution Laboratories, NTT Corporation, Japan
{kohjima.masahiro,matsubayashi.tatsushi,sawada.hiroshi}@lab.ntt.co.jp

Abstract. In this paper, we consider a generalized variant of *inverse reinforcement learning* (IRL) that estimates both a cost (negative reward) function and a transition probability from observed optimal behavior. In theoretical studies of standard IRL, which estimates only the cost function, it is well known that IRL involves a non-identifiable problem, i.e., the cost function cannot be determined uniquely. This problem has been solved by using a new class of Markov decision process (MDP) called a linearly solvable MDP (LMDP). In this paper, we investigate whether a non-identifiable problem occurs in the generalized variant of IRL (gIRL) using the framework of LMDP and construct a new gIRL method. The contributions of this study are summarized as follows: (i) We point out that gIRL with LMDP suffers from a non-identifiable problem. (ii) We propose a Bayesian method to escape the non-identifiable problem. (iii) We validate the proposed method by performing an experiment on synthetic data and real car probe data.

Keywords: Inverse reinforcement learning, Linearly solvable MDP, Bayesian method

1 Introduction

Inverse reinforcement learning (IRL) is a method that estimates the cost (negative reward) function of a certain class of Markov decision process (MDP) from an agent's optimal behavior. Since designing a truly effective cost function is regarded as a difficult problem in various applications of *reinforcement learning* (RL) including robot control tasks, IRL attracted the attention of robotics researchers from an early stage [1]. Its application area is now spreading and the effectiveness of IRL has been reported for taxi driver destination prediction [2], preferred route estimation after a natural disaster [3], and natural language processing [4]. These studies show that IRL can estimate the cost functions of entities, such as people and animals, whose internal structure is unobservable and whose preferences remain vague.

In this paper, we consider a generalized variant of IRL that simultaneously estimates both the cost function and transition probability. Since this problem is a generalization of existing IRL methods that estimate only the cost function, we call it *generalized IRL* (gIRL). Figure 1 shows the input and output of RL,

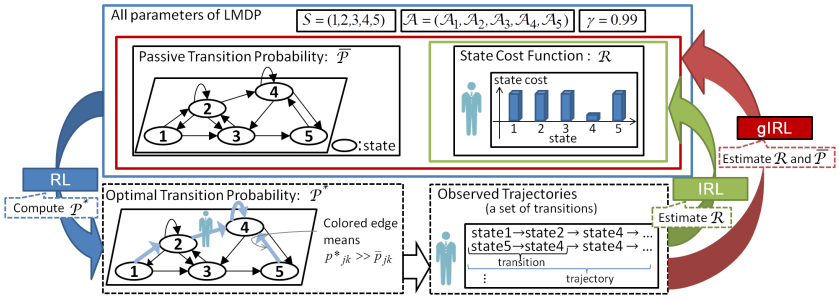


Fig. 1. Input and output of RL, IRL and gIRL with linearly solvable MDP.

IRL and gIRL. Specifically, as shown in the figure, we tackle the gIRL using the framework of the linearly solvable MDP (LMDP) [5].

LMDP has been proposed as a new class of MDP where a forward problem (RL) is more easily solved than with standard MDP [5]. Dvijotham showed that IRL with the LMDP has a unique solution [6], i.e., the cost function generating agent behavior is uniquely identified. It is regarded as important result in IRL. The first IRL paper [7] proved the existence of a non-identifiable problem with standard MDP, and therefore the cost function is not unique and that a cost function with entirely zero values is always one of the solutions. Until Dvijotham's paper was published it had remained an open issue as to whether it was possible to avoid a non-identifiable problem. [1] However, a gIRL with the LMDP has not yet been studied. Since the number of transition probability parameters of the LMDP is smaller than that of a standard MDP, the use of the LMDP is suitable for gIRL.

The study most closely related to ours is the work reported by Makino and Takeuchi [8], which considers gIRL on the partially observable MDP for constructing efficient apprenticeship learning methods. It is experimentally confirmed that gIRL formulation contributes to the realization of a more effective policy [8]. However, the theoretical aspect of gIRL remained unknown and there is no gIRL method for the LMDP.

In this paper, we provide a theoretical analysis and a new method for gIRL. Beginning with an investigation as to whether a non-identifiable problem occurs in gIRL with LMDP, we establish a new formulation of gIRL and new gIRL methods. We apply the proposed method with both synthetic data and real car probe data collected in Yokohama City, Japan. The contributions of this paper can be summarized as follows:

– We point out that generalized IRL using the framework of LMDP involves a non-identifiable problem; the cost function and transition probability cannot be uniquely estimated. This is because we cannot distinguish between

[1] Although Ziebart et al. [2] also solve the non-identifiable problem by using the maximum entropy principle, Dvijotham and Todorov show that Ziebart's formulation is equivalent to the special case of an inverse problem of LMDP [6].

the effect of the cost function and that of the transition probability on the observed transitions.
- To avoid the non-identifiable problem, we adopt a Bayesian approach with hyperparameters, which is also used approach for IRL with a standard MDP [8–11]. We also extend the LMDP to a multiple intention setting [11, 12] and use it to formulate generalized IRL. This enables us to apply the proposed method to many practical problems such as traffic data analysis. Our new Bayesian gIRL method with the extended LMDP can estimate the cost functions, the transition probability, and the hyperparameters.
- We confirm the effectiveness of the proposed method by performing numerical experiments using both synthetic data and real car probe data. The result of our car probe data experiment shows that the proposed method can estimate the LMDP parameters, which reflect car drivers' behavior.

The rest of this paper is organized as follows. In §2, we introduce the LMDP. The non-identifiable problem of the LMDP is illustrated in §3. §4 presents the extended LMDP and §5 introduces the proposed gIRL method. §6 is devoted to the experimental evaluation and §7 concludes the paper.

2 Linearly Solvable MDP (LMDP)

In this section, a basic property of the LMDP [5] is introduced. Although the definition of the LMDP is similar to that of the MDP, its difference is critical in creating solutions to the forward and inverse problems. Note that this work focuses on an "infinite horizon discounted cost" case [13]; however, its application to other settings is straightforward.

Definition of LMDP: The LMDP is defined by the quintuplet $\{\mathcal{S}, \mathcal{A}, \bar{\mathcal{P}}, \mathcal{R}, \gamma\}$, where $\mathcal{S} = \{1, 2, \cdots, S\}$ is a finite set of *states* and S is the number of states. $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_S\}$ is a set of admissible actions at each states. $\bar{\mathcal{P}} = \{\bar{p}_{jk}\}_{j,k=1}^{S}$ indicates *passive transition probabilities*, each element of which defines the transition probability from state j to state k when an action is *not* executed. $\mathcal{R} : \mathcal{S} \to \mathbb{R}$ is a *state cost function* (*negative reward function*) and we denote the state cost at state j as r_j. $\gamma \in [0, 1)$ is a *discount factor*.

In the LMDP, *action* $\boldsymbol{a}$ is a continuous valued $\mathbb{R}^S$ dimensional vector and the *action transition probability* from state j to state k when action $\boldsymbol{a}_j = \{a_{jk}\}_{k=1}^{S}$ is executed is defined by

$$p_{jk}(\boldsymbol{a}_j) = \bar{p}_{jk} \exp(a_{jk}). \tag{1}$$

Note that any action executed at state j, $\boldsymbol{a}_j$, must belong to a set of admissible actions, $\mathcal{A}_j$, which is defined as

$$\mathcal{A}_j = \left\{ \boldsymbol{a}_j \in \mathbb{R}^S \middle| \sum\nolimits_k p_{jk}(\boldsymbol{a}_j) = 1; \bar{p}_{jk} = 0 \to a_{jk} = 0 \right\}, \tag{2}$$

so that the sum of the probabilities equals one. Therefore, the transition probability itself can be controlled by an action. To execute a certain action, it is

necessary to pay the action cost defined by *action cost function*. The action cost when action $\boldsymbol{a}_j$ is executed in state j is defined as

$$q_j(\boldsymbol{a}_j) = KL(\boldsymbol{p}_j(\boldsymbol{a}_j)||\boldsymbol{p}_j(\boldsymbol{0})), \tag{3}$$

where $KL(\cdot||\cdot)$ is the Kullback-Leibler divergence and $\boldsymbol{p}_j(\boldsymbol{a}) = \{p_{jk}(\boldsymbol{a})\}_{k=1}^S$. Thus, the action cost increases as $p_{jk}(\boldsymbol{a})$ deviates further from a passive transition $\bar{p}_{jk}$. Note that when the action is a zero vector, $\boldsymbol{a} = \boldsymbol{0}$, $p_{jk}(\boldsymbol{0})$ equals the passive transition probability $\bar{p}_{jk}$ and the action cost $q_j(\boldsymbol{0}) = 0$. Intuitively, the LMDP is a class of MDP in which the transition probability itself can be controlled by the payment of the action cost. Unlike standard MDP where the transition probability is defined separately for each action, the passive transition probability substantially determines the transition probability of all the actions. Thus, we consider that it is suitable to use the LMDP for gIRL.

Let $\boldsymbol{\pi} = \{\boldsymbol{a}_j\}_{j=1}^S$ be a policy whose element $\boldsymbol{a}_j$ indicates the action executed in state j. The *value function* of policy π, $\boldsymbol{v}^\pi = \{v_j^\pi\}_{j=1}^S$, is defined such that element v_j^π indicates the expected sum of the future cost from state j when following policy $\boldsymbol{\pi}$,

$$v_j^\pi = \lim_{T\to\infty} \mathbb{E}_{\boldsymbol{d}^T}\left[\sum_{t=1}^T \gamma^{t-1}\left\{r_{s_t} + q_{s_t}(\boldsymbol{a}_{s_t})\right\}\middle| s_1 = j\right]. \tag{4}$$

Here $\mathbb{E}_{\boldsymbol{d}^T}$ denotes the expectation over trajectory $\boldsymbol{d}^T = \{s_t\}_{t=1}^T$, the transitions from $t = 1$ to T where s_t denotes the visit state at time t, which follow probability $P(\boldsymbol{d}^T|\bar{\mathcal{P}}, \boldsymbol{\pi}) = p_{s_1}^{ini}\prod_{t=1}^{T-1} p_{s_t s_{t+1}}(\boldsymbol{a}_{s_t})$. p^{ini} is the initial state distribution.

Forward Problem with LMDP: The forward problem with the LMDP is to obtain the optimal policy $\boldsymbol{\pi}^* = \{\boldsymbol{a}_j^*\}_{j=1}^S$ that minimizes the expected sum of the future cost. The optimal action in state j is given by

$$\boldsymbol{a}_j^* = \arg\min_{\boldsymbol{a}_j\in\mathcal{A}_j}\left\{r_j + q_j(\boldsymbol{a}_j) + \gamma\sum_{k=1}^S p_{jk}(\boldsymbol{a}_j)v_k\right\} = -\gamma v_j - \log\Big(\sum_{k=1}^S \bar{p}_{jk}\exp(-\gamma v_k)\Big), \tag{5}$$

where $\boldsymbol{v} = \{v_j\}_{j=1}^S$ is the optimal value function $v_j = \min_\pi v_j^\pi$ that can be computed by solving the optimal equation [5]. Inserting Eq. (5) into Eq. (1), *optimal transition probability*, the action transition probability when the optimal action being executed is written as

$$p_{jk}^* = p_{jk}(\boldsymbol{a}_j^*) = \frac{\bar{p}_{jk}\exp(-\gamma v_k)}{\sum_\ell \bar{p}_{j\ell}\exp(-\gamma v_\ell)}. \tag{6}$$

We emphasize that the above form of optimal transition probability is a direct consequence of the LMDP unlike Bayesian IRL, which uses the value function as a potential function [9].

3 Generalized IRL and the Non-identifiable Problem

3.1 Generalized Inverse Reinforcement Learning

This section illustrates the non-identifiable problem of generalized IRL (gIRL) with the LMDP. The purpose of gIRL is to estimate the state cost function and passive transition probability of the LMDP from a transition log that follows the optimal transition probability Eq. (6). Figure 1 illustrates the forward and inverse problems of the LMDP.

A key motivation for gIRL can be explained as follows. Let us consider a case where the cost function must be estimated only from the past movements of a person who is interested in a certain place in a city. In this case, the passive transition probability between places, which can be interpreted as the transition probability of a person who has a uniform state cost function (same degree of interest in each place), is of course unknown and cannot be observed. That is, gIRL is useful for estimating a state cost function when only a set of past movements is available, which is a common setting in various machine learning problems.

To determine whether the state cost function and passive transition probability can be uniquely estimated or not, we consider a case where the amount of available data is sufficiently large. In this case, the optimal transition probability itself can be observed. Therefore, we need to seek the corresponding relation between the optimal transition probability and a pair consisting of a state cost function and a passive transition probability.

3.2 Toy Example of Non-identifiable Problem

Figure 2 shows a toy example in which LMDPs with different passive transitions and value functions provide equivalent optimal transition probabilities. In the left dotted box in Fig. 2, the passive transition from state-1 to state-1, $\bar{p}_{11}$, is p and the value function of state-1, v_1, is v. Similarly, in the right dotted box, they are $p' = p/(K - pK + p)$ and $v' = -\log(K)/\gamma + v$. We can easily confirm that the optimal transition probabilities for both LMDPs are equivalent for arbitrary constant K. This means that we cannot identify the passive transition probability and value function simultaneously from the optimal transition probability. By considering the transition probability and value function to be parameters and the optimal transition probability to be a probabilistic model, this is a case where the model is *non-identifiable* [2]. Note that the above examples are compatible with the claim made by Dvijotham and Todorov [6]. Since they consider a case where the passive transition probability is known, the value function can be uniquely estimated.

Non-identifiability implies the impossibility of estimating the state cost function uniquely. Let us consider the setting of the constant value $K = \exp(\gamma v)$ in

[2] The probabilistic model P_θ is called *identifiable* in statistics if parameter $\theta_1 \neq \theta_2$, then distributions P_{θ_1} and P_{θ_2} are different [14]. A model which is not *identifiable* is called *non-identifiable*.

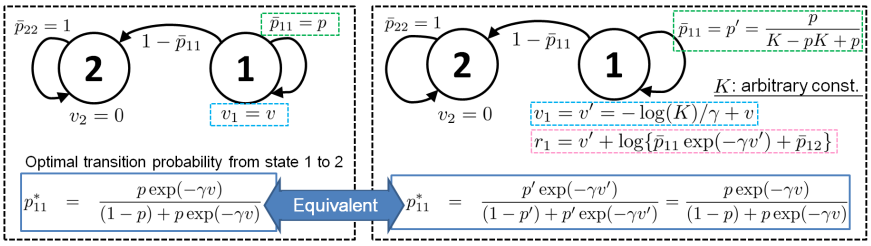


Fig. 2. An example that indicates that LMDPs with different passive transitions and value functions provides an equivalent optimal transition probability. This implies that the passive transition probability and the value function cannot be uniquely estimated even if the optimal transition probability itself is observed. We call this problem the non-identifiable problem of gIRL with LMDP.

the previous toy example. In this case, the value and the state cost of state 1 become 0, $v_1 = r_1 = 0$, and the passive transition equals the optimal transition, $\bar{p}_{11} = p^*_{11}$. This means that the optimal transition probability of the left LMDP can be reproduced by the right LMDP with a state cost function whose values are all zero. This fact immediately leads to the following theorem:

Theorem 1 (Non-identifiability of gIRL with LMDP) *Let $\mathcal{S}$ and γ be a set of states and a discount factor. Then, the mapping from a pair consisting of passive transition probability $\bar{\mathcal{P}} \in [0,1]^{S\times S}$ and state cost function $\mathcal{R} \in \mathbb{R}^S$ to the optimal transition probability of the LMDP $(\mathcal{S}, \mathcal{A}, \bar{\mathcal{P}}, \mathcal{R}, \gamma)$ is not one-to-one.*

Proof *When $\mathcal{R} = \mathbf{0}$, the passive transition probability and optimal transition probability are identical. Then, for any LMDP, there exists an LMDP that has an all zero state cost function and a passive transition probability that is identical to the optimal transition probability of a given LMDP.* □

This is an obviously unacceptable result because the transition probability and cost function have different roles in RL; cost is a target of the agent to be minimized, the transition probability determines the possible movements of the agent. Their two roles should not be mixed.

It is well-known for IRL with standard MDP that the cost function with entirely zero values is always one of the solutions [7]. Therefore, our observation indicates that the generalized IRL problem with the LMDP also raises similar theoretical concerns.

Remark 1 Note that we do not view this as a problem that the optimal transition is consistent by transformation $v'_i = v_i + c$ for all states i using a common constant value c while the passive transition probability remains fixed; this is because the magnitude relation of the value function holds. This type of degree of freedom can be removed by, for example, setting the value function of a certain state at zero. The problem tackled in this paper is the non-identifiability of the value function and the passive transition probability.

3.3 Approach for Non-identifiable Problem

We confirmed above that gIRL with LMDP suffers from non-identifiability. This subsection introduces an idea that can avoid this problem. A promising approach is to introduce hyperparameters. This approach is the same as that used by Ng and Russel for IRL with MDP to avoid non-identifiability [7]. They introduce hyperparameters to make the cost function sparse, i.e., the cost becomes zero in many states. However, as stated in their paper, a remaining problem was that the estimated result strongly depends on the manual setting of the hyperparameters. Therefore, we construct a gIRL method with a Bayesian framework that can estimate hyperparameters. By automatically estimating the hyperparameters, their dependency is weakened. The Bayesian approach is promising since its effectiveness has already been confirmed for standard IRL with an MDP [9–11]. We also introduce a new gIRL formulation for a later experiment.

Our new formulation considers a collection of LMDPs that share state and passive transition probabilities. The setting is referred to as multiple intention or multitask IRL in the literature [11, 12]. The following example explains the motivation behind using this new formulation. Again, let us consider a case where the state cost function of a certain person needs to be extracted. If the trajectories of several people are available, the task seems obvious when we consider that only the cost function alone depends on each person and the passive transition probability is not person dependent. Thus, our new gIRL is formulated as the problem of estimating everybody's cost functions and common passive transition probabilities from observed trajectories. Thanks to this formulation, hyperparameters for the cost function are defined as common parameters among all people; this may contribute to performance improvement similar to that described in [11]. The next two sections present a rigorous formulation and an estimation algorithm.

4 Shared-parameter LMDPs

In this section, we re-formulate gIRL. We consider a collection of LMDPs that share states $\mathcal{S}$, passive transition probability $\bar{\mathcal{P}}$, and discount factor γ. Each LMDP has its own state cost function, $\mathcal{R}_i$, where i is the index of the LMDP. We call this collection of LMDPs, *shared-parameter LMDPs* (SP-LMDPs). We formulate gIRL as an inverse learning problem to estimate the passive transition probability and all the state cost functions in the SP-LMDPs. Figure 3 shows all the parameters of the SP-LMDPs. gIRL with the SP-LMDPs is a natural extension of IRL because gIRL with SP-LMDPs can be seen as a setting in which multiple state cost functions are estimated. Each cost function may be the cost function of a different person, animal and so on. We emphasize that the setting at which multiple cost functions are defined on a *standard* MDP has already been studied [11, 12] but it has not been studied for an LMDP.

We provide a formal definition of SP-LMDPs as follows. SP-LMDPs are defined by the quintuplet $\{\mathcal{S}, \mathcal{A}, \bar{\mathcal{P}}, \boldsymbol{\mathcal{R}}, \gamma\}$. The definitions of $\mathcal{S}$, $\mathcal{A}$, $\mathcal{P}$ and γ are equivalent to those for an LMDP while that of $\boldsymbol{\mathcal{R}}$ is different. $\boldsymbol{\mathcal{R}} = (\mathcal{R}_1, \cdots, \mathcal{R}_I)$

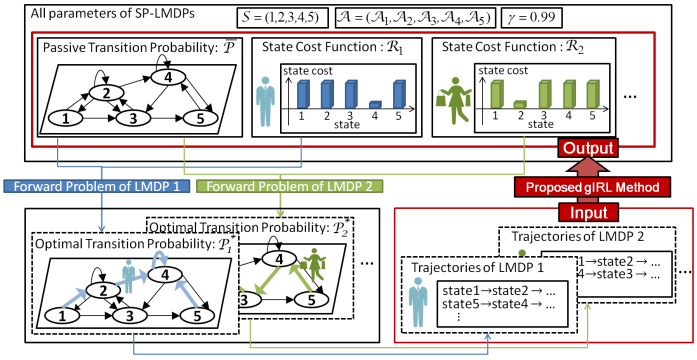


Fig. 3. Forward and inverse problems of SP-LMDPs. We call this inverse problem, which estimates all the state cost functions and passive transition probabilities of SP-LMDPs, gIRL with SP-LMDPs.

is a set of *state cost functions* and $\mathcal{R}_i = \{r_{ij}\}_{j=1}^S$. I is the number of functions in the set. From this definition, we can construct I LMDP; the i-th LMDP is defined as $\{\mathcal{S}, \mathcal{A}, \bar{\mathcal{P}}, \mathcal{R}_i, \gamma\}$. Note that SP-LMDPs with $I = 1$ reduce to an LMDP.

Since the forward problem of the i-th LMDP can be solved independently following the method explained in §2, SP-LMDPs pose no difficulty in solving the forward problem. Let us define $\boldsymbol{V} = \{\boldsymbol{v}_i\}_{i=1}^I$. $\boldsymbol{v}_i = \{v_{ij}\}_{j=1}^S$ is the optimal value function of the i-th LMDP. This optimal value function satisfies the following optimal equation.

$$v_{ij} = \min_{\boldsymbol{a}_{ij} \in \mathcal{A}_j} \left\{ r_{ij} + q_j(\boldsymbol{a}_{ij}) + \gamma \sum\nolimits_k p_{jk}(\boldsymbol{a}_{ij}) v_{ik} \right\} = r_{ij} - \log\left(\sum\nolimits_k \bar{p}_{jk} \exp(-\gamma v_{ik}) \right). \tag{7}$$

Then, the optimal transition probability from state j to state k is, for the i-th LMDP, given by

$$p^*_{ijk} = \frac{\bar{p}_{jk} \exp(-\gamma v_{ik})}{\sum_\ell \bar{p}_{j\ell} \exp(-\gamma v_{i\ell})}. \tag{8}$$

The above optimal transition probability shows that the agent executing the optimal policy tends to move adjacent states whose value functions are small.

5 Proposed generalized IRL method

5.1 Bayesian Modeling

This subsection details the proposed gIRL method, which can estimate both the state cost functions and passive transition probabilities with SP-LMDPs from observed transitions. We denote the transition logs of the i-th LMDP as

$\mathcal{D}_i$ and the number of observed transitions from state j to state k in the i-th LMDP as n_{ijk}. We also denote all the transition logs as $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^{I}$. Our gIRL method is naturally derived by considering that each transition is generated by the probability defined in Eq. (8) which has parameters $\boldsymbol{V}, \bar{\mathcal{P}}$. In this section, we re-parametrize $\bar{p}_{jk}$ as $w_{jk} = -\log \bar{p}_{jk}$. We define $\boldsymbol{W} = \{\boldsymbol{w}_j\}_{j=1}^{S}$ and $\boldsymbol{w}_j = \{w_{jk}\}_{k=1}^{S}$. Then, the probability that transition log $\mathcal{D}$ is generated given parameter $\boldsymbol{V}, \boldsymbol{W}$ can be written as

$$P(\mathcal{D}|\boldsymbol{V}, \boldsymbol{W}) = \prod_i \prod_{j,k \in \mathcal{S}} \Big(\frac{\exp(-w_{jk} - \gamma v_{ik})}{\sum_\ell \exp(-w_{j\ell} - \gamma v_{i\ell})} \Big)^{n_{ijk}}. \tag{9}$$

We can avoid the ill-posedness of gIRL, and also obtain the full parameter-estimation procedure by adopting a Bayesian approach. We define a Gaussian prior distribution on $\boldsymbol{v}_i$ and $\boldsymbol{w}_j$ for all i, j given by

$$P(\boldsymbol{V}|\alpha) = \prod_{i,j=1}^{I,S} \mathcal{N}(v_{ij}|0, \frac{1}{\alpha}),\ P(\boldsymbol{W}|\beta) = \prod_{j=1}^{S} \prod_{k \in \Omega_j^{\mathrm{fr}}} \mathcal{N}(w_{jk}|0, \frac{1}{\beta}). \tag{10}$$

Note that Ω_j^{fr} denotes a set of reachable states "from" state j by a one step transition.[3] We also used a conjugate gamma prior on the hyper-parameters, similar to [15]:

$$P(\alpha) = \mathcal{G}(\alpha|a_0, b_0) = \frac{a_0^{b_0}}{\Gamma(a_0)} \alpha^{a_0 - 1} e^{-b_0 \alpha},\ P(\beta) = \mathcal{G}(\beta|a_0, b_0) = \frac{a_0^{b_0}}{\Gamma(a_0)} \beta^{a_0 - 1} e^{-b_0 \beta}. \tag{11}$$

We set $a_0 = 10^{-1}$ and $b_0 = 10^{-2}$ in an experiment described later. Summarizing the above, we denote the joint distribution of all the parameters and the set of trajectories as

$$P(\mathcal{D}, \boldsymbol{V}, \boldsymbol{W}, \alpha, \beta) = P(\mathcal{D}|\boldsymbol{V}, \boldsymbol{W}) \underbrace{P(\boldsymbol{V}|\alpha) P(\boldsymbol{W}|\beta) P(\alpha) P(\beta)}_{P(\boldsymbol{V}, \boldsymbol{W}, \alpha, \beta)}. \tag{12}$$

Figure 4(a) shows a graphical model representation. The posterior distribution of parameters is given by

$$P(\boldsymbol{V}, \boldsymbol{W}, \alpha, \beta|\mathcal{D}) = P(\mathcal{D}, \boldsymbol{V}, \boldsymbol{W}, \alpha, \beta)/P(\mathcal{D}), \tag{13}$$

where $P(\mathcal{D})$ is the marginal likelihood $P(\mathcal{D}) = \int P(\mathcal{D}, \boldsymbol{V}, \boldsymbol{W}, \alpha, \beta) d\boldsymbol{V} d\boldsymbol{W} d\alpha d\beta$. Since the exact computation of the marginal likelihood is infeasible, we adopt the variational Bayesian (VB) approach [16] to obtain the posterior distribution.

[3] If such adjacency information is not available, consider Ω_j^{fr} as a set of all states $\mathcal{S}$.

5.2 Variational Bayes

The VB algorithm is designed to obtain the variational distributions that approximate the posterior distribution. The variational distribution $q(\boldsymbol{V}, \boldsymbol{W}, \alpha, \beta)$ is estimated by minimizing functional $\tilde{\mathcal{F}}[q, \boldsymbol{\eta}, \boldsymbol{\xi}]$, which is defined by

$$\tilde{\mathcal{F}}[q, \boldsymbol{\eta}, \boldsymbol{\xi}] := \mathbb{E}_q\Big[\log \frac{q(\boldsymbol{V}, \boldsymbol{W}, \alpha, \beta)}{h(\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\eta}, \boldsymbol{\xi}) P(\boldsymbol{V}, \boldsymbol{W}, \alpha, \beta)}\Big] \tag{14}$$

under the constraint that the parameters are independent: $q(\boldsymbol{V}, \boldsymbol{W}, \alpha, \beta) = q(\boldsymbol{V})q(\boldsymbol{W})q(\alpha)q(\beta)$. Note that $h(\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\eta}, \boldsymbol{\xi})$ is a lower bound of the likelihood function (Eq. (9)), i.e., $h(\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\eta}, \boldsymbol{\xi}) \leq P(\mathcal{D}|\boldsymbol{V}, \boldsymbol{W})$ for all $\boldsymbol{V}, \boldsymbol{W}$. $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are auxiliary variables. The functional $\tilde{\mathcal{F}}[q, \boldsymbol{\eta}, \boldsymbol{\xi}]$ is an upper bound of the negative log marginal likelihood $-\log P(\mathcal{D})$. By minimizing $\tilde{\mathcal{F}}[q, \boldsymbol{\eta}, \boldsymbol{\xi}]$, we can indirectly minimize the Kullback-Leibler (KL) divergence between the variational distributions and posterior distribution.

Figure 4(b) makes it easier to understand our optimization scheme. We define functional $\bar{\mathcal{F}}$ as follows:

$$\bar{\mathcal{F}}[q] := \mathbb{E}_q\Big[\log \frac{q(\boldsymbol{V})q(\boldsymbol{W})q(\alpha)q(\beta)}{P(\mathcal{D}, \boldsymbol{V}, \boldsymbol{W}, \alpha, \beta)}\Big]. \tag{15}$$

This is also an upper bound of the negative log marginal likelihood, and its difference is given by the KL divergence between variational distributions and the posterior distribution (See the green box in Fig. 4(b)). $\tilde{\mathcal{F}}[q, \boldsymbol{\eta}, \boldsymbol{\xi}]$ is always greater than $\bar{\mathcal{F}}[q]$, and its difference is given by the average log ratio of function h and the likelihood function (See the blue box). Since log marginal likelihood does not depend on variational distributions, minimizing $\tilde{\mathcal{F}}$ w.r.t. variational distribution q corresponds to minimizing the sum of the KL divergence and the average log bound-likelihood ratio. Minimizing $\tilde{\mathcal{F}}$ w.r.t. auxiliary variables $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ corresponds to minimizing the average log bound-likelihood ratio. Iterating this procedure yields a variational distribution.

Remark 2 For probabilistic models belonging to an exponential family, the VB algorithm is derived by using $\bar{\mathcal{F}}[q]$ as the objective functional. However, since the softmax function in Eq. (9) breaks the conjugate-exponential structure in our model, we make use of upper bound function h. The use of a bound function in VB can be found in logistic regression [17], mixture of experts [15] and the correlated topic model [18].

There are various choices for function h since several bounds of the softmax function have been derived [18–21]. From here, we use the following definition of function h, which is a quadratic form with respect to v_{ij}, w_{jk}, by using the bound described by Bouchard [20]. This choice yields an analytical update equation that is easy to implement.

$$\begin{aligned}\log h(\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\eta}, \boldsymbol{\xi}) = &\sum\nolimits_{j,k} -n_{\cdot jk} w_{jk} + \sum\nolimits_{ij} -n_{i\cdot j}\gamma v_{ij} \\ &- \sum\nolimits_{ij} n_{ij\cdot}\big\{\eta_{ij} + \sum\nolimits_{\ell} f(-w_{j\ell} - \gamma v_{i\ell}, \eta_{ij}, \xi_{ij\ell})\big\},\end{aligned} \tag{16}$$

$$f(x_\ell, \eta, \xi_\ell) = \log(1 + e^{\xi_\ell}) + (x_\ell - \eta - \xi_\ell)/2 + \lambda(\xi_\ell)\{(x_\ell - \eta)^2 - \xi_\ell^2\}, \tag{17}$$

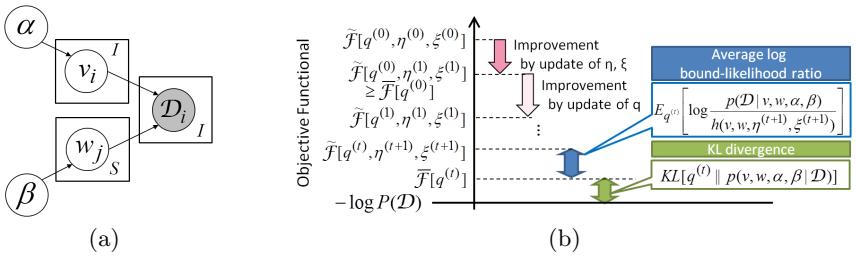


Fig. 4. (a) Graphical model. Shaded nodes indicate observed variables. Dependency on a_0 and b_0 is omitted for clarity. (b) Optimization scheme of the proposed algorithm.

where $\lambda(\xi_\ell) = \frac{1}{2\xi_\ell}(\sigma(\xi_\ell) - 1/2)$ and $\sigma(\cdot)$ is a sigmoid function. The dot index means that the corresponding index is summed out: $n_{\cdot jk} = \sum_i n_{ijk}$, $n_{i\cdot k} = \sum_j n_{ijk}$, $n_{ij\cdot} = \sum_k n_{ijk}$. We can easily confirm that this h is a lower bound of likelihood $P(\mathcal{D}|\boldsymbol{V}, \boldsymbol{W})$ by the following theorem.

Theorem 2 (Bouchard) [20] *For any $x \in \mathbb{R}^L$, any $\eta \in \mathbb{R}$ and any $\xi \in [0, \infty)^L$, the following inequality holds:* $\log\left(\sum_{\ell=1}^L e^{x_\ell}\right) \leq \eta + \sum_{\ell=1}^L f(x_\ell, \eta, \xi_\ell)$.

We construct an algorithm that iteratively updates the variational distribution q and auxiliary variables $\boldsymbol{\xi}, \boldsymbol{\eta}$. Algorithm 1 summarizes the parameter estimation procedure. Parameter update is explained as follows.

Update of Variational Distribution q:
With the variational method, the optimal variational distribution must satisfy the following optimal equation:

$$q(\boldsymbol{V}) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{W})q(\alpha)}\left[\log h(\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\eta}, \boldsymbol{\xi}) p(\boldsymbol{V}|\alpha)\right]\right), \tag{18}$$

$$q(\boldsymbol{W}) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{V})q(\beta)}\left[\log h(\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\eta}, \boldsymbol{\xi}) p(\boldsymbol{W}|\beta)\right]\right), \tag{19}$$

$$q(\alpha) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{V})}\left[\log p(\boldsymbol{V}|\alpha) p(\alpha)\right]\right), \tag{20}$$

$$q(\beta) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{W})}\left[\log p(\boldsymbol{W}|\beta) p(\beta)\right]\right). \tag{21}$$

The above distributions are given by elementwise Gaussian distribution $q(v_{ij}) = \mathcal{N}(v_{ij}|\mu^v_{ij}, (\sigma^v_{ij})^2)$, $q(w_{jk}) = \mathcal{N}(w_{jk}|\mu^w_{jk}, (\sigma^w_{jk})^2)$ and gamma distributions $q(\alpha) = \mathcal{G}(\alpha|a_\alpha, b_\alpha)$, $q(\beta) = \mathcal{G}(\beta|a_\beta, b_\beta)$, where $\mu^v_{ij}, \sigma^v_{ij}, \mu^w_{jk}, \sigma^w_{jk}, a_\alpha, b_\alpha, a_\beta, b_\beta$ are variational parameters.

$$\mu^v_{ij} = \left[-n_{i\cdot j} + \sum\nolimits_{k \in \Omega^{\mathrm{to}}_j} \left\{\frac{n_{ik\cdot}}{2} - 2n_{ik\cdot}\lambda(\xi_{ikj})(\bar{w}_{kj} + \eta_{ik})\right\}\right]\gamma(\sigma^v_{ij})^2, \tag{22}$$

$$\sigma^v_{ij} = \left\{\bar{\alpha} + \sum\nolimits_{k \in \Omega^{\mathrm{to}}_j} 2n_{ik\cdot}\lambda(\xi_{ikj})\gamma^2\right\}^{-\frac{1}{2}}, \tag{23}$$

$$\mu^w_{jk} = \left[-n_{\cdot jk} + \frac{n_{\cdot j\cdot}}{2} + \sum\nolimits_i 2n_{ij\cdot}\lambda(\xi_{ijk})(-\gamma\bar{v}_{ik} - \eta_{ij})\right](\sigma^w_{jk})^2, \tag{24}$$

$$\sigma^w_{jk} = \left\{\bar{\beta} + \sum\nolimits_i 2n_{ij\cdot}\lambda(\xi_{ijk})\right\}^{-\frac{1}{2}}, \tag{25}$$

$$a_\alpha = a_0 + \frac{IS}{2},\ b_\alpha = b_0 + \frac{1}{2}\sum\nolimits_{ij} \mathbb{E}_{q(\boldsymbol{V})}[v^2_{ij}]. \tag{26}$$

Algorithm 1 Proposed VB Algorithm for gIRL

input $\mathcal{D}$: observed transitions, γ: discount factor
output $\mu^v_{ij}, \sigma^v_{ij}, \mu^w_{jk}, \sigma^w_{jk}, a_\alpha, b_\alpha, a_\beta, b_\beta$: variational parameters.

```
Initialization.
repeat
    //parameters for variational distribution q
    Update μ^v_ij, σ^v_ij following Eq. (22)(23).
    Update μ^w_jk, σ^w_jk following Eq. (24)(25).
    Update a_α, b_α, a_β, b_β following Eq. (26)(27).
    //auxiliary variables ξ, η
    Update ξ_ijℓ, η_ij following Eq. (29)(30).
until converge
```

$$a_\beta = a_0 + \frac{\sum_j |\Omega_j^{\mathrm{fr}}|}{2},\ b_\beta = b_0 + \frac{1}{2}\sum\nolimits_j \sum\nolimits_{k\in\Omega_j^{\mathrm{fr}}} \mathbb{E}_{q(\boldsymbol{W})}[w_{jk}^2]. \tag{27}$$

Note that Ω_j^{to} denotes a set of states that can reach "to" state j by a one-step transition and some statistics are given by the following equations: $\bar{v}_{ij} = \mu^v_{ij}$, $\bar{w}_{jk} = \mu^w_{jk}$, $\bar{\alpha} = a_\alpha/b_\alpha$, $\bar{\beta} = a_\beta/b_\beta$, $\mathbb{E}_{q(\boldsymbol{V})}[v_{ij}^2] = (\sigma^v_{ij})^2 + (\mu^v_{ij})^2$, $\mathbb{E}_{q(\boldsymbol{W})}[w_{jk}^2] = (\sigma^w_{jk})^2 + (\mu^w_{jk})^2$.

The proposed algorithm works by iteratively updating the variational parameters. Note that the objective functional is monotonically decreased by the updates and thus converges to a local minimum.

Update of Auxiliary Parameter $\boldsymbol{\eta}, \boldsymbol{\xi}$:
Since only the term $\log h(\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\eta}, \boldsymbol{\xi})$ depends on $\boldsymbol{\xi}$ in the objective functional Eq.(14), at the optimal point, its partial derivative must satisfy

$$\begin{aligned} &\frac{\partial}{\partial \xi_{ij\ell}} \mathbb{E}_{q(\boldsymbol{V})q(\boldsymbol{W})q(\alpha)q(\beta)}\Big[-\log h(\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\eta}, \boldsymbol{\xi})\Big] = 0 \\ &\Leftrightarrow (\sigma^w_{j\ell})^2 + \gamma^2(\sigma^v_{i\ell})^2 + (-\bar{w}_{j\ell} - \gamma\bar{v}_{i\ell} - \eta_{ij})^2 - \xi^2_{ij\ell} = 0. \end{aligned} \tag{28}$$

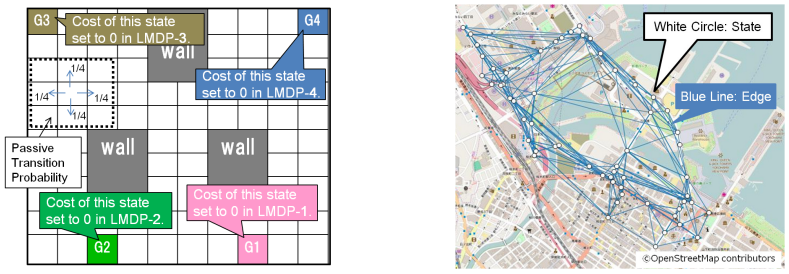
Therefore, we develop the following update rule:

$$(\xi^{\mathrm{new}}_{ij\ell})^2 \leftarrow (\sigma^w_{j\ell})^2 + \gamma^2(\sigma^v_{i\ell})^2 + (-\bar{w}_{j\ell} - \gamma\bar{v}_{i\ell} - \eta_{ij})^2. \tag{29}$$

Similarly, the update rule for η is given by

$$\eta^{\mathrm{new}}_{ij} \leftarrow \Big\{\frac{1}{2}\Big(\frac{|\Omega_j^{\mathrm{fr}}|}{2} - 1\Big) + \sum_{\ell\in\Omega_j^{\mathrm{fr}}} \lambda(\xi_{ij\ell})(-\bar{w}_{j\ell} - \gamma\bar{v}_{i\ell})\Big\} \Big/ \Big\{\sum_{\ell\in\Omega_j^{\mathrm{fr}}} \lambda(\xi_{ij\ell})\Big\}. \tag{30}$$

In the process of parameter estimation, state cost function $\boldsymbol{\mathcal{R}}$ need not be considered. However, if necessary, using optimal equation (7), the estimated function $\hat{\boldsymbol{\mathcal{R}}} = \{\hat{r}_{ij}\}$ is obtained as $\hat{r}_{ij} = \bar{v}_{ij} + \log\big(\sum_k \exp(-\bar{w}_{jk} - \gamma\bar{v}_{ik})\big)$ after parameter estimation. Then, the estimated value function is the optimal value function of the LMDP with the above estimated state cost function.



(a) Grid world (b) Yokohama

Fig. 5. Settings for (a) grid world where uniform passive transition probability and four types of state cost functions are set and (b) yokohama using real car probe data

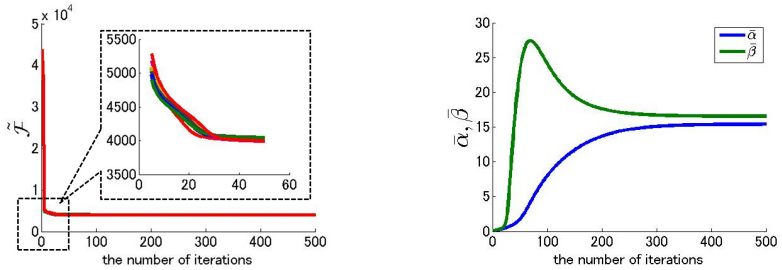
6 Numerical Experiment

6.1 Experimental Settings

This section confirms the validity of the proposed method. We conduct a numerical experiment to determine (i) convergence property, (ii) predictive performance and (iii) parameter visualization.

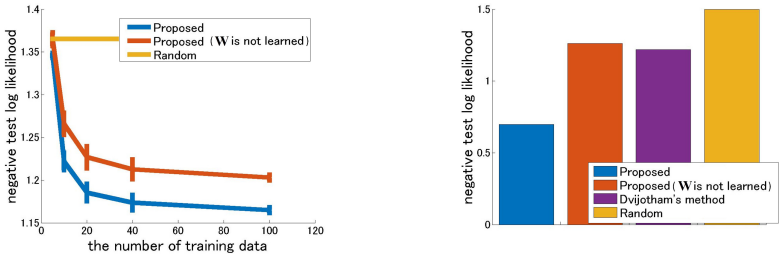
Data Description: We prepare two experiment settings: grid-world and yokohama. In the grid-world experiment, we set the passive transition probability of each state (vertical and horizontal) at a uniform probability (if walls or obstacles exist, self-transition is to be considered) and prepared four different types of state cost functions: $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ and $\mathcal{R}_4$. The cost of each function is set at 0 just for the corresponding goal state shown in Fig. 5 (a) and at 1 for the other states. By computing the true optimal transition probability of each LMDP, we generate training and test data in an *iid* manner in each state. In the yokohama experiment, we use real car probe data provided by NAVITIME JAPAN Co, Ltd. This dataset is a collection of GPS trajectories of users who used a car navigation application on smartphones in Kanagawa Prefecture, Japan. In particular, we used the trajectories for the Minato-Mirai-21 district in Yokohama. We use the log data recorded during the holiday period from 2015.4.13 to 2015.5.1 (5 days in total) as training data and the log data of 2015.5.2 as test data. By applying a landmark graph construction algorithm [22], we construct the abstract street network as shown in Fig. 5 (b). We convert the GPS into transition data between the nodes (states) of this graph. We treat the logs of 10:00-12:59, 14:00-16:59, 17:00-19:59 as the logs of LMDP1, 2 and 3, respectively.

Predictive Performance Measurement: To evaluate the predictive performance, we use the negative test log likelihood. A lower value indicates that the method extracts the parameter that reflects the agent's behavior more precisely. The negative test log likelihood is defined as $(1/\mathcal{T}) \sum_{i=1}^{I} \sum_{j,k \in \mathcal{S}} -n_{ijk}^{\text{test}} \log \hat{p}_{ijk}^{*}$, where $\mathcal{T}$ is the number of test datasets and n_{ijk}^{test} indicates the number of transitions from state j to state k in the i-th LMDP. $\hat{p}_{ijk}^{*}$ is computed by substituting $\bar{v}_{ij}$ and $\bar{w}_{kj}$ into Eq. (8). We compare the proposed method with Random



(a) Objective function (b) Hyperparameters

Fig. 6. Convergence behavior of (a) objective function and (b) hyperparameters in grid world experiment with $n_{ij\cdot} = 5$. (a) shows the result of 10 random initialization settings and (b) shows one of the paths from the initial point.



(a) Grid world (b) Yokohama

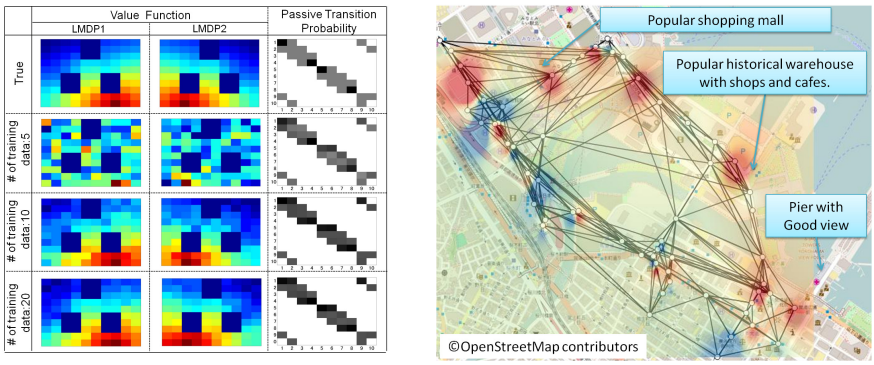
Fig. 7. Comparison of predictive performance of (a) grid world and of (b) yokohama experiment. Lower values are better.

and Dvijotham's method [6]. Since Dvijotham's method can estimate only the cost function, we set the passive transition probability at a uniform probability. Moreover, to investigate the effect of passive transition probability estimation, we also make a comparison with the proposed method, which does not learn the passive transition probability (fixed at a random initial value).

6.2 Results

Convergence Behavior: Figure 6 shows the convergence behavior of the objective function and hyperparameters. We can confirm that they both converge to certain values by iterating the update process. This shows that the proposed method can estimate hyperparameters. In terms of convergence speed, Fig. 6 (a) shows that the objective function basically converges within 50 iterations. In contrast, Fig. 6 (b) shows that more than 200 iterations are needed for hyperparameter convergence. These results imply that a relatively longer running time is required in order to learn the hyperparameters.

Predictive Performance: Figure 7 (a) shows the predictive performance in the grid world experiment. In comparison with the proposed method without



(a) Grid world (b) Yokohama

Fig. 8. (a) True and estimated value functions of LMDP 1,2 and passive transition probabilities for states 1∼10 of grid world for various numbers of observed transitions $n_{ij\cdot} = 5, 10, 20$. (b) Estimated value function of LMDP1 of yokohama. Value functions are visualized by a heat map with colors ranging from red to blue.

learning the passive transition probability, the proposed method shows better predictive performance. This result shows that estimating the passive transition probability contributes to better performance. Figure 7 (b) shows the predictive performance in the yokohama experiment. Dvijotham's method is competitive with the proposed method that does not learn the passive transition probability but the proposed method outperforms them. This also confirms the effectiveness of the proposed method.

Parameter Visualization: Figure 8(a) shows the estimated parameters in the grid world experiment for various numbers of observed transitions $n_{ij\cdot}$ (visualization of LMDP-3 and 4 is omitted due to lack of space). We can confirm that as the number of observed transitions increases, the estimated parameters more closely approach the true parameters. Figure 8(b) shows the estimated parameters of the yokohama experiment[4]. Although we are unable to know the true parameters behind the real car probe data, we observe that the state near attractive locations has a lower value function value. We can predict that the agent (car driver) tends to move to the locations. This result implies that the proposed method estimates parameters that reflect car drivers' behavior.

7 Conclusion and Future Work

In this paper, we tackled the gIRL problem to estimate both the state cost function and transition probability from the observed optimal behavior of agents. We showed that gIRL with an LMDP suffers from a non-identifiable problem and, in response, we proposed a variational Bayesian gIRL algorithm with SP-LMDPs. The result of our experiment shows the effectiveness of the proposed

[4] This figure is drawn by QGIS using the data interpolation plugin.

method. Since the application area of our method is not limited to traffic data, we plan a further investigation into practical applications. We also consider that analyzing the theoretical performance constitutes important future research.

References

1. Abbeel, P., Coates, A., Quigley, M., Ng, A.Y.: An application of reinforcement learning to aerobatic helicopter flight. In: NIPS. pp. 1–8 (2007)
2. Ziebart, B.D., Maas, A.L., Bagnell, J.A., Dey, A.K.: Maximum entropy inverse reinforcement learning. In: AAAI. pp. 1433–1438 (2008)
3. Song, X., Zhang, Q., Sekimoto, Y., Shibasaki, R.: Intelligent system for urban emergency management during large-scale disaster. In: AAAI. pp. 458–464 (2014)
4. Neu, G., Szepesvári, C.: Training parsers by inverse reinforcement learning. Machine learning 77(2-3), 303–337 (2009)
5. Todorov, E.: Linearly-solvable markov decision problems. In: NIPS. pp. 1369–1376 (2006)
6. Dvijotham, K., Todorov, E.: Inverse optimal control with linearly-solvable MDPs. In: ICML. pp. 335–342 (2010)
7. Ng, A.Y., Russell, S.: Algorithms for inverse reinforcement learning. In: ICML. pp. 663–670 (2000)
8. Makino, T., Takeuchi, J.: Apprenticeship learning for model parameters of partially observable environments. In: ICML. pp. 1495–1502 (2012)
9. Ramachandran, D., Amir, E.: Bayesian inverse reinforcement learning. In: IJCAI. pp. 2586–2591 (2007)
10. Rothkopf, C.A., Dimitrakakis, C.: Preference elicitation and inverse reinforcement learning. In: ECML PKDD. pp. 34–48 (2011)
11. Lazaric, A., Ghavamzadeh, M.: Bayesian multi-task reinforcement learning. In: ICML. pp. 599–606 (2010)
12. Babes, M., Marivate, V., Subramanian, K., Littman, M.L.: Apprenticeship learning about multiple intentions. In: ICML. pp. 897–904 (2011)
13. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming (Wiley Series in Probability and Statistics) (2005)
14. Van der Vaart, A.W.: Asymptotic statistics. Cambridge University Press (2000)
15. Bishop, C.M., Svenskn, M.: Bayesian hierarchical mixtures of experts. In: UAI. pp. 57–64 (2002)
16. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. Machine learning 37(2), 183–233 (1999)
17. Jaakkola, T., Jordan, M.I.: A variational approach to Bayesian logistic regression models and their extensions. In: AISTATS (1997)
18. Blei, D.M., Lafferty, J.D.: A correlated topic model of science. The Annals of Applied Statistics pp. 17–35 (2007)
19. Böhning, D.: Multinomial logistic regression algorithm. Annals of the Institute of Statistical Mathematics 44(1), 197–200 (1992)
20. Bouchard, G.: Efficient bounds for the softmax function and applications to approximate inference in hybrid models. In: NIPS 2007 workshop for approximate Bayesian inference in continuous/hybrid systems (2007)
21. Jebara, T., Choromanska, A.: Majorization for CRFs and latent likelihoods. In: NIPS. pp. 557–565 (2012)
22. Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., Huang, Y.: T-drive: driving directions based on taxi trajectories. In: SIGSPATIAL. pp. 99–108 (2010)