Inverse Optimal Control from Demonstrations with Mixed Qualities

Kyungjae Lee, Yunho Choi, and Songhwai Oh

***Abstract*— This paper proposes an inverse optimal control (IOC) framework which incorporates demonstrations with mixed qualities. The proposed method utilizes the benefits of sub-optimal demonstrations which can provide information about *what not to do* and supplies training data near states unvisited by optimal demonstrations. The main idea of the proposed method is to find the value function which satisfies the optimality condition over optimal demonstrations and violates it over sub-optimal demonstrations. We conduct experiments on three environments and empirically show that the proposed method outperforms the original IOC algorithm, which uses only optimal demonstrations.**

I. Introduction

Inverse optimal control (IOC) has enabled for a robot to perform complex tasks, such as inverted helicopter flight [1], autonomous driving [2], [3], and manipulation [4]. An IOC framework learns both reward and policy function from expert's demonstrations. In general, IOC aims to recover a reward function which can generate expert's demonstrations.

Since most existing IOC algorithms utilize only optimal demonstrations which are often located near the high reward regions, the resulting reward function has no capability to estimate the low reward regions correctly. This issue is highly investigated in several studies [5]–[7]. In [5], the authors argued that insufficient demonstrations near the low reward region will cause a catastrophic failure. To handle this issue, [6] and [7] utilize failed and sub-optimal demonstrations, respectively. In [6], the method of incorporating optimal demonstrations and failed demonstrations is proposed by adding constraints that make the resulting behavior maximally different from the failed demonstrations. In [7], Lee et al. incorporate both optimal and sub-optimal demonstrations where the algorithm assigns a high reward near the optimal demonstrations and a low reward near the sub-optimal demonstrations. The sub-optimal demonstrations give an information about *what not to do* near low reward regions. While [6] and [7] have successfully utilized sub-optimal demonstrations for an IOC problem, [6] and [7] require to solve the reinforcement learning (RL) problem as a subroutine to check whether expert's demonstrations become an optimal solution or not.

To handle both sample inefficiency and the lack of demonstrations near the low reward regions, we employ the necessary and sufficient condition of Hamilton-Jacobi-Bellman (HJB) equation. The HJB equation is a well known partial differential equation where the solution is the maximum cumulative reward, also knwon as the optimal value. By assuming control affine dynamics and a control quadratic reward function, a closed-loop optimal policy can be obtained from HJB, i.e., $u = R^{-1}G(x)^T\nabla_x V(x)$. We observe that, by using this optimal policy, the gradient information of the value function can be extracted from expert's actions. The main idea of the proposed method is to find the value function which satisfies the optimal policy for expert's optimal demonstrations and violates its optimality for sub-optimal demonstrations. In this regards, a value function is directly learned from optimal and sub-optimal control data and the resulting policy can be obtained from HJB. Hence, our method avoids solving optimal control problem at every iteration while incorporating both optimal and sub-optimal demonstrations.

K. Lee, Y. Choi, and S. Oh are with the Department of Electrical and Computer Engineering and ASRI, Seoul National University, Seoul 08826, Korea (e-mail: {kyungjae.lee, yunho.choi}@cpslab.snu.ac.kr, songhwai@snu.ac.kr).

II. Inverse Optimal Control from Demonstrations with Mixed Qualities

In this section, we propose a novel inverse optimal control from demonstrations with mixed qualities (IOCfDMQ) which has a capability to incorporate both optimal and sub-optimal demonstrations. In our problem, it is assumed that the positive demonstrations are a set of sequences that consists of state and optimal control pairs $\mathcal{D}^+$ and the sub-optimal demonstrations consists of a sequence of state and sub-optimal control pairs, i.e., $\mathcal{D}^- = \{(x_{i,t}, u_{i,t})_{t=0}^{T_i}\}_{i=0}^{N^-}$, where N^- is the number of sub-optimal demonstrations, $+$ and $-$ indicate optimal and sub-optimal data, respectively. We consider the problem of finding the value function to generates given positive demonstrations and not to create negative demonstrations. The proposed IOC framework can be formulated as follows:

$$\underset{\theta}{\text{minimize}} \quad \mathcal{L}^+(\mathcal{D}^+, \hat{V}_\theta) + \lambda^- \mathcal{L}^-(\mathcal{D}^-, \hat{V}_\theta, \epsilon) + \lambda B(\hat{V}_\theta), \tag{1}$$

where θ is parameters of a value estimator $\hat{V}_\theta$, $B(\hat{V}_\theta)$ is a regularization function, $\lambda > 0$ is a regularization coefficient, $\mathcal{L}(\mathcal{D}, \hat{V}) \triangleq \sum_{i=0}^{N}\sum_{t=0}^{T} ||u_{i,t} - R^{-1}G(x_{i,t})^T\nabla_x\hat{V}(x_{i,t})||_2^2$, and $\mathcal{L}^-(\mathcal{D}^-, \hat{V}_\theta, \epsilon) \triangleq \frac{1}{2}\sum_{i=0}^{N^-}\sum_{t=0}^{T}\left[\epsilon - ||u_{i,t} - R^{-1}G(x_{i,t})^T\nabla_x\hat{V}_\theta(x_{i,t})||_2^2\right]_+$, where $G(x)$ is a control affine dynamics, i.e., $\dot{x} = F(x) + G(x)u$. Unlike to existing IOC algorithms, we newly add $\mathcal{L}^-$ that requires the Euclidean distance between the control from sub-optimal demonstrations and the estimated policy to be greater than the threshold ϵ. Note that $\mathcal{L}^-$ becomes zero when the distance between the estimated policy and the sub-optimal control is below the predefined threshold. In other words, minimizing $\mathcal{L}^-$ ensures that the gradient of the estimated value function

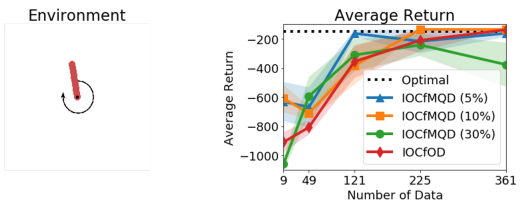

Fig. 1: An illustration of the pendulum environment and average returns of different algorithms. The number in the parentheses shows the ratio of the number of sub-optimal demonstrations to the total number of demonstrations.

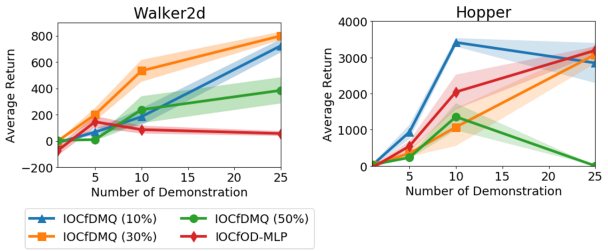

Fig. 2: Average returns of the trained policy on the hopper and walker2d problems.

does not align in the direction similar to the sub-optimal control. In this regards, the estimated value function does not increase near $\mathcal{D}^-$. Conversely, since minimizing $\mathcal{L}^+$ is equivalent to align the gradient of the value estimator to the direction of the optimal control, the estimated value function increases near $\mathcal{D}^+$. To obtain a smooth estimation result, we minimize the output of the value function using the following regularization: $B(\hat{V}_\theta) \triangleq \sum_{x\in\mathcal{D}} ||\hat{V}_\theta(x)||_2^2$, where $\mathcal{D}$ is entire training demonstrations. Since we only utilize the gradient information of the value function, the constant of the indefinite integral cannot be determined in our optimization. In this regards, the regularization helps to handle such illposedness and penalizes the scale of the output value to increase.

While the proposed method is well defined when control affine dynamics are known, in many complex problems, control affine dynamics are often unavailable. In this case, we learn a control affine dynamics using a neural network, which is called a dynamic network. The network output consists of three components: passive dynamics $F_\nu(x)$, control dynamics $G_\nu(x)$ and variance $\Sigma_\nu(x)$ of dynamics, where $\Sigma_\nu(x)$ is a diagonal matrix and ν is the parameter of a dynamic network.

III. Experiments and Conclusion

To validate the effectiveness of sub-optimal demonstrations, we prepare two experiments: a model based problem and a model free problem. In the model based problem, we demonstrate that the proposed method has a capability of recovering the underlying optimal policy. In the model free problem, we first verify that sub-optimal demonstrations help to train a dynamic network in terms of the generalization performance and compare the proposed method with the method using only optimal demonstrations [8]. We model the value function using a multi-layered perceptron where a *tanh* and *cosh* are used for the activation functions. In the model free problem, the dynamic network is also modeled as a perceptron. Both the proposed and the compared method use the same network architecture for the entire experiments.

A. Model Based Problem

We compare the proposed method to the method using only optimal demonstrations. We call the latter an inverse optimal control from optimal demonstrations (IOCfOD) [8]. We test all algorithms on the pendulum problem as shown in the left figure in Figure 1. The optimal demonstrations are generated by controlling the agent with the optimal policy which is obtained by solving the HJB equation. The sub-optimal demonstrations are generated by flipping the optimal control, where the flipped control is clearly sub-optimal. We prepare three different sets of demonstrations by mixing optimal and sub-optimal demonstrations with three different ratios: (95%, 5%), (90%, 10%), and (70%, 30%) and measure the performance while increasing the number of given state action pairs.

The results are shown in Figure 1. The proposed methods with ratio 5% and 10% outperform IOCfOD which uses only optimal demonstrations in terms of the number of demonstrations required to achieve the expert's performance. Especially, the proposed method with 5% sub-optimal demonstrations achieves the expert's performance using the smallest number of data.

B. Model Free Problem

In this experiment, we demonstrate that the proposed method can be applied to model free problems. The simulations are conducted in the MuJoCo simulator [9], which is a physics-based simulator, using two environments with unknown model dynamics: *Walker2d* and *Hopper*.

We apply both IOCfDMQ and IOCfOD to different numbers of mixed demonstrations with different ratios: $10\%, 30\%$, and 50% and measure the average returns over 100 consecutive episodes using the resulting policy. The five simulations are conducted with different random seeds. The results are shown in Figure 2. In both problems, the proposed method outperforms IOCfOD. IOCfDMQ with 10% sub-optimal demonstrations show the best performance and the IOCfDMQs with other ratios are worse than IOCfOD in the hopper problem. In the walker2d problem, IOCfDMQ generally outperforms IOCfOD. The first reason why IOCfDMQ shows better performance than IOCfOD is that the dynamic model trained with the demonstrations with mixed quality shows better generalization performance. Furthermore, this result supports that mixing optimal and sub-optimal demonstrations has benefits over using only optimal demonstrations. One interesting observation is that IOCfDMQ with 30% sub-optimal demonstrations shows poor performance than IOCfOD in both problems. We believe that this performance drop is caused due to the insufficient number of optimal demonstrations. In other words, if the number of optimal demonstrations is not enough, then using sub-optimal data has a less benefit.

REFERENCES

[1] P. Abbeel, A. Coates, and A. Y. Ng, "Autonomous helicopter aerobatics through apprenticeship learning," *The International Journal of Robotics Research*, vol. 29, no. 13, pp. 1608–1639, 2010.

[2] D. Vasquez, B. Okal, and K. Arras, "Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison," in *Proc. of the International Conference on Intelligent Robots and Systems*, 2014, pp. 1341–1346.

[3] B. Okal and K. O. Arras, "Learning socially normative robot navigation behaviors with bayesian inverse reinforcement learning," in *Proc. of the International Conference on Robotics and Automation*. IEEE, 2016, pp. 2889–2895.

[4] C. Finn, S. Levine, and P. Abbeel, "Guided cost learning: Deep inverse optimal control via policy optimization," in *Proc. of the International Conference on Machine Learning*, Jun 2016, pp. 49–58.

[5] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proc. of the international conference on artificial intelligence and statistics*, Apr 2011, pp. 627–635.

[6] K. Shiarlis, J. V. Messias, and S. Whiteson, "Inverse reinforcement learning from failure," in *Proc. of the International Conference on Autonomous Agents & Multiagent Systems*, May 2016, pp. 1060–1068.

[7] K. Lee, S. Choi, and S. Oh, "Inverse reinforcement learning with leveraged gaussian processes," in *Proc. of the International Conference on Intelligent Robots and Systems*, Oct 2016, pp. 3907–3912.

[8] W. Li, E. Todorov, and D. Liu, "Inverse optimality design for biological movement systems," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 9662–9667, 2011.

[9] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *Proc. of the International Conference on Intelligent Robots and Systems*, Oct 2012.