Supplementary Material: Continuous Space Imitation Learning with Density Matching Policy Learning

Sungjoon Choi, Kyungjae Lee, Andy H. Park, and Songhwai Oh

I. Analysis

In this supplementary material, we present full proofs of Theorem 1, 2, and 3. The proposed density matching policy learning (DMPL) first estimate the proximal reward function and used the optimized reward function for modeling the policy function. The proximal reward function can be estimated from the following optimization problem:

$$\begin{aligned} \underset{R}{\text{maximize}} \quad & V(R) = \langle \hat{\mu}, R \rangle \\ \text{subject to} \quad & \|R\|_2 \leq 1, \end{aligned} \tag{1}$$

where the norm ball constraint $\|R\|_2 \leq 1$ is introduced to handle the scale ambiguity of the reward function [1] and $\langle \hat{\mu}, R \rangle = \int_{\mathcal{S}\times\mathcal{A}} \hat{\mu}(s,a) R(s,a)\, ds\, da$. Once the proximal reward $\hat{R}$ is achieved, the policy function is achieved by model predictive control with respect to the logarithm of the reward $\hat{R}$, i.e.,

$$\hat{\pi}(a|s) = \underset{\substack{a=a_1,\ldots,a_T,\, s_1=s \\ s_{t+1}=f(s_t,\, a_t)}}{\operatorname{argmax}} \sum_{t=1}^{T} \gamma^{t-1} \log \hat{R}(s_t,\, a_t) \tag{2}$$

where $s_{t+1} = f(s_t, a_t)$ is a dynamic model and T is the time horizon.

Theorem 1: Given exact occupancy measure of an expert μ_E, the policy function $\hat{\pi}$ achieved from (2) with any $T \geq 1$ recovers the original policy function of an experts.

Proof: We first consider the occupancy measure and policy function of an expert is derived from the infinite horizon discounted setting of Markov decision process (MDP) which is widely used in both reinforcement learning and inverse reinforcement learning problems. The MDP problem is formulated as follows:

$$\begin{aligned} \underset{\pi}{\text{maximize}} \quad & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t,\, a_t) \middle| \pi, d, T \right] \\ \text{subject to} \quad & \forall\, s \quad \sum_{a'} \pi(a'|s) = 1, \\ & \forall\, s, a \quad \pi(a'|s) \geq 0 \end{aligned} \tag{3}$$

where $r(\cdot)$ is a reward function, γ is a discounted factor, $\pi(a|s)$ is a policy function, $d(s)$ is an initial state distribution, and $T(s', a, a) = p(s'|s, a)$ is a state transition probability. We assume both policy and occupancy measure of an expert are derived from (3).

S. Choi, K. Lee, and S. Oh are with the Department of Electrical and Computer Engineering and ASRI, Seoul National University, Seoul 08826, Korea (e-mail: {sungjoon.choi, kyungjae.lee, songhwai.oh}@cpslab.snu.ac.kr). H. Park is with Rethink Robotics, Boston, MA, USA (e-mail: apark@rethinkrobotics.com).

The optimal solution of (3) can be derived from solving Karush-Kuhn-Tucker conditions:

$$\pi(a|s) = \underset{a}{\operatorname{argmax}}\ Q(s,\, a) \tag{4}$$

where the action value function $Q(s,\, a)$ is defined as

$$Q_\pi(s,a) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s, a_0 = a \right]. \tag{5}$$

(4) tells us that given state s the optimal action a is uniquely determined, i.e., a deterministic policy.

The original MDP formulation in (3) can be reformulated as an optimization problem of finding occupancy measure $\rho(s,a)$ which is often referred to as a dual problem of an MDP [2]:

$$\begin{aligned} \underset{\rho}{\text{maximize}} \quad & \sum_{s,\, a} r(s,\, a)\rho(s,\, a) \\ \text{subject to} \quad & \forall\, s \quad \sum_{a'} \rho(s,\, a') = d(s) + \sum_{s',\, a'} \rho(s',\, a') T(s',\, a,\, s) \\ & \forall\, s, a \quad \rho(s,\, a) \geq 0 \end{aligned} \tag{6}$$

where the occupancy measure $\rho(s,\, a)$ and policy $\pi(a|s)$ have one-to-one correspondence [2]:

$$\pi(a|s) = \frac{\rho(s,\, a)}{\sum_{a'} \rho(s,\, a')}. \tag{7}$$

Combining (7) and the fact that the optimal policy function is deterministic from (4), one can easily obtain

$$\pi(a|s) = \underset{a}{\operatorname{argmax}}\ \rho(s,\, a) \tag{8}$$

which clearly indicates that given true occupancy measure $\rho(s,\, a)$ one can recover the expert's policy function by the greedy selection of an action a given a state s to maximize $\rho(s,\, a)$.

Once a true occupancy measure ρ_E of an expert is given, the resulting density measure $\hat{\rho}$ from solving (1) is proportional to ρ_E, i.e., $\hat{\rho} = k\rho_E$ where $k = \frac{1}{\|\rho_E\|_2}$. This can be easily shown with the Cauchy-Schwarz inequality states that $|\langle a,\, b \rangle| \leq \|a\|_2 \|b\|_2$ where the equality holds if and only if a is a scalar multiple of b.

Finally, as a greedy selection of an action will recover optimal policy ρ_E, the model predictive control in (2) will also recover the optimal policy. This can be easily shown with the proof by contradiction. Suppose that (2) is not the optimal policy functions, i.e., given a state s_1 there exists an action a_1' that maximizes $\rho_E(s_1, a_1')$ and an action a_1 that maximizes $\sum_{t=1}^{T} \gamma^{t-1} \log(\rho(s_t,\, a_t))$ where $a_1' \neq a_1$.

However, given a state s_1 the optimal occupancy measure is only nonzero for a unique a_1, i.e., $a_1 = \underset{a}{\operatorname{argmax}}\, \rho_E(s_1, a)$, and zero otherwise. As $\log(0)$ is minus infinity, a_1' should be identical to a_1 which completes the proof. ∎

Theorem 2: Let $\widehat{R}$ be computed from $\widehat{R} = \arg\max_R \langle \hat{\mu}, R\rangle$ s.t. $\|R\|_2 \leq 1$ and $\bar{R}$ be computed from $\bar{R} = \arg\max_R \langle \bar{\mu}, R\rangle$ s.t. $\|R\|_2 \leq 1$, where $\hat{\mu}$ and $\bar{\mu}$ are estimated and true densities, respectively. Then

$$|\langle \bar{\mu}, \widehat{R}\rangle - \langle \bar{\mu}, \bar{R}\rangle| \leq 3(R_{max} - R_{min}) d_{var}(\bar{\mu}, \hat{\mu})$$

where R_{max} and R_{min} are the largest and smallest reward values, respectively, and $d_{var}(\cdot, \cdot)$ is the variational distance[1] between two probability distributions.

Theorem 3: Suppose Assumption 1–8 in Appendix I hold. Let N and n be the number of samples and the number of basis functions for kernel density estimation, respectively. Then, for any $\delta \geq 0$ and N and n be sufficiently large, we have

$$|\langle \bar{\mu}, \widehat{R}\rangle - \langle \bar{\mu}, \bar{R}\rangle| \leq 3\sqrt{2}(R_{max} - R_{min}) \times \sqrt{O\left(\frac{1}{N}\right) + O\left(\frac{Nd}{n}(1 + \frac{1}{\sqrt{\delta}})\right)} \tag{9}$$

with probability at least $(1 - \delta)$.

Before proving Theorem 2 and 3, let us first introduce useful lemmas.

Lemma 1: [3] Let P and Q be probability distributions over $\mathcal{X}$ and f be a bounded function on $\mathcal{X}$. Then,

$$|\langle P, f\rangle - \langle Q, f\rangle| \leq (\sup f - \inf f) d_{var}(P, Q).$$

Proof:

$$\begin{aligned}
\langle P,\, f\rangle - \langle Q,\, f\rangle &= \langle P - Q,\, f - \inf f\rangle + \langle P - Q,\, \inf f\rangle \\
&= \sum_{x: P(x) > Q(x)} (f(x) - \inf f)(P(x) - Q(x)) + \\
&\quad \sum_{x: P(x) \leq Q(x)} (f(x) - \inf f)(P(x) - Q(x)) \\
&\leq \sum_{x: P(x) > Q(x)} (f(x) - \inf f)(P(x) - Q(x)) \\
&\leq (\sup f - \inf f) d_{var}(P, Q)
\end{aligned}$$

Since $d_{var}(P,\, Q) = d_{var}(Q,\, P)$ this completes the proof. ∎

Lemma 2: [4] Suppose P and Q are probability distributions. Then,

$$d_{var}(P,\, Q) \leq \sqrt{2} d_{\mathcal{H}}(P,\, Q),$$

where $d_{\mathcal{H}}(\cdot)$ is the Hellinger distance.[2]

[1]The variational distance between two probability distributions, P and Q, is defined as $d_{var}(P, Q) = \frac{1}{2}\sum_{x \in \mathcal{X}} |P(x) - Q(x)|$. It is also known that $d_{var}(P, Q) = \sum_{x \in \mathcal{X}: P(x) > Q(x)} |P(x) - Q(x)|$.

[2]The Hellinger distance between two probability distributions P and Q is defined as $d^2_{\mathcal{H}}(P, Q) = \frac{1}{2}\sum_{x \in \mathcal{X}} (\sqrt{P(x)} - \sqrt{Q(x)})^2$

Proof:

$$\begin{aligned}
|d_{var}(P, Q)|^2 &= \frac{1}{4}\left(\sum_{x \in \mathcal{X}} |P(x) - Q(x)|\right)^2 \\
&\leq \frac{1}{4}\left(\sum_{x \in \mathcal{X}} (\sqrt{P(x)} - \sqrt{Q(x)})^2\right) \times \\
&\qquad \left(\sum_{x \in \mathcal{X}} (\sqrt{P(x)} + \sqrt{Q(x)})^2\right) \\
&= \frac{1}{2} d^2_H(P, Q)\left(2 + 2\sum_{x \in \mathcal{X}} \sqrt{P(x)Q(x)}\right) \\
&= d^2_H(P, Q)\left(2 - d^2_H(P, Q)\right) \\
&\leq 2 d^2_H(P, Q)
\end{aligned}$$

∎

Lemma 3: [5] Suppose $\|\cdot\|$ is a proper norm defined on index set $\mathcal{X}$. Then, for any $a, b \in \mathcal{X}$,

$$|\|a\| - \|b\|| \leq \|a - b\|.$$

Proof: The Lemma 3 is often referred to as a reverse triangle inequality. As a proper norm function satisfies triangle inequality, following two equations are correct:

$$\|a\| + \|b - a\| \geq \|b\| \tag{10}$$

$$\|b\| + \|a - b\| \geq \|a\| \tag{11}$$

Rearranging (10) and (11), we get

$$\|b - a\| \geq \|b\| - \|a\| \tag{12}$$

$$\|a - b\| \geq \|a\| - \|b\| \tag{13}$$

As we $\|a - b\| = \|b - a\|$, combining (12) and (13), we get the reverse triangle inequality. ∎

We also introduce a theorem from [6], which is used to prove Theorem 3.

Theorem 4 ([6]): Suppose Assumption 1–8 in Appendix I hold. Let N and n be the number of samples and the number of basis functions for kernel density esimation, respectively, and $\hat{f}_{N,n}$ be the estimated density function from kernel density estimation with N samples and n basis functions. Then, for any $\delta \geq 0$ and N and n sufficiently large, we have

$$d^2_{\mathcal{H}}(f, \hat{f}_{N,n}) \leq O\left(\frac{1}{N}\right) + O\left(\frac{Nd}{n}(1 + \frac{1}{\sqrt{\delta}})\right),$$

where $d^2_{\mathcal{H}}(\cdot)$ is a squared Hellinger distance between two probability distributions.

We are now ready to prove Theorem 2 and Theorem 3.

Proof: (Theorem 2) Let $\bar{\mu}$ and $\hat{\mu}$ be the true and estimated density functions, respectively, and $\bar{R}$ and $\widehat{R}$ be the reward functions estimated by DMRL with $\bar{\mu}$ and $\hat{\mu}$, respectively. The absolute difference between the optimal

value $\bar{V}$ and the estimated value $\widehat{V}$ can be expressed as:

$$\begin{aligned}|\langle\bar{\mu},\widehat{R}\rangle - \langle\bar{\mu},\bar{R}\rangle| &= |\langle\bar{\mu},\widehat{R}\rangle - \langle\bar{\mu},\bar{R}\rangle + \langle\hat{\mu},\widehat{R}\rangle - \langle\hat{\mu},\widehat{R}\rangle| \\ &\leq |\langle\bar{\mu},\widehat{R}\rangle - \langle\hat{\mu},\widehat{R}\rangle| + |\langle\bar{\mu},\bar{R}\rangle - \langle\hat{\mu},\widehat{R}\rangle| \\ &\leq (R_{max} - R_{min}) d_{var}(\bar{\mu},\,\hat{\mu}) \\ &\quad + |\langle\bar{\mu},\,\bar{R}\rangle - \langle\hat{\mu},\,\widehat{R}\rangle|,\end{aligned}$$

where we used the triangular inequality and Lemma 1.

$|\langle\bar{\mu},\,\bar{R}\rangle - \langle\hat{\mu},\,\widehat{R}\rangle|$ can be further bounded by

$$\begin{aligned}|\langle\bar{\mu},\bar{R}\rangle - \langle\hat{\mu},\widehat{R}\rangle| &= \left|\max_{\|R\|\leq 1}\langle\bar{\mu},R\rangle - \max_{\|R\|\leq 1}\langle\hat{\mu},R\rangle\right| \\ &\leq \max_{\|R\|\leq 1}\langle\bar{\mu}-\hat{\mu},R\rangle \\ &\leq (R_{max} - R_{min}) \sum_{x\in\mathcal{X}} |\mu(x) - \hat{\mu}(x)| \\ &= 2(R_{max} - R_{min}) d_{var}(\bar{\mu},\hat{\mu}),\end{aligned}$$

where we used the definition of optimal and estimated values, Lemma 3 for a dual norm, the definition of an inner product, and the definition of $d_{var}(\cdot)$. ■

Proof: (Theorem 3) Combining aforementioned two inequalities and Lemma 2, we get

$$\begin{aligned}|\langle\bar{\mu},\widehat{R}\rangle - \langle\bar{\mu},\bar{R}\rangle| &\leq 3(R_{max} - R_{min}) d_{var}(\bar{\mu},\,\hat{\mu}) \\ &\leq 3\sqrt{2}(R_{max} - R_{min})\times \\ &\quad \sqrt{O\left(\frac{1}{N}\right) + O\left(\frac{Nd}{n}(1+\frac{1}{\sqrt{\delta}})\right)}.\end{aligned}$$

■

II. Implementation Details

A. Constructing MDPs for Track Driving Experiments

In this section, we introduce implementation details that have been used in track driving experiments. For the compared inverse reinforcement learning algorithms, The state and action consist of the pose (x, y, and θ) of a car following a unicycle dynamic model [7] and directional and angular velocities (v, w), respectively. The demonstrations are collected from three-lane tracks with random car configurations as shown in Figure 1. Both state and action spaces are discretized for model based IRL methods by dividing the state space into 720 states, 6 for vertical axis, 20 for horizontal axis, and 6 for heading, and the action space is divided into 12 actions, 4 for directional and 3 for angular velocities. We also tried with more fine-grained MDPs, however, it only deteriorates the reward inference performance. We believe it is mainly due to the curse of dimensionality.

The transition probability $T(\mathbf{s}'|\mathbf{a},\mathbf{s})$ is defined by a three-dimensional tensor $T \in \mathbb{R}^{720\times 12\times 720}$ where each element $T(\mathbf{s}_i,\mathbf{a}_j,\mathbf{s}_k)$ is proportional to $\mathcal{N}(\mathbf{s}_i|\mathbf{f}(\mathbf{s}_k,\mathbf{a}_j),\Sigma)$ where $f(\mathbf{s},\mathbf{a})$ is a vector-valued function whose output is the next position starting from the state $\mathbf{s}$ with the action $\mathbf{a}$ following a unicycle dynamic model and Σ controls the stochasticity. Σ is $\text{diag}\left([(\frac{x_{max}-x_{min}}{x_{res}})^2, (\frac{y_{max}-y_{min}}{y_{res}})^2, (\frac{\theta_{max}-\theta_{min}}{\theta_{res}})^2]/3\right)$

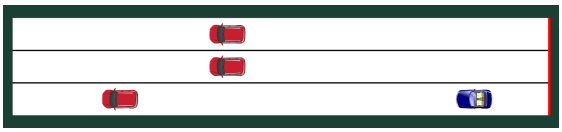


Fig. 1: The track configuration where the demonstrations are manually collected. The car to be controlled is depicted with a blue convertible and other cars are shown as red vans.

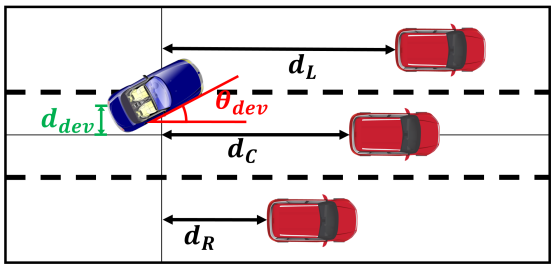


Fig. 2: Descriptions of features used in track driving experiments.

where 3 is manually selected. We found that appropriate stochasticity stabilizes the MDP learning process.

We used six dimensional features to represent the reward function: (1) the lane deviation distance, $dist_{dev}$, (2) the lane deviation degree, θ_{dev}, the closest frontal distance to another car in the (3) left lane, $dist_L$, (4) center lane, $dist_C$, and (5) right lane, $dist_R$, and (6) the directional velocity, v, as depicted in Figure 2. The mapping between the dicretized state-action space and the feature space is defined by a three-dimensional tensor $F \in \mathbb{R}^{720\times 12\times 6}$ where $[F]^k_{i,j}$ indicates k-th feature of i-th state and j-th action.

The constructed MDP is used for training of MaxEnt and GPIRL. In the other hand, RelEnt is a model-free IRL method that does not require MDP. However, RelEnt requires an additional forward-path stage of sampling demonstrations from the baseline policy during the optimization process. The required base line demonstrations are generated by applying random control to the unicycle model.

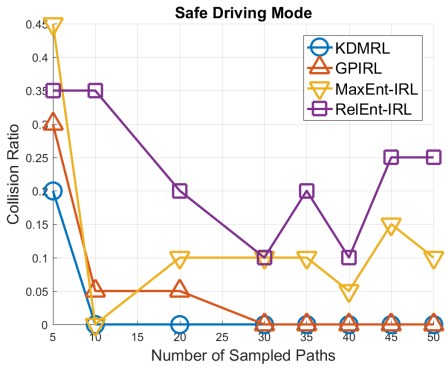
B. Hyperparameters of KDMRL

The proposed KDMRL has two hyperparameters, β and λ. These hyperparameters are optimize with a simple grid search on log scale to maximize the average variational distances. The resulting values of β and λ are 0.1 and 1.0, respectively.

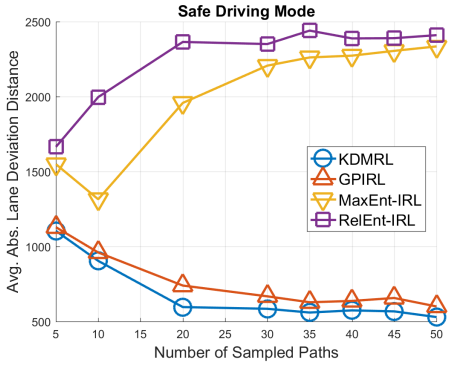
C. Sample-based Random Steer Controller

In the track driving experiments, a simple sample-based random steering method [8] is used to control the car. Specifically, we sample 37 trajectories with 0.8 second long by randomly sampling angular velocities from $[-100, +100]$ deg/s every 0.4 second. The score of each trajectory is evaluated by the sum of traversed rewards and the final control is obtained from the initial control of the trajectory with the highest score.

In this subsection, we vary the number of sampled trajectories and tested the performance of a sample-based controller



(a)



(b)

Fig. 3: (a) Collision rates and (b) average absolute lane divation distances of each reward learning algorithm in safe driving mode as a function of the number of sampled paths.

in safe driving mode. Collision rates and average absolute lane deviation distances as a function of the number of sampled trajectories are shown in Figure 3. Both collision rates and average absolute lane deviation distances are computed from 20 independent run on a large track with five lanes with six randomly deployed static cars. One can see that the performance of sample-based controller saturates when the number of samples exceeds 30.

APPENDIX I

We list assumptions required by Theorem 3. Note that assumption 1 to 7 are the same assumptions used in [6], but we list them here for the sake of completeness. We note that these are widely-used assumptions for kernel density estimation [9]. Assumption 8 is newly added.

Assumption 1: The random variables $\{x_i\}_i^N$ are sampled independent and identically distributed according to $f(x)$.

Assumption 2: The density estimator function f_n^θ is piecewise continuous for each parameter θ.

Assumption 3: (a) $E[\log f(x)]$ exists and $|\log f_n^\theta(x)| \leq m(x)$ $\forall \theta \in \Theta$ for some $m(x)$ where $m(x)$ is an integrable function with respect to f. (b) $E[\log(f/f_n^\theta)]$ has a unique minimum.

Assumption 4: $\frac{\partial \log f_n^\theta(x)}{\partial \theta}$ is integrable with respect to x and continuously differentiable.

Assumption 5: $|\frac{\partial^2 \log f_n^\theta(x)}{\partial \theta_i \partial \theta_j}|$ and $|\frac{\partial f_n^\theta(x)}{\partial \theta_i}\frac{\partial f_n^\theta(x)}{\partial \theta_j}|$ is dominated by functions integrable with respect to f $\forall x \in \mathcal{X}$, $\theta \in \Theta$.

Assumption 6: (a) θ^* is interior to Θ. (b) $B(\theta^*)$ is non-singular. (c) θ^* is a regular point of $A(\theta)$.

Assumption 7: The convex model $f_n^\theta = \mathcal{G}_n$ obeys η positivity requirement.

Assumption 8: The target density function $f \in \{\sum_{i=1}^\infty \alpha_i \phi_\rho(\cdot;\theta_i)\}$, where $\phi_\rho(\cdot;\theta_i)$ is a basis density function.

REFERENCES

[1] A. Y. Ng and S. Russell, "Algorithms for inverse reinforcement learning." in *Proc. of the 17th International Conference on Machine Learning*, June 2000.

[2] U. Syed, M. Bowling, and R. E. Schapire, "Apprenticeship learning using linear programming," in *Proceedings of the international conference on Machine learningProceedings of the international conference on Machine learning*. ACM, 2008, pp. 1032–1039.

[3] W. R. Gilks, *Markov chain monte carlo*. Wiley Online Library, 2005.

[4] D. Pollard, "Asymptopia," *Unpublished book*, 2000.

[5] P. Alfeld, "Understanding mathematics," *Utah: Departemen of Mathematics. University of Utah.(Online: April 20014) Tersedia: http:/www math utah edu/-alfeld/math html.(Tanggal:)*, 2004.

[6] A. J. Zeevi and R. Meir, "Density estimation through convex combinations of densities: approximation and estimation bounds," *Neural Networks*, vol. 10, no. 1, pp. 99–109, 1997.

[7] T.-C. Lee, K.-T. Song, C.-H. Lee, and C.-C. Teng, "Tracking control of unicycle-modeled mobile robots using a saturation feedback controller," *IEEE transactions on control systems technology*, vol. 9, no. 2, pp. 305–318, 2001.

[8] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 33–55, 2016.

[9] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica: Journal of the Econometric Society*, pp. 1–25, 1982.