Imitation Learning via Multi-Step Occupancy Measure Matching

Minghuan Liu, Hangyu Wang, Yangtian Zhang, Minkai Xu,
Zhengbang Zhu, Weinan Zhang[†]
[*]equal contribution [†]corresponding author
Department of Computer Science
Shanghai Jiao Tong University

Abstract

Imitation learning (IL) aims to learn a policy from expert demonstrations without reward signals. Previous methods such as behavior cloning (BC) work by learning one-step predictions, but seriously suffer from the compounding error problem; recent generative adversarial solution, though alleviates such problems in a discrepancy minimization view, is still limited in only matching single-step state-action distributions instead of long-term trajectories. To address the long-range effect, in this paper, we explore the potential to boost the performance of IL by regularizing the multi-step discrepancies. We first propose the multi-step occupancy measure matching formulation, where we extend the idea of matching single state-action pairs to sequences of multiple steps. Interestingly, theoretical analysis of the proposed multi-step algorithm reveals a trade-off between the rollout discrepancy and the sampling complexity, making it non-trivial to select an appropriate step length T for the practical implementation. Inspired by the recent progress of integrating multi-armed bandits in curriculum learning, we further propose an automated curriculum multi-step occupancy measure matching algorithm named AutoGAIL, which automatically selects the appropriate step length during the training procedure. Compared with various multi-step GAIL baselines, AutoGAIL consistently achieves superior performance with satisfactory learning efficiency given different amount of demonstrations.

1 Introduction

Imitation learning (IL) approaches solve the learning from demonstrations task, where the reward is unknown and the agent can only get access to the expert's demonstrations. Naive solutions, such as behavior cloning (BC) Bain and Sammut [1995], simply treat it as a one-step supervised learning problem. Another kind of solutions, inverse reinforcement learning (IRL) Abbeel and Ng [2004], tries to first estimate the reward from the demonstrations and then train an online RL agent to induce the optimal policy. Recently, inspired by the progress of generative models, Ho and Ermon [2016] proposes generative adversarial imitation learning (GAIL), which views IL as a discrepancy minimization problem and imitated the expert policy by an occupancy measure (OM) matching procedure.

Perfect imitation corresponds to match the long-term rollout sequence of the expert. Unfortunately, algorithms mentioned above, though practically effective for a range of IL tasks, remain problematic on long-term imitation. The classic BC method heavily relies on the *i.i.d* assumption of each single step to learn the one-step predictions, which omits the long-term sequence matching. Thanks to the online training scheme, IRL and GAIL methods are capable of alleviating the long-term sequence matching problem by learning the policy from the estimated reward function or the discriminator.

Preprint.

However, they are essentially limited in matching the single-step OM, which obstructs the policy to learn to match the long-term trajectory distribution. Though theoretically matching the 1-step OMs is able to match the long-term sequence, in practice this optimal matching cannot be always achieved, where the small 1-step mistakes will eventually lead to an unacceptable large error from the long-step view.

In this paper, to explicitly mitigate such long-term problems in IL, we extend the definition of the state or state-action occupancy from simple single-step to pieces of sequences, and introduce the idea of T-step OM. T-step OM is defined as the distribution of multi-step sequences of states and actions. This concept enables us to derive a novel algorithm where we can alleviate the long-term effect by directly conducting multi-step OM matching (*i.e.*, match the sequence-level OM instead of the step-level OM), named multi-step GAIL (MS-GAIL). Intuitively, compared with the one-step OM discrepancy, the divergence of a multi-step OM can be much more informative and lead to better optimization of the gap between the agent and the expert. Starting from the intuitive idea, we further conduct a theoretical analysis of both rollout discrepancy and sample complexity, and demonstrate that the MS-GAIL algorithm always holds a tighter bound on single-step occupancy discrepancy. Interestingly, the theoretical results also reveal a trade-off between the rollout discrepancy and the sample complexity, *i.e.*, we can further alleviate the long-term effect by minimizing a longer-step OM, but there is no free lunch and we have to pay much more training samples for that. Therefore, it is challenging to determine the best multi-step length given the fixed size of training samples. Inspired by the recent progress of combining multi-armed bandits with curriculum learning, we further propose an automated curriculum measure matching algorithm named AutoGAIL. AutoGAIL provides a flexible framework for multi-step OM matching, which can automatically select the appropriate sequence length to improve the sample efficiency as well as the final performance.

In a nutshell, the contributions of this paper can be summarized as follows:

1. We introduce the idea of T-step OM and propose the practical multi-step OM matching algorithm, *i.e.*, MS-GAIL (Section 3).
2. We analyze the rollout discrepancy and the sample complexity of MS-GAIL, and reveal a non-trivial trade-off between them (Section 4).
3. We further propose an auto-curriculum framework of T-step OM matching algorithm that can automatically choose the sequence length to improve the final performance (Section 5).

Finally, we evaluate MS-GAIL and AutoGAIL with various step length on several continuous control benchmarks. Comprehensive experiments verify the trade-off between the rollout discrepancy and the sample complexity, and demonstrate the potential for improving the imitation performance via multi-step OM matching. Results also show that AutoGAIL successfully handles the challenge on determine the appropriate multi-step length and can always achieve the best performance as well as the sample efficiency.

2 Preliminaries

Notation. A Markov Decision Process (MDP) is defined by a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, M, \rho_0, r, \gamma \rangle$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the action space of the agent, $M : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the environment dynamics, $\rho_0 : \mathcal{S} \to [0, 1]$ is the distribution of the initial state s_0, and $\gamma \in [0, 1]$ is the discounted factor. The agent holds the policy $\pi(a|s) : \mathcal{S} \times \mathcal{A} \to [0, 1]$ to make decisions and receive rewards defined as $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. The objective is to find the optimal policy that maximize the expected sum of the discounted rewards with the entropy at each visited state:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[r(s, a) \right] + \alpha H(\pi) \, , \tag{1}$$

where $H(\pi) \triangleq \mathbb{E}_{\pi}[-\log \pi(a|s)]$ is the γ-discounted casual entropy Bloem and Bambos [2014] and α is the temperature hyperparameter to determine the relative importance of the entropy term. In this work we use the subscript to denote the timestep, *e.g.*, s_t and the superscript is the order in a sequence, *e.g.*, a^t.

Many recent IL methods are built upon the concept of *Occupancy Measure* (OM), which is also the foundation of our approach. Formally, OM is defined as the discounted occurrence probability of states or state-action pairs when the agent interacts with the environment using policy π:

$$\rho_{\pi}(s, a) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s, a_t = a|\pi) = \pi(a|s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi) = \pi(a|s)\rho_{\pi}(s). \tag{2}$$

Note that ρ is unormalized and the normalization can be easily achieved by $d_\pi = (1-\gamma)\rho_\pi$. With such a definition we can write down that $\mathbb{E}_\pi[\cdot] = \sum_{s,a} \rho_\pi(s,a)[\cdot] = \mathbb{E}_{(s,a)\sim\rho_\pi}[\cdot]$.

Imitation Learning as 1-step Occupancy Measure Matching. Here we briefly review the conclusions from Ho and Ermon [2016] and Ghasemipour et al. [2019], which analyzed the connection between the IL problem and the 1-step OM matching problem. These conclusions help us to construct a theoretical analysis of our proposed method.

Proposition 1 (Proposition 3.2 of Ho and Ermon [2016])**.** *Given the definition of RL procedure as Eq. (1) and IRL procedure as* $\mathrm{IRL}_\psi(\pi_E) = \arg\max_r -\psi(r) + (\min_{\pi\in\Pi} -H(\pi) - \mathbb{E}_\pi[r(s,a)]) + \mathbb{E}_{\pi_E}[r(s,a)]$*, we have:*

$$\mathrm{RL}\circ\mathrm{IRL}_\psi(\pi_E) = \arg\min_\pi -H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E}) \tag{3}$$

This proposition indicates that RL with the reward recovered by a ψ-regularized IRL can actually learn a policy whose 1-step OM matches the expert's measured by the convex function ψ^*, *i.e.*, optimizing certain distance metrics of OM between the policy and expert can solve the IL problem.

Proposition 2 (Appendix D of Ghasemipour et al. [2019])**.** *Considering the reward function regularizer as:* $\psi_f(r) = \mathbb{E}_{\rho_{\pi_E}(s,a)}\left[f^*(s,a) + r(s,a)\right]$ *where* f^* *is the convex conjugate of* f*, then we have:*

$$\psi_f^*(\rho_\pi(s,a) - \rho_{\pi_E}(s,a)) = D_f(\rho_\pi(s,a)\|\rho_{\pi_E}(s,a)) \tag{4}$$

$$\mathrm{RL}\circ\mathrm{IRL}_\psi(\pi_E) = \arg\min_\pi -H(\pi) + D_f(\rho_\pi(s,a)\|\rho_{\pi_E}(s,a)) \tag{5}$$

This proposition illustrates that any f-divergence can be used for IL as long as we choose a specific ψ_f. For example, GAIL Ho and Ermon [2016] minimizes the JS divergence $\mathrm{D}_{\mathrm{JS}}(\rho_\pi\|\rho_{\pi_E})$ while AIRL Fu et al. [2017] minimizes the KL divergence $\mathrm{D}_{\mathrm{KL}}(\rho_\pi\|\rho_{\pi_E})$.

3 Imitation Learning as T-step Occupancy Measure Matching

In this section, we first analyze the limitations of single-step discrepancy by two illustrative examples. Then we propose the definition of T-step OM, and further extend the previous 1-step OM matching to T-step OM matching to overcome the shortages of previous methods.

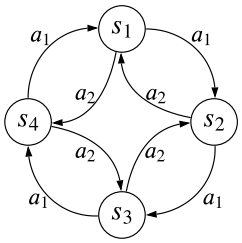
3.1 Limitations of 1-Step Discrepancy

Ambiguity for determining a better policy in a single-step view. The first example is constructed on a simple ring MDP example to show the ambiguity for determining the optimality in a one-step view (Fig. 1(a)). As shown in the table, two behavior policies π_1 and π_2 differ with the expert policy π_E only on a single state, resulting in the same KL divergence of the 1-step OM $\mathrm{D}_{\mathrm{KL}}(P_i(s,a)\|P_E(s,a))$. However, the optimality can be determined in a multi-step view as we should match the long-term sequence. For example, π_1 keeps a smaller 2-step divergence (*i.e.* , $\mathrm{D}_{\mathrm{KL}}(P_i(s^1,a^1,s^2,a^2)\|P_E(s^1,a^1,s^2,a^2))$) and is better than π_2. As the step gets longer, the optimality is more significant.

Weak capacity for measuring single-step discrepancies. This example is constructed on a simple one-dimensional environment, aiming to describe how one-step error spreads to long-step. Specifically, the agent moves along the x-axis from the point 0.5 to the point 10 within an action space [0,1] (shown in Fig. 1(b)). The expert policy is a rectangular window function (blue), and the sub-optimal policies are Blackman window functions with different parameters (orange, red and green). With respect to the expert policy, the sub-optimal policies get worse as the index increases. We analyze the KL divergence ratio $\frac{\mathrm{D}_{\mathrm{KL}}(P_i(s,a)\|P_E(s,a))}{\mathrm{D}_{\mathrm{KL}}(P_1(s,a)\|P_E(s,a))}$, $\frac{\mathrm{D}_{\mathrm{KL}}(P_i(s^1,a^1,s^2,a^2)\|P_E(s^1,a^1,s^2,a^2))}{\mathrm{D}_{\mathrm{KL}}(P_1(s^1,a^1,s^2,a^2)\|P_E(s^1,a^1,s^2,a^2))}$ and $\frac{\mathrm{D}_{\mathrm{KL}}(P_i(s^1,a^1,\cdots,s^3,a^3)\|P_E(s^1,a^1,\cdots,s^3,a^3))}{\mathrm{D}_{\mathrm{KL}}(P_1(s^1,a^1,\cdots,s^3,a^3)\|P_E(s^1,a^1,\cdots,s^3,a^3))}$ for different sub-optimal index i on 1-step, 2-step and 3-step state-action sequence distribution. Apparently, even if the sub-optimalilties for different policies seems similar in a single-step view (the 1-step bars shown in Fig. 1(b) right), the discrepancy can deteriorate much more as the step gets longer (the 2 and 3-step bars shown in Fig. 1(b) right).

[3] The OM shown here is unnormalized, while we use the normalized ones to calculate the divergence.

[4] k-s denotes k-step for short.



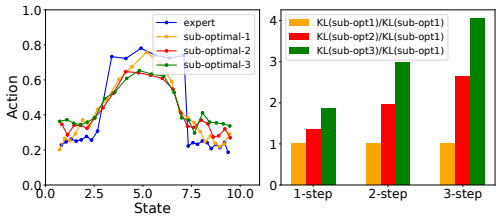
ρ_{π_E}	a_1	a_2
s_1	0.2	0.8
s_2	0.5	0.5
s_3	0.4	0.6
s_4	0.2	0.8

ρ_{π_1}	a_1	a_2
s_1	0.3	0.7
s_2	0.5	0.5
s_3	0.4	0.6
s_4	0.2	0.8

ρ_{π_2}	a_1	a_2
s_1	0.2	0.8
s_2	0.5	0.5
s_3	0.4	0.6
s_4	0.3	0.7

$KL_{\leq 2}$	π_1	π_2
1-s	0.7	0.7
2-s	1.1	1.5
3-s	1.7	2.3
4-s	2.2	3.1

(a) Ring MDP example. Left: environment transitions. Right: 1-step OM for different policies[3] and the KL divergence in different step views[4].



(b) One-dimensional environment. Left: different policies. Right: Sub-optimality divergence ratios on 1∼3-step views.

Figure 1: Illustrative examples.

3.2 From 1-step to T-step

To mitigate the limitation of the single-step discrepancy and solve the long-term effect, we explore the potential of directly regularizing the multi-step discrepancy. We start with the definition of the T-step OM:

Definition 1 (T-step Occupancy Measure)**.** *The T-step OM is defined as the discounted occurrence probability of a T-step trajectory $\tau^T = \{s^0, a^0, s^1, a^1, \cdots, s^{T-1}, a^{T-1}\}$ that begins with s^0, a^0:*

$$\rho_\pi^T(\tau^T) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s^0, a_t = a^0, \cdots, s_{t+T-1} = s^{T-1}, a_{t+T-1} = a^{T-1} | \pi) \,. \tag{6}$$

For simplicity, we will use the notation ρ to denote the 1-step OM ρ^1 in the following paper. An easy conclusion is that starting from the same state-action pair s^0, a^0, its corresponding T-step OM ρ_π^T and H-step OM ρ_π^H ($H \leq T$) are connected by the policy π and the dynamics M as:

$$\rho_\pi^T(\tau^T) = \rho_\pi^H(\tau^H) \prod_{t=H}^{T-1} \pi(a^t|s^t) M(s^t|s^{t-1}, a^{t-1}) \,. \tag{7}$$

In particular, when $H = 1$, we have:

$$\rho_\pi^T(\tau^T) = \rho_\pi(s^0, a^0) \prod_{t=1}^{T-1} \pi(a^t|s^t) M(s^t|s^{t-1}, a^{t-1}) \tag{8}$$

By definition, we extend the expectation *w.r.t.* the policy π as the expectation under the T-step OM: $\mathbb{E}_\pi[\cdot] \triangleq \sum_{\tau^T} \rho_\pi^T(\tau^T)[\cdot] = \mathbb{E}_{\tau_T \sim \rho_\pi^T}[\cdot]$. Therefore the RL objective can be written in a T-step form:

$$\pi^* = \arg\max_\pi \mathbb{E}_{\rho_\pi^T} \left[r(\tau^T) \right] + \alpha H(\pi) \,, \tag{9}$$

where $r(\tau^T) \triangleq r(s^0, a^0)$.

Now we are ready to extend 1-step conclusions to T-step OM matching. Specifically, we first show the one-one mapping between the policy set Π and the set of τ^T-step OMs $\mathcal{D}(T) \triangleq \{\rho_\pi^T : \pi \in \Pi\}$, which enables us to construct the policy with the T-step OM:

Lemma 1 (Theorem 2 of Syed et al. [2008], Lemma 3.1 of Ho and Ermon [2016])**.** *If $\rho \in \mathcal{D}(1)$, then ρ is the OM for $\pi_\rho \triangleq \rho(s,a) / \sum_{a'} \rho(s,a')$, and π_ρ is the only policy whose OM is ρ.*

For $T > 1$, we are able to induce ρ from ρ^T according to Eq. (7):

Lemma 2. *If $\rho^T \in \mathcal{D}(T)$, then $\rho \in \mathcal{D}(1)$ is the unique 1-step OM corresponding to ρ^T.*

Given Lemma 1 and Lemma 2, now we draw the following conclusions:

Theorem 1 (Extension of Lemma 1)**.** *If $\rho^T \in \mathcal{D}(T)$, then ρ^T is the T-step OM for $\pi_{\rho^T} \triangleq \rho(s^0, a^0) / \sum_{a'} \rho(s^0, a')$ where ρ is the corresponding 1-step OM found by Lemma 2, and π_{ρ^T} is the only policy whose T-step OM is ρ^T.*

Theorem 1 indicates that we can recover the policy if we can match a T-step OM. Thence, similar properties of 1-step OM (Proposition 1 and Proposition 2) still hold for T-step OM:

Proposition 3 (Extension of Proposition 1)**.** *Given the definition of RL procedure as Eq. (9) IRL procedure as:* $\mathrm{IRL}_\psi(\pi_E) = \arg\max_r -\psi(r) + \left(\min_{\pi\in\Pi} -H(\pi) - \mathbb{E}_{\rho_\pi^T}[r(\tau^T)]\right) + \mathbb{E}_{\rho_{\pi_E}^T}[r(\tau^T)]$*, we have*

$$\mathrm{RL}\circ\mathrm{IRL}_\psi(\pi_E) = \arg\min_\pi -H(\pi) + \psi^*(\rho_\pi^T - \rho_{\pi_E}^T) \tag{10}$$

Proposition 4 (Extension of Proposition 2)**.** *Consider the reward function regularizer as:* $\psi_f(r) = \mathbb{E}_{\rho_{\pi_E}^T(s,a)}\left[f^*(\tau^T) + r(\tau^T)\right]$*, where* f^* *is the convex conjugate of* f*, then we have*

$$\psi_f^*(\rho_\pi^T - \rho_{\pi_E}^T) = D_f(\rho_\pi^T(\tau^T)\|\rho_{\pi_E}^T(\tau^T)) \tag{11}$$

$$\mathrm{RL}\circ\mathrm{IRL}_\psi(\pi_E) = \arg\min_\pi -H(\pi) + D_f(\rho_\pi^T\|\rho_{\pi_E}^T) \tag{12}$$

These extended conclusions suggest that it is theoretically sound to generalize previous IL solutions from 1-step OM matching to T-step OM matching, by replacing the state-action pair (s, a) as the T-step trajectory τ^T.

Practical T-step Imitation Learning. In this part we provide an practical algorithm for optimizing the proposed T-step OM matching objective. Motivated by Ho and Ermon [2016], we can simply derive an adversarial algorithm by choosing a specific regularizer ψ, which actually minimizes the JS divergence $\mathrm{D_{JS}}(\rho_\pi^T\|\rho_{\pi_E}^T)$ between the T-step OM of the agent and the expert. In this way, the algorithm alternately updates the discriminator D_w and the policy π_θ following the gradients:

$$\nabla_w\mathcal{L}^D = \hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w^T(\tau^T))] + \hat{\mathbb{E}}_{\tau_E^T}[\nabla_w \log(1 - D_w^T(\tau^T))] \tag{13}$$

$$\nabla_\theta\mathcal{L}^\pi = \hat{\mathbb{E}}_{\rho^T}\left[\nabla_\theta \log \pi_\theta(a|s) Q^T(s,a)\right] \ , \tag{14}$$

where $Q^T(s,a) = \hat{\mathbb{E}}_\pi[\hat{r}^T(s_t,a_t)\,|\,s_0 = s, a_0 = a]$. Note that, instead of operating on single-step state-action pairs, the discriminator now classifies whether a T-step sequence τ^T is drawn from the expert distribution. And the reward function can be constructed using the discriminator as:

$$\hat{r}^T(s,a) \triangleq \hat{r}^T(\tau^T|s^t = s, a^t = a) = \log D(\tau^T) \tag{15}$$

This learning procedure actually can be viewed as a T-step generalization of GAIL Ho and Ermon [2016], which we call multi-step GAIL (MS-GAIL). A detailed description of the algorithm can be found in Algo. 1.

4 Theoretical Analysis

While the proposed multi-step OM matching is conceptually simple, it is still important to investigate the underlying properties of the algorithm. In this section, we analyze rollout discrepancy and sample complexity of the proposed MS-GAIL.

4.1 Discrepancy Analysis on Rollout Sequences

We first study whether matching multi-step OMs results a better solution in a shorter-step view, through analyzing the discrepancy on rollout sequences. Let us rollout the policy π from s_t^0 at timestep t for H step. Our desired goal is to match a long-term rollout sequence τ^H of the expert. Then the discrepancy of the H-step OM between the agent and the expert can be given by the following theorem.

Theorem 2 (Rollout Discrepancy of Multi-Step OM Matching)**.** *If the Kullback–Leibler divergence of two* T*-step normalized OM is smaller than a certain error* ϵ_π*, i.e.,* $D_{KL}\left(\rho_1^T(\tau^T)\|\rho_2^T(\tau^T)\right) \le \epsilon_\pi$*, then we have that the discrepancy of the* H*-step normalized OM (*$H \le T$*) is bounded by*

$$D_{TV}\left(\rho_1^H(s,a)\|\rho_2^H(s,a)\right) \le \sqrt{2\epsilon_\pi}(T - H + 1) \ . \tag{16}$$

The proof is shown in Appendix B.1. Theorem 2 indicates that matching a long step discrepancy significantly contributes to matching a shorter sequence. Specifically, with the same error, minimizing a multi-step OM benefits much more than directly matching the single-step OM. This verifies our observation in the motivating example shown in Fig. 1(a), which suggests that utilizing the multi-step OM matching can benefit the IL tasks.

4.2 Sample Complexity Analysis on Different Length

In this section we concentrate on the sample complexity of our proposed algorithm. Based on the generalization theory in GAN Arora et al. [2017], Zhang et al. [2017] and previous sample complexity analysis in IL Xu et al. [2019], we first introduce the definition of the generalization of multi-step OM matching:

Definition 2 (Generalization of Occupancy Measure Matching)**.** *Given T-step OM $\rho^{\hat{T}}_{\pi_E}$, an empirical distribution of $\rho^T_{\pi_E}$ with m_T samples obtained by π_E, and a T-step OM ρ^T_π generalizes under the distance between two distributions $d(\cdot,\cdot)$ with error ϵ if the following holds with high probability:*

$$|d(\rho^T_\pi, \rho^T_{\pi_E}) - d(\hat{\rho}^T_\pi, \hat{\rho}^T_{\pi_E})| \leq \epsilon \,, \tag{17}$$

where $\hat{\rho^T_\pi}$ and $\hat{\rho}^T_{\pi_E}$ are the empirical distributions of m_T sequence-level samples from ρ_π and ρ_{π_E} respectively.

With the above definition, now we start to analyze the sample complexity of the proposed MS-GAIL, which trains the policy and the discriminator within the generative adversarial framework. We first present the sample complexity theorem here:

Theorem 3 (Lemma 6.3 of Xu et al. [2019])**.** *Assume that the policy π optimizes GAIL objective up to an ϵ error and all discriminator nets D in the discriminator set $\mathcal{C}$ are bounded by Δ, i.e., $\|D\|_\infty \leq \Delta$, $\forall D \in \mathcal{C}$. Let $\hat{\mathcal{R}}_{\rho^{(m)}_{\pi_E}}(\mathcal{C})$ denote the empirical Rademacher complexity of $\mathcal{C}$. Then with probability at least $1-\delta$, the following inequality holds:*

$$D_{TV}(\rho^T_\pi \| \rho^T_{\pi_E}) \leq \sqrt{2\Lambda_{\mathcal{F},\Pi}}\left(\inf_{\pi\in\Pi}\sqrt{\Delta D_{TV}(\rho^T_\pi \| \rho^T_{\pi_E})} + \sqrt{\epsilon} + 2\sqrt{\hat{\mathcal{R}}^{(m_1/T)}_{\rho^T_{\pi_E}}(\mathcal{C})} + 2\Delta\sqrt{\frac{2Tlog(1/\delta)}{m_1}}\right), \tag{18}$$

where $\Lambda_{\mathcal{C},\Pi} = \sup_{\pi\in\Pi} \| \log(\frac{\rho^T_\pi}{\rho^T_{\pi_E}}) \|_{\mathcal{C},1} < \infty$ and m_1 is the number of state-action pairs. The proof of Theorem 3 can be found in Xu et al. [2019], where the only difference is the sample number. Since we are matching the T-step OM, the acquired samples have to be T-step sequences such that $m_1 = Tm_T$. Therefore, to get the same bound, for a larger step length T, it will need much more training samples proportional to the number of samples in 1-step OM matching.

Summary. Combining the conclusions from Theorem 2 and Theorem 3, we derive a trade-off between the performance and the sample complexity. Particularly, given the same capacity on the error of the optimization, we would like to match a multi-step OM (with a step length T as long as it can be) to get a better final performance; unfortunately, to converge to such an optimal policy, a T-step OM matching objective requires T times number of samples compared with a 1-step OM matching algorithm. Thus, to achieve a better result, we need to carefully choose the appropriate step length T. In the next part, we will elaborate on how to effectively mitigate the trade-off via an automated curriculum strategy.

5 Automated Curriculum Multi-Step Imitation Learning

The trade-off between the performance and the sample complexity makes it challenging to determine an appropriate step length T under different scenarios. As shown in Section 4, we require much more state-action training samples if we want to match longer OMs for better imitation results. Therefore, multi-step OM matching can be hard at the beginning of the online training schedule when the training samples are quite limited. A natural solution for the challenge is to expand the step length as the agent collects more samples during interacting with the environment. Specifically, the step length should be selected as the one which provides the most informative momentum for updating the policy, which motivates us to provide a syllabus of curriculum along with the training procedure. Curriculum learning automatically designs and constructs a *curriculum* as a sequence of tasks $K_1, \ldots, K_N$ to be learned, so that the efficiency or performance on the target task K_t can be improved. In our setting, correspondingly, the target task is imitation learning, and the curriculum at each training iteration is the chosen step T of T-step OM matching. Instead of appropriate handcraft curriculum, we apply an elegant automated curriculum framework from Graves et al. [2017] for T-step OM matching, which also provides efficient and flexible training for IL tasks.

Specifically, this formulation takes a curriculum containing N tasks as an adversarial N-armed bandit Bubeck and Cesa-Bianchi [2012], where an agent selects a sequence of arms $a_1 \dots a_T$ over T rounds of play ($a_t \in \{1, \dots, N\}$) and yields a reward r_t for that arm after each round. The goal is to maximize the sum of rewards with an adaptive policy, and they employed a classic adversarial bandits algorithm named Exp3.S Auer et al. [2002] to handle the problem, where the optimal arm is only responsible for a portion of history. On round t, the stochastic policy π_t for selecting an arm i is defined by a set of incrementally multiplicative weight $w_{t,i}$:

$$\pi_t^{\text{EXP3.S}}(i) = \frac{e^{w_{t,i}}}{\sum_{j=1}^{N} e^{w_{t,j}}} + \frac{\epsilon}{N} \tag{19}$$
$$w_{t,i} = \log\left[(1-\alpha_t)\exp\left\{w_{t-1,i}+\eta\tilde{r}_{t-1,i}^{\beta}\right\} + \frac{\alpha_t}{N-1}\sum_{j\neq i}\exp\left\{w_{t-1,j}+\eta\tilde{r}_{t-1,j}^{\beta}\right\}\right] \quad ,$$

where $w_{1,i} = 0$, $\alpha_t = t^{-1}$, $\tilde{r}_{t,i}^{\beta} = \frac{r_t \mathbb{I}_{[a_s=i]}+\beta}{\pi_s(i)}$ is the reward for selecting arm i, and η is the step size. In practice, the received reward $\hat{r}_t$ is adaptively rescaled to lie in the interval $[-1, 1]$ as:

$$r_t = \begin{cases} -1 & \text{if } \hat{r}_t < q_t^{\text{lo}} \\ 1 & \text{if } \hat{r}_t > q_t^{\text{hi}} \\ \frac{2(\hat{r}_t - q_t^{\text{lo}})}{q_t^{\text{hi}} - q_t^{\text{lo}}} - 1 & \text{otherwise} , \end{cases} \tag{20}$$

where q_t^{lo} and q_t^{hi} are 20$^{\text{th}}$ and 80$^{\text{th}}$ percentiles of historical unscaled rewards up to time t: $\mathcal{R}_t = \{\hat{r}_i\}_{i=1}^{t-1}$.

Therefore, to learn an adaptive policy for selecting the curriculum, we need to devise a reward to guide the policy to select the appropriate task. Ideally, we would like to choose a curriculum that can maximize the optimization rate of the target objective, and thus the constructed reward should reflect this optimization rate, *e.g.*, the decreased value of the loss function. Since the target objective of method is represented by a certain distance for T-step OM matching, a natural idea is to utilize the JS divergence between two T-step OMs $\text{D}_{\text{JS}}(\rho_\pi^T \| \rho_{\pi_E}^T)$ as the loss measure, which could be estimated by the loss of the discriminator:

$$-\text{D}_{\text{JS}}(\rho_\pi^T \| \rho_{\pi_E}^T) \approx \mathcal{L}(D_w^T) = \hat{\mathbb{E}}_{\tau_i}[\log(D_w^T(\tau^T))] + \hat{\mathbb{E}}_{\tau_E}[\log(1 - D_w^T(\tau^T))] \, . \tag{21}$$

Similar to GANs Goodfellow et al. [2014], the objective of the discriminator $\mathcal{L}(D_w)$ can reflect the JS divergence of the trajectories between the agent and then expert. Hence, we can take changing range of the loss value before and after each training iteration to evaluate the optimization rate, and take improvement margin as reward to guide the policy optimization. Formally, we abuse the symbol r_k as the reward for the curriculum selection policy at k^{th} training iteration:

$$r_k = \mathcal{L}_k(D_w^T) - \mathcal{L}_{k-1}(D_w^T) \, , \tag{22}$$

where T is the task selected at k^{th} iteration by the policy, *i.e.*, $\pi^{\text{EXP3.S}}(k) = T$.

It is worth noting that the chosen curriculum indeed reflects the learning progress. Intuitively, if the reward for curriculum T is higher than the others, we would like to believe that T-step OM has the most significant discrepancy and should be optimized in priority. We find it useful for improving the sample efficiency and alleviate the training instability and the gradient vanishing problem, which is common in adversarial training Brock et al. [2018]. In practice, we keep and train a limited number of T discriminators as T curriculums, and update the weight of each curriculum following the rules of the EXP3.S algorithm (Eq. (19)). These weights is used to construct the higher-level policy $\pi^{\text{EXP3.S}}$ whose outputs are taken as a syllabus for different step lengths T, enabling it to automatically create stages of curriculum. To prevent the randomness during early training period, we simply adopt an initialized curriculum instead of utilizing $\pi^{\text{EXP3.S}}$ until n training iterations. This resulting algorithm is named automated curriculum GAIL (AutoGAIL). Details of the algorithm can be found in Algo. 2.

6 Related Work

Perfect imitation of single-step behaviors corresponds to match the long-term trajectories of the expert. However, once there is a gap in the single-step, the discrepancy enlarges much more as

the sequence becomes longer due to the error accumulation. This can be understood as the long-term effect of the imitation learning (IL) tasks, which existed in most of the previous solutions. For example, behaviour cloning (BC) methods Pomerleau [1991], Bain and Sammut [1995], adopts supervised training which leads to the notorious compounding error problem Ross et al. [2011], Ross and Bagnell [2010]. The recent popular generative adversarial methods GAIL Ho and Ermon [2016], benefits from interacting with the environment with less compounding error Xu et al. [2019]. However, GAIL essentially matches a single-step occupancy measure (OM), instead of matching a sequence. In our work, we further ease the long-term effect by proposing multi-step OM matching, along with an automated curriculum framework for selecting appropriate step length for optimization.

Curriculum learning (CL) adopts a curriculum of progressive tasks to accelerate the neural network's training Elman [1993], Bengio et al. [2009], which has been widely used in complicated tasks Reed and De Freitas [2015], Graves et al. [2016]. A typical CL solution is using hand-crafted curriculum Zaremba and Sutskever [2014] by assuming the difficulty order of all the tasks, which is usually hard to be quantified. As an improvement, Schmidhuber [2004] proposes automatic curriculum generation and utilizes program search to construct an asymptotically optimal algorithm for this problem. Our automated strategy is built upon the work of Graves et al. [2017], which proposes automated curriculum learning by learning a policy to adaptively decide the task during the training, based on the so-called learning progress Oudeyer et al. [2007] and multi-armed bandit algorithm Bubeck and Cesa-Bianchi [2012]. In our setting, we let the agent choose the appropriate step length as the best curriculum through the training time on multi-step OM matching, so as to improve the sample efficiency and the final performance.

7 Experiments

We conduct several experiments to investigate the following research questions:

RQ1 Does multi-step occupancy measure matching have the potential for improvement?
RQ2 Does and how does the automated strategy of AutoGAIL enhance the performance?
RQ3 What are the key ingredients of AutoGAIL that contribute to the improvements?

To answer RQ1, we evaluate 1$\sim$4-step GAIL and AutoGAIL on various continuous control tasks with different numbers of trajectories. Regarding RQ2, we compare AutoGAIL with a random curriculum selection strategy by showing the learning efficiency with the corresponing curriculum during the training procedure. Finally, we conduct ablation studies on two key hyperparameters (the maximum step length T and the explration ratio ϵ of the high-level policy) of AutoGAIL to address RQ3. Due to the space limit, we leave experimental details and additional results in Appendix C.

Potential in multi-step. Quantitative experiments are conducted to investigate how multi-step GAIL affects the performance when the step length T varies. In particular, we test $1 \sim 4$-step GAIL on continuous control benchmarks: Hopper, Walker2d, HalfCheetah and Ant. For all environments, we first train an Soft Actor-Critic (SAC) Haarnoja et al. [2018] agent to collect expert demonstrations with varying trajectory counts and then train the imitation agents with such data. All algorithms are trained with exactly the same amount of environment interaction and evaluated by a deterministic policy. To measure the imitation efficacy over the sequence, we use the relative return accumulated over the trajectories compared with expert. Fig. 2 depicts the results, which illustrates that there does exist a trade-off between the rollout discrepancy and the sample complexity, according to the chosen step T. Specifically, as is observed, with sufficient expert trajectories, 4-step GAIL can always achieve the best performance than the other GAIL baselines (except AutoGAIL), but it has no advantage when there are fewer trajectories. Besides, the optimal choice of the step length T also varies with different numbers of trajectories on different tasks. However, in most of times, a multi-step solution improves the performance over 1-step GAIL, showing the potential for better sequence-level imitation.

Analysis for automated strategy. Notice that we have illustrated the superior performance of AutoGAIL among different numbers of demonstrations in Fig. 2, where we provide AutoGAIL with 4 ($1 \sim 4$) kinds of curriculum choice. Benifiting from the automated curriculum selection strategy that balances the trade-off between the rollout discrepancy and the sample complexity, AutoGAIL reaches the best performance on almost all tasks. The complete learning curves, shown in Appendix C.3, also provides strong evidence on its good learning efficiency against above multi-step GAIL baselines. Beyond the performance, we also analyze if the curriculum provides instructive

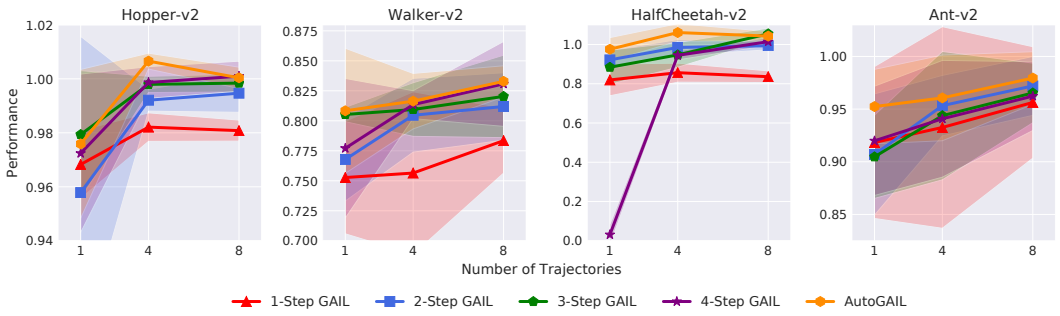


Figure 2: Performance of different T-step GAILs. The y-axis is average return over 5 random seeds, scaled so that the expert achieves 1 and a random policy achieves 0.

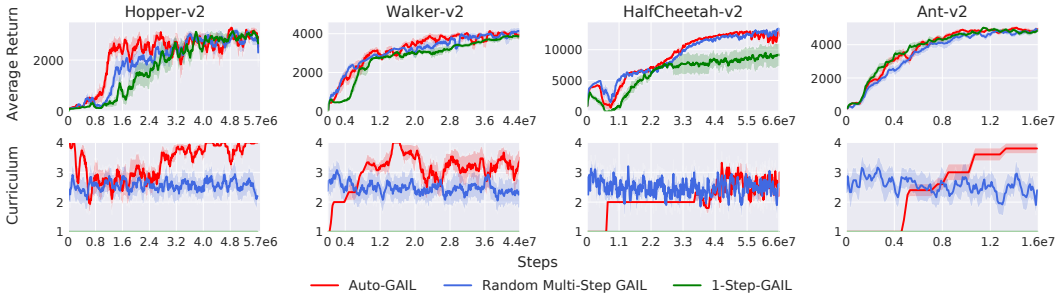


Figure 3: Curriculum selection alongside the training on 4 expert demonstrations over 5 random seeds.

guidance through the training of the imitation agents. To this end, based on 4 expert trajectories, we compare AutoGAIL with a random strategy algorithm (denoted as *Random Multi-step GAIL*) that selects the step length T in a randomized way at each training iteration. As shown in Fig. 3, that the high-level policy is well learned to provide reasonable choices on the selection of the curriculum, which accomplishes the higher learning efficiency against the random strategy in most of the cases. To explain the rationality of the curriculum, let us first conclude from Fig. 2: on Hopper, Walker and Ant, 3-step GAIL and 4-step GAIL have the ability of reaching the better performances than 1-step and 2-step methods; on the contrary, 2-step GAIL and 3-step GAIL are good enough on HalfCheetah. As a result, AutoGAIL tends to choose a longer-step curriculum as the training goes on Hopper, Walker and Ant, when the number of samples are no longer limited; on the other hand, AutoGAIL does not even take a 4-step curriculum on HalfCheetah but stays at the 2-step curriculum for a long time. To our surprise, on Walker, a randomized strategy can achieve a similar result as good as AutoGAIL. This indicates the advantage of using a multi-step OM matching objective that even a random step length (instead of 1-step always) is beneficial for imitation learning. Complete training results on different numbers of demonstrations are available in Appendix C.3.

Ablation study. AutoGAIL has two important hyperparameters, namely, the maximum step length T and the exploration ratio ϵ of the high-level policy π^{EXP3}. To go deep into the algorithm, we further perform a diverse set of analyses on assessing the impact of these two hyperparameters under 4 expert trajectories. The comparison results are plotted in Fig. 4 and the detailed quantitive results is provided in Appendix C.3. A brief conclusion is that 1) a small value of T limits the ability of AutoGAIL and the performance can hardly improve when T is large enough for the task; and 2) the exploration ratio slightly affects the final performance of AutoGAIL and a greedy choice ($\epsilon = 0$) also keeps a good result. Nevertheless, all the variants consistently outperform 1-step GAIL.

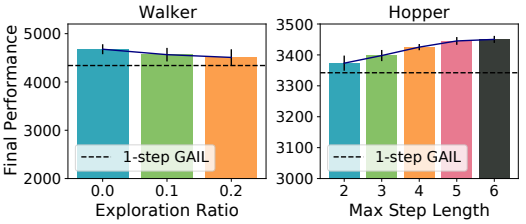


Figure 4: Ablation study.

8 Conclusion

In this paper we propose multi-step occupancy measure (OM) matching to alleviate the long-term effect in imitation learning tasks. Based on the analysis of the trade-off between the sample com-

plexity and the rollout discrepancy, we find it challenging to determine appropriate step length in practice. Therefore, we further propose AutoGAIL that constructs automated curriculum learning for multi-step OM matching by learning a high-level policy. AutoGAIL chooses the curriculum on the current level of the agent and is able to provide a good result both on the learning efficiency and the final performance.

References

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573*, 2017.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

M. Bain and C. Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pages 103–129, 1995.

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

M. Bloem and N. Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. In *53rd IEEE Conference on Decision and Control*, pages 4911–4916. IEEE, 2014.

A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

J. L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

J. Fu, K. Luo, and S. Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

S. K. S. Ghasemipour, R. Zemel, and S. Gu. A divergence minimization perspective on imitation learning methods. *arXiv preprint arXiv:1911.02256*, 2019.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. G. Colmenarejo, E. Genstette, T. Ramalho, J. Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

A. Graves, M. G. Bellemare, J. Menick, R. Munos, and K. Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pages 1311–1320. PMLR, 2017.

T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573, 2016.

P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2):265–286, 2007.

D. A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991.

S. Reed and N. De Freitas. Neural programmer-interpreters. *arXiv preprint arXiv:1511.06279*, 2015.

S. Ross and D. Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 661–668, 2010.

S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635, 2011.

J. Schmidhuber. Optimal ordered problem solver. *Machine Learning*, 54(3):211–254, 2004.

U. Syed, M. Bowling, and R. E. Schapire. Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on Machine learning*, pages 1032–1039. ACM, 2008.

T. Xu, Z. Li, and Y. Yu. On value discrepancy of imitation learning. *arXiv preprint arXiv:1911.07027*, 2019.

W. Zaremba and I. Sutskever. Learning to execute. *arXiv preprint arXiv:1410.4615*, 2014.

P. Zhang, Q. Liu, D. Zhou, T. Xu, and X. He. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017.

Appendices

A Algorithm

Algorithm 1 Multi-Step Generative Adversarial Imitation Learning (MS-GAIL)

Input: Sequence length T, expert demonstration data $\tau_E = \{(s_i, a_i)\}_{i=1}^N$, parameterized discriminator D_w^T, parameterized policy π_θ.
for $k = 0, 1, 2, \dots$ **do**
  Sample T-step length sequence $\tau^T \sim \pi_\theta$.
  Optimize w with the gradient:

$$\hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w^T(\tau^T))] + \hat{\mathbb{E}}_{\tau_E^T}[\nabla_w \log(1 - D_w^T(\tau^T))] \ .$$

  Update θ with the gradient:

$$\hat{\mathbb{E}}_{\rho^T}\left[\nabla_\theta \log \pi_\theta(a|s) Q^T(s,a)\right] \ .$$

  where $Q^T(s,a) = \hat{\mathbb{E}}_\pi[\hat{r}^T(s_t, a_t) \,|\, s_0 = s, a_0 = a]$, and

$$\hat{r}^T(s,a) = \hat{r}^T(\tau^T | s^t = s, a^t = a) = \log D(\tau^T) \ .$$

end for
return π

Algorithm 2 Automated Curriculum Generative Adversarial Imitation Learning (AutoGAIL)

Input: Maximum sequence length T, expert demonstration data $\tau_E = \{(s_i, a_i)\}_{i=1}^N$, parameterized discriminator D_w^T, parameterized policy π_θ, initial curriculum id i_0, least selection iteration n.
for $k = 0, 1, 2, \dots$ **do**
  for $t = 0, \cdots, T$ **do**
    Sample t-step length sequence $\tau^t \sim \pi_\theta$.
    Optimize w with the gradient:

$$\hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w^t(\tau^t))] + \hat{\mathbb{E}}_{\tau_E^t}[\nabla_w \log(1 - D_w^t(\tau^t))] \ .$$

    Compute the loss and construct the reward of EXP3.S $r_k^{\text{EXP3.S}} = L^t(D^t)$ following Eq. (22).
    Rescale the reward follows Eq. (22).
    Update the high-level policy $\pi^{\text{EXP3.S}}(t)$ using the rescaled reward following Eq. (19).
    if $k < n$ **then**
      Set the current curriculum id $i = i_0$.
    else
      Set the current curriculum id $i = \pi^{\text{EXP3.S}}(k)$.
    end if
  end for
  Update θ with the gradient w.r.t the current curriculum i:

$$\hat{\mathbb{E}}_{\rho^T}\left[\nabla_\theta \log \pi_\theta(a|s) Q^i(s,a)\right] \ .$$

  where $Q^i(s,a) = \hat{\mathbb{E}}_\pi[\hat{r}^i(s_t, a_t) \,|\, s_0 = s, a_0 = a]$, and

$$\hat{r}^i(s,a) = \hat{r}^i(\tau^i | s^t = s, a^t = a) = \log D(\tau^i) \ .$$

end for
return π

B Proofs

We introduce useful lemmas before providing our proof.

Lemma 3 (Relation between the Kullback–Leibler divergence and the total variation distance)**.** *Given two distributions $p_1(x)$ and $p_2(x)$, the relationship between their Kullback–Leibler divergence and total variation distance is:*

$$D_{TV}(p_1(x)\|p_2(x)) = \left(\frac{1}{2}D_{KL}(p_1(x)\|p_2(x))\right)^{\frac{1}{2}}$$

Lemma 4 (Total variation distance of joint distributions)**.** *Given two joint distributions $p_1(x,y) = p_1(y|x)p_1(x)$ and $p_2(x,y) = p_2(y|x)p_2(x)$, then the total variation distance has the following bound:*

$$\mathbb{E}_{x\sim p_1(x)}\left[D_{TV}(p_1(y|x)\|p_2(y|x))\right] \leq D_{TV}(p_1(x,y)\|p_2(x,y)) + D_{TV}(p_1(x)\|p_2(x))$$

Proof.

$$\begin{aligned}
\mathbb{E}_{x\sim p_1}\left[\mathrm{D_{TV}}(p_1(y|x)\|p_2(y|x))\right] &= \frac{1}{2}\sum_{x,y} p_1(x)\,|p_1(y|x) - p_2(y|x)| \\
&= \frac{1}{2}\sum_{x,y} |p_1(x)p_1(y|x) - p_1(x)p_2(y|x)| \\
&= \frac{1}{2}\sum_{x,y} |p_1(x)p_1(y|x) - p_1(x)p_2(y|x) + p_2(x)p_2(y|x) - p_2(x)p_2(y|x)| \\
&= \frac{1}{2}\sum_{x,y} |p_1(x,y) - p_2(x,y) + (p_2(x) - p_1(x))p_2(y|x)| \\
&\leq \frac{1}{2}\sum_{x,y} |p_1(x,y) - p_2(x,y)]| + \frac{1}{2}\sum_{x} |p_2(x) - p_1(x)| \\
&= \mathrm{D_{TV}}(p_1(x,y)\|p_2(x,y)) + \mathrm{D_{TV}}(p_1(x)\|p_2(x))
\end{aligned}$$

□

Lemma 5. *Given two joint distributions $p_1(x,y) = p_1(y|x)p_1(x)$ and $p_2(x,y) = p_2(y|x)p_2(x)$, then the total variation distance has the following bound:*

$$D_{TV}(p_1(x)\|p_2(x)) \leq D_{TV}(p_1(x,y)\|p_2(x,y))$$

Proof.

$$\begin{aligned}
\mathrm{D_{TV}}(p_1(x)\|p_2(x)) &= \frac{1}{2}\sum_{x} |p_1(x) - p_2(x)| \\
&= \frac{1}{2}\sum_{x} \left|\sum_{y} p_1(x,y) - p_2(x,y)\right| \\
&\leq \frac{1}{2}\sum_{x,y} |p_1(x,y) - p_2(x,y)| \\
&= \mathrm{D_{TV}}(p_1(x,y)\|p_2(x,y))
\end{aligned}$$

□

B.1 Proof of Theorem 2

To give a proof of Theorem 2, we start with the policy discrepancy bound if we bound the discrepancy of a 1-step OM.

Lemma 6. *Assume the Kullback–Leibler divergence of two 1-step normalized OM is bounded by $D_{KL}\left(\rho_1(s,a)\|\rho_2(s,a)\right) \leq \epsilon_\pi$, then we have that the policy discrepancy is bounded by*

$$\mathbb{E}_{s\sim\pi_1}\left[D_{TV}(\pi_1(a|s)\|\pi_2(a|s))\right] \leq (2\epsilon_\pi)^{\frac{1}{2}} \tag{23}$$

Proof. Follows Lemma 3, Lemma 4 and Lemma 5, we have that

$$\begin{aligned}\mathbb{E}_{s\sim\rho_1(s)}\left[\mathrm{D_{TV}}(\pi_1(a|s)\|\pi_2(a|s))\right] &\leq \mathrm{D_{TV}}(\rho_1(s,a)\|\rho_2(s,a)) + \mathrm{D_{TV}}(\rho_1(s)\|\rho_2(s)) \\ &\leq 2\mathrm{D_{TV}}(\rho_1(s,a)\|\rho_2(s,a)) \\ &\leq 2\left(\frac{1}{2}\mathrm{D_{KL}}(\rho_1(s,a)\|\rho_2(s,a))\right)^{\frac{1}{2}} \\ &\leq (2\epsilon_\pi)^{\frac{1}{2}}\end{aligned} \tag{24}$$

□

After that, we also need to know the conclusion when it is extended to T-step OM.

Lemma 7. *Assume the Kullback–Leibler divergence of two T-step normalized OM is bounded by $D_{KL}\left(\rho_1^T(\tau^T)\|\rho_2^T(\tau^T)\right) \leq \epsilon_\pi$, then we have that the policy discrepancy is bounded by*

$$\mathbb{E}_{s\sim\pi_1}\left[D_{TV}(\pi_1(a|s)\|\pi_2(a|s))\right] \leq (2\epsilon_\pi)^{\frac{1}{2}} \tag{25}$$

Proof. We begin the deviation by showing the relation of the Kullback–Leibler divergence of two T-step normalized occupancy measure and the Kullback–Leibler divergence of two corresponding 1-step normalized occupancy measure:

$$\begin{aligned}\mathrm{D_{KL}}\left(\rho_1^T(\tau^T)\|\rho_2^T(\tau^T)\right) &= \sum_{\tau^T}\rho_1^T(\tau^T)\log\frac{\rho_1^T(\tau^T)}{\rho_2^T(\tau^T)} \\ &= \sum_{\tau^T}\rho_1(s^0,a^0)\prod_{t=1}^{T-1}\pi_1(a^t|s^t)M(s^t|s^{t-1},a^{t-1})\log\frac{\rho_1(s^0,a^0)\prod_{t=1}^{T-1}\pi_1(a^t|s^t)}{\rho_2(s^0,a^0)\prod_{t=1}^{T-1}\pi_2(a^t|s^t)} \\ &= \sum_{\tau^T}\rho_1(s^0,a^0)\log\frac{\rho_1(s^0,a^0)}{\rho_2(s^0,a^0)}\prod_{t=1}^{T-1}\pi_1(a^t|s^t)M(s^t|s^{t-1},a^{t-1}) \\ &\quad + \sum_{\tau^T}\rho_1(s^0,a^0)\prod_{t=1}^{T-1}\pi_1(a^t|s^t)M(s^t|s^{t-1},a^{t-1})\log\prod_{t=1}^{T-1}\frac{\pi_1(a^t|s^t)}{\pi_2(a^t|s^t)}\,.\end{aligned} \tag{26}$$

For the second term in Eq. (26), we denote $P^{1,T} = \prod_{t=1}^{T-1}\pi(a^t|s^t)M(s^t|s^{t-1},a^{t-1})$, then we have that:

$$\begin{aligned}&\sum_{\tau^T}\rho_1(s^0,a^0)\prod_{t=1}^{T-1}\pi_1(a^t|s^t)M(s^t|s^{t-1},a^{t-1})\log\prod_{t=1}^{T-1}\frac{\pi_1(a^t|s^t)}{\pi_2(a^t|s^t)} \\ &= \sum_{\tau^T}\rho_1(s^0,a^0)\prod_{t=1}^{T-1}\pi_1(a^t|s^t)M(s^t|s^{t-1},a^{t-1})\left(\sum_{t=1}^{T-1}\log\frac{\pi_1(a^t|s^t)}{\pi_2(a^t|s^t)}\right) \\ &= \sum_{s^0,a^0}\rho_1(s^0,a^0)\sum_{t=1}^{T-1}\sum_{s^1,a^1,s^2,a^2,\cdots,a^{T-1}}P^{1,T}\log\frac{\pi_1(a^t|s^t)}{\pi_2(a^t|s^t)} \\ &= \sum_{s^0,a^0}\rho_1(s^0,a^0)\sum_{t=1}^{T-1}\sum_{s^1,a^1,s^2,a^2,\cdots,a^{T-1}}P^{1,T}\log\frac{P^{1,T}}{P^{1,T}\frac{\pi_2(a^t|s^t)}{\pi_1(a^t|s^t)}} \\ &= \sum_{s^0,a^0}\rho_1(s^0,a^0)\sum_{t=1}^{T-1}\mathrm{D_{KL}}(P^{1,T}\|P^{1,T}\frac{\pi_2(a^t|s^t)}{\pi_1(a^t|s^t)}) \\ &\geq 0\,.\end{aligned} \tag{27}$$

For the first term in Eq. (26), we have that:

$$\begin{aligned}
&\sum_{\tau^T} \rho_1(s^0,a^0)\log\frac{\rho_1(s^0,a^0)}{\rho_2(s^0,a^0)}\prod_{t=1}^{T-1}\pi_1(a^t|s^t)M(s^t|s^{t-1},a^{t-1}) \\
&= \sum_{s^0,a^0} \rho_1(s^0,a^0)\log\frac{\rho_1(s^0,a^0)}{\rho_2(s^0,a^0)}\sum_{\tau-(s^0,a^0)}\prod_{t=1}^{T-1}\pi_1(a^t|s^t)M(s^t|s^{t-1},a^{t-1}) \\
&= \sum_{s^0,a^0} \rho_1(s^0,a^0)\log\frac{\rho_1(s^0,a^0)}{\rho_2(s^0,a^0)}\sum_{\tau} p(\tau|s^0,a^0) \\
&= \mathrm{D}_{\mathrm{KL}}\left(\rho_1(s,a)\|\rho_2(s,a)\right) \ .
\end{aligned} \tag{28}$$

Therefore, we conclude that $\mathrm{D}_{\mathrm{KL}}\left(\rho_1(s,a)\|\rho_2(s,a)\right) \le \mathrm{D}_{\mathrm{KL}}\left(\rho_1^T(\tau^T)\|\rho_2^T(\tau^T)\right) \le \epsilon_\pi$. Combining Lemma 6 completes the proof. □

Now we are ready to give the proof of Theorem 2.

Theorem 2 (Rollout Discrepancy of Multi-Step OM Matching)**.** *If the Kullback–Leibler divergence of two T-step normalized OM is smaller than a certain error ϵ_π, i.e., $D_{KL}\left(\rho_1^T(\tau^T)\|\rho_2^T(\tau^T)\right) \le \epsilon_\pi$, then we have that the discrepancy of the H-step normalized OM ($H \le T$) is bounded by*

$$D_{TV}\left(\rho_1^H(s,a)\|\rho_2^H(s,a)\right) \le \sqrt{2\epsilon_\pi}(T-H+1)\ . \tag{29}$$

Proof. Before start, let us denote $P^{H,T} = \prod_{t=H}^{T-1}\pi(a^t|s^t)M(s^t|s^{t-1},a^{t-1})$, $M^{H,T} = \prod_{t=H}^{T-1}M(s^t|s^{t-1},a^{t-1})$ and $\pi^{H,T} = \prod_{t=H}^{T-1}\pi(a^t|s^t)$, then we have that:

$$\mathrm{D}_{\mathrm{TV}}(\rho_1^T(s,a)\|\rho_2^T(s,a)) \le \mathrm{D}_{\mathrm{TV}}(\rho_1^H(s,a)\|\rho_2^H(s,a)) + \max \mathrm{D}_{\mathrm{TV}}\left(P_1^{H,T}\|P_2^{H,T}\right) \tag{30}$$

We continue the deviation on the second term, which can be further decomposed as:

$$\mathrm{D}_{\mathrm{TV}}\left(P_1^{H,T}\|P_2^{H,T}\right) \le \mathrm{D}_{\mathrm{TV}}\left(M_1^{H,T}\|M_2^{H,T}\right) + \max \mathrm{D}_{\mathrm{TV}}\left(\pi_1^{H,T}\|\pi_2^{H,T}\right)\ . \tag{31}$$

Since we always choose to rollout in the same environment for all policies, therefore there are no difference between M_1 and M_2, which leads Ineq. (31) to a simpler form which contains only the total variation of the policies.

$$\begin{aligned}
\mathrm{D}_{\mathrm{TV}}\left(P_1^{H,T}\|P_2^{H,T}\right) &\le \max \mathrm{D}_{\mathrm{TV}}\left(\pi_1^{H,T}\|\pi_2^{H,T}\right) \\
&\le \sum_{t=H}^{T-1}\max(\mathrm{D}_{\mathrm{TV}}\left(\pi_1(a^t|s^t)\|\pi_2(a^t|s^t)\right)) \\
&\le \sqrt{2\epsilon_\pi}(T-H)\ .
\end{aligned} \tag{32}$$

By the relation of the total variation and the Kullback–Leibler divergence we also have that:

$$\mathrm{D}_{\mathrm{TV}}(\rho_1^H(s,a)\|\rho_2^H(s,a)) \le (2\epsilon_\pi)^{\frac{1}{2}} \tag{33}$$

Therefore combining Ineq. (32) and Ineq. (33) we have completed the proof. □

C Experiments

C.1 Implementation

To yield a more fair comparison, the implementation of all the algorithms is based on a same open-source PyTorch framework[5]. The expert used for collecting the demonstration is trained via Soft Actor-Critic (SAC) Haarnoja et al. [2018] based on the original implementaion[6]. For all policy and value functions, we use a 2-layer MLP as the network structure. Particularly, for multi-step GAILs, we concatenate the multi-step sequence as the input.

[5]https://github.com/KamyarGh/rl_swiss

[6]https://github.com/rail-berkeley/softlearning

C.2 Important Hyperparameters

We list all important hyperparameters in Tab. 1. Most of these hyperparameters are the default without any finetuning, but we tune our experiments among the exploration ratio ϵ in a small range.

Table 1: Important Hyperparameters.

	Environments	Hop.	Walk.	Half.	Ant
Hyperparameters of MS-GAIL	Trajectory maximum length	1000			
	Optimizer	AdamOptimizer			
	Discount factor γ	0.99			
	Replay buffer size	2e5			
	Batch size	256			
	Generative adversarial reward form	$\log D$	$-\log(1-D)$		
	Q learning rate	3e-4			
	π learning rate	3e-4			
	D learning rate	3e-4			
	Gradient penalty weight	4.0	8.0	16.0	0.5
	Reward scale	2.0			
Hyperparameters of AutoGAIL	Exploration Ratio Epsilon ϵ	[0.0, 0.05, 0.1, 0.2, 0.3]			
	Step Length T	4			

C.3 Additional Results

Quantitative results.

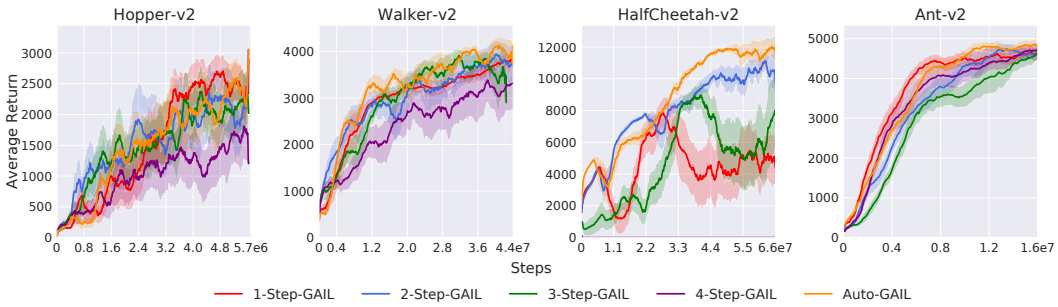
The exact quantitative experiment results of Fig. 2 are shown in Tab. 2, where all methods are evaluated on $\{1, 4, 8\}$ demonstrations over more than 5 random seeds. On most of the tasks, MS-GAIL achieves better performance than the normal 1-step GAIL, indicating the potential of minimizing a multi-step objective. However, the optimal step length T is hard to be determined under different settings. As a comparison, AutoGAIL offers a great choice for balancing the trade-off between the rollout discrepancy and the sample complexity, which is able to reach the best or close to the best performance among all tasks.

Learning curves.

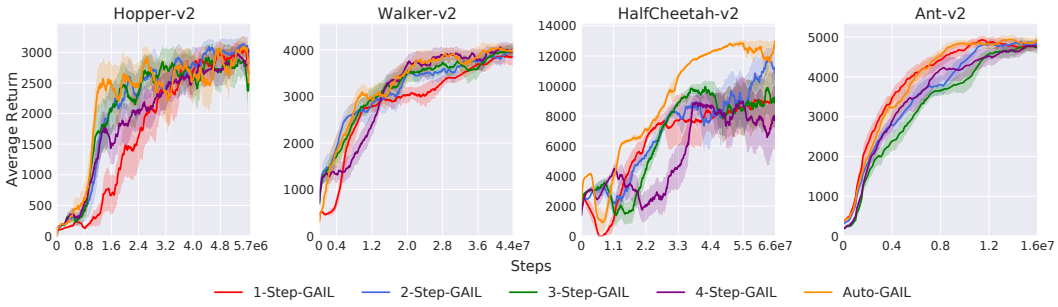
We illustrate the complete learning curves on all tasks to compare the learning efficiency of these methods in Fig. 5. The curves provide strong evidence that AutoGAIL owns competitive learning efficiency against all multi-step GAIL baselines. Although on different environments with different counts of demonstration, the leader always changes hands, AutoGAIL is stable and is not affected much by the setting. Specifically, AutoGAIL wins on the HalfCheetah of 1 and 8 expert trajectories with large margins; 1-step GAIL always results in a quick convergence on Ant, but the final performance leaves a gap between the other MS-GAIL and AutoGAIL; on Hopper and Walker, AutoGAIL always shares the fast efficiency with MS-GAIL of different step lengths.

Table 2: Quantitative results for all methods on different count of demonstrations. The means and the standard deviations are evaluated over more than 5 random seeds.

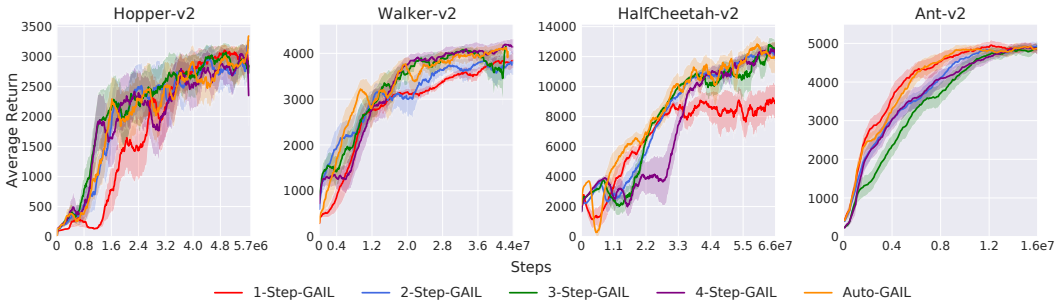
		Hopper	Walker2d	HalfCheetah	Ant
	Random	13.09 ± 0.10	7.07 ± 0.13	74.48 ± 12.39	713.59 ± 203.92
1 Demo	1-Step GAIL	3295.41 ± 42.76	4246.59 ± 264.35	11249.05 ± 978.32	5021.4 ± 338.5
	2-Step GAIL	3260.08 ± 196.27	4331.55 ± 195.65	12642.07 ± 787.04	4968.07 ± 269.98
	3-Step GAIL	**3333.04 ± 80.23**	4542.61 ± 33.89	12131.73 ± 1616.28	4597.46 ± 185.34
	4-Step GAIL	3309.43 ± 99.53	4383.99 ± 327.86	499.35 ± 711.79	5028.79 ± 243.5
	Auto-GAIL	3321.24 ± 93.5	**4559.91 ± 293.77**	**13380.29 ± 1014.81**	**5181.33 ± 165.73**
4 Demo	1-Step GAIL	3342.29 ± 17.95	4267.63 ± 381.81	11751.97 ± 1077.34	5088.27 ± 449.07
	2-Step GAIL	3376.11 ± 13.77	4538.97 ± 173.3	13504.01 ± 620.4	5185.44 ± 134.84
	3-Step GAIL	3395.95 ± 10.23	4565.8 ± 125.85	12988.54 ± 1528.38	5140.87 ± 286.02
	4-Step GAIL	3398.34 ± 19.96	4588.79 ± 65.7	12956.38 ± 1342.46	5126.84 ± 260.11
	Auto-GAIL	**3425.21** ± 10.01	**4605.17 ± 130.45**	**14540.69 ± 103.59**	**5219.97 ± 190.87**
8 Demo	1-Step GAIL	3337.9 ± 13.24	4419.84± 152.7	11473.72 ± 749.37	5199.73 ± 248.32
	2-Step GAIL	3384.91 ± 24.03	4580.16 ± 158.37	13622.72 ± 719.07	5273.29 ± 131.1
	3-Step GAIL	3397.58 ± 11.24	4627.5 ± 193.67	14450.49 ± 373.77	5242.05 ± 132.86
	4-Step GAIL	**3406.24** ± 18.88	4686.19 ± 199.41	13907.51 ± 633.84	5227.89 ± 153.35
	Auto-GAIL	3404.24 ± 13.4	**4697.99** ± 73.29	**14295.69 ± 324.88**	**5308.34 ± 117.38**
	Expert (SAC)	3402.94 ± 446.48	5639.32 ± 29.97	13711.64 ± 111.47	5404.55 ± 1520.49



(a) 1 demonstration.



(b) 4 demonstration.

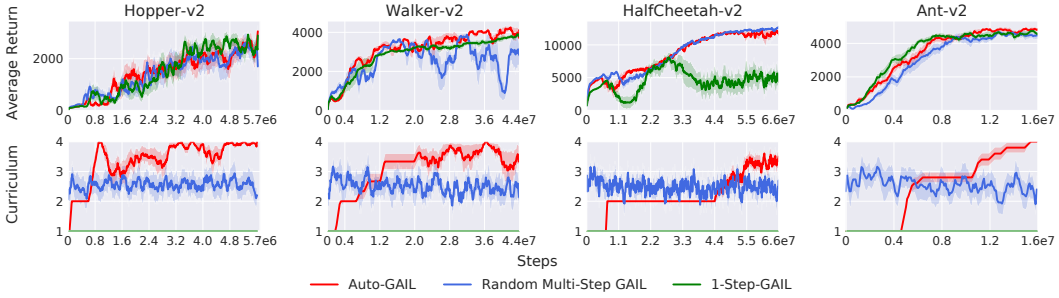


(c) 8 demonstration.

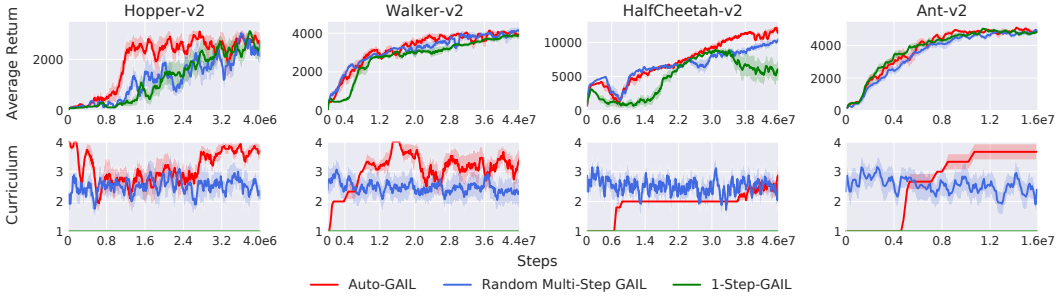
Figure 5: Learning curves with different numbers of demonstration.

Curriculum selection.

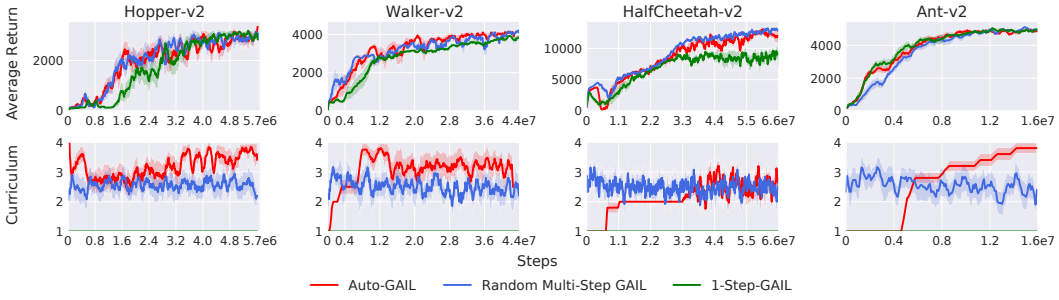
We also provide the complete curves of the selection of the curriculum during the training process on each task against the random MS-GAIL baselines, which selects the curriculum randomly at each training iteration. Interestingly, the random strategy is competitive with AutoGAIL on a considerable number of tasks, showing that the imitation learning results benefit a lot from even a random multi-step strategy. However, examples as on Walker with 1 demonstration, on Hopper with 4 demonstrations and on every Ant task, still indicate that the random strategy causes instability or inefficiency. Notably, on HalfCheetah, AutoGAIL tends to choose the step length as 2 or 3, this meets the averaged choice of random MS-GAIL, and therefore the training processes are very similar.



(a) 1 demonstration.



(b) 4 demonstration.



(c) 8 demonstration.

Figure 6: Curriculum selection with different numbers of demonstration.