Self-Adaptive Imitation Learning: Learning Tasks with Delayed Rewards from Sub-Optimal Demonstrations

Zhuangdi Zhu,[1] **Kaixiang Lin,** [1] **Bo Dai,** [2] **Jiayu Zhou** [1]

[1] Michigan State University
[2] Google Brain
zhuzhuan@msu.edu, lkxcarson@gmail.com, bodai@google.com, jiayuz@msu.edu

Abstract

Reinforcement learning (RL) has demonstrated its superiority in solving sequential decision-making problems. However, heavy dependence on immediate reward feedback impedes the wide application of RL. On the other hand, imitation learning (IL) tackles RL without relying on environmental supervision by leveraging external demonstrations. In practice, however, collecting sufficient expert demonstrations can be prohibitively expensive, yet the quality of demonstrations typically limits the performance of the learning policy. To address a practical scenario, in this work, we propose *Self-Adaptive Imitation Learning (SAIL)*, which, provided with a few demonstrations from a sub-optimal teacher, can perform well in RL tasks with extremely delayed rewards, where the only reward feedback is trajectory-wise ranking. *SAIL* bridges the advantages of IL and RL by interactively exploiting the demonstrations to catch up with the teacher and exploring the environment to yield demonstrations that surpass the teacher. Extensive empirical results show that not only does *SAIL* significantly improve the sample efficiency, but it also leads to higher asymptotic performance across different continuous control tasks, compared with the state-of-the-art.

Introduction

Reinforcement Learning (RL) is notably advantageous in learning sequential decision-making problems in simulated environments, such as game-playing (Mnih et al. 2015; Silver et al. 2017), where massive samples with dense rewards can be accessed at a negligible cost. However, it is challenging to upscale RL to real-world scenarios due to its dependence on immediate reward feedback. For practical applications where rewards are usually delayed in time and sparse in value, RL agents may struggle with high sample complexity, facing difficulties of connecting a long sequence of actions to the feedback received in the far future.

In fact, the ability to learn from delayed feedback is crucial for realizing advanced artificial intelligence (Christiano et al. 2017; Reddy, Dragan, and Levine 2019). On the one hand, reducing the frequency of reward sampling contributes to a lower interaction complexity for practical applications, such as autonomous-driving (Pham et al. 2018) and UAV navigation (Kiran et al. 2020). On the other hand, learning from coarse-grained supervision, such as human preference (Kupcsik, Hsu, and Lee 2018), is rather useful when it is easy to recognize the desired behavior but difficult to explain its rationale by designing delicate reward functions (Palan et al. 2019).

Recent advances of Imitation Learning (IL) can effectively provide remedies when the environment feedbacks are delayed or even unavailable, by referencing expert demonstrations (Ho and Ermon 2016; Kostrikov et al. 2019; Kostrikov, Nachum, and Tompson 2020) or policies (Ross, Gordon, and Bagnell 2011; Sun et al. 2017). In spite of their success, a major limitation of such IL approaches is that the learned performance is bounded by the given expert. Consequently, when the provided demonstrations are sub-optimal, which is a practical yet more challenging scenario, the IL approaches will induce a sub-optimal policy. In the meantime, some work has been proposed for learning from sub-optimal guidance in delayed rewarded tasks (Kang, Jie, and Feng 2018; Sun, Bagnell, and Boots 2018; Wu et al. 2019; Zhang et al. 2019; Gao et al. 2018). A shared rationale among them is to augment the environment rewards with synthetic rewards derived from the demonstrations, after which an actor-critic algorithm can take over the policy learning. Although technically effective, these approaches are inherent with twofold limitations. First, *the sub-optimality of teacher demonstrations has not been fully resolved*. Once the learning agent reaches a reasonable performance, the demonstrations will become a bottleneck, leading to negative guidance that contradicts environment feedbacks (Jing et al. 2020). Second, *the environment feedback is not well leveraged*. Learning a *critic* function relying on delayed environment rewards can be sample costly, which may provide weak signals to compensate for the sub-optimal demonstrations.

In this paper, we formally consider a problem setting, where an RL agent only has access to a limited number of *sub-optimal* demonstrations in a task with highly *delayed* rewards. Our goal hence is to combine the merits of RL and IL, by exploring the sub-optimal demonstrations that are easier to access in practice, while preserving the chance to explore for better policies guided by the coarse-grained environment feedbacks.

Noticing the challenges of the proposed problem and limitations in prior arts, we propose *Self-Adaptive Imitation Learning (SAIL)*, an off-policy imitation learning approach

that strikes a balance between exploitation and exploration to reach high performance. More concretely, we formulate our objective as exploration-driven IL. On the one hand, our approach minimizes the discrepancy between the teacher and the learning policy; on the other hand, it encourages the learning policy to deviate from its previously learned predecessors for better exploration. Specifically, we leverage the delayed feedback from the environment to explore superior self-generated trajectories that surpass the teacher's performance. Those self-generated trajectories are used to replace the suboptimal teachers to constructs a dynamic target distribution that gradually converges to optimality. An overview of our proposed approach is provided in Figure 1. Extensive empirical studies have shown that *SAIL* achieves significant improvement regarding both sample efficiency and asymptotic performance on various popular benchmarks.

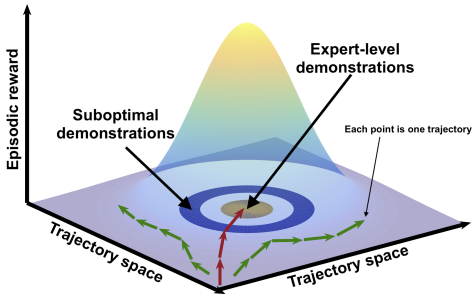

Figure 1: Illustraion of *SAIL*: Navigations in red arrows follow the exploration-driven IL objective, which approaches to teacher's density distribution while deviating from previous learned ones. It explores more efficiently to reach expertise, compared with random explorations (green arrows).

Background

Markov Decision Process (MDP) is an ideal environment to formulate RL, which can be defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma, \mu_0, \mathcal{S}_0)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action space, $\mathcal{T}(s'|s,a)$ denotes the probability of the environment transitioning from state s to s' upon action a is taken, $r(s,a)$ is the environment reward received by taking action a on state s, $\gamma \in (0,1])$ is a discounted factor, μ_0 is the initial state distribution, and $\mathcal{S}_0$ is the set of terminal states or *absorbing states*. Any absorbing state always transits to itself and yields a reward of zero (Sutton and Barto 2018). Given a trajectory $\tau = \{(s_t, a_t)\}_{t=0}^{\infty}$, we define its return as $R(\tau) = \sum_{k=0}^{\infty} \gamma^k r(s_k, a_k)$. For an episodic task with a *finite* horizon, its return can be written as $R(\tau) = \sum_{k=0}^{T} \gamma^k r(s_k, a_k)$, where T is the number of steps to reach an absorbing state.

The objective of RL is to learn a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the expected return of its trajectories. Equivalently, this objective can be rephrased as finding a distribution $d_\pi(s,a)$:

$$\max_\pi \ \eta(\pi) := \mathbb{E}_{(s,a)\sim d_\pi(s,a)}\big[r(s,a)\big], \qquad (1)$$

in which $d_\pi(s,a)$ is the *normalized stationary state-action distribution* of π: $d_\pi(s,a) = (1-\gamma)\mu^\pi(s,a)$, and $\mu^\pi(s,a)$ is the *occupancy measure* of a policy π, defined as: $\mu^\pi(s,a) = \sum_{t=0}^{\infty} \gamma^t Pr\big(s_t = s, a_t = a | s_0 \sim \mu_0, a_t \sim \pi(s_t), s_{t+1} \sim \mathcal{T}(s_t, a_t)\big)$ (Ho and Ermon 2016). Without ambiguity, we use *density* and *normalized stationary state-action distribution* interchangeably to refer $d_\pi(s,a)$ in this paper.

Adversarial Imitation Learning addresses IL from the perspective of distribution matching. A representative work along this line is *Generative Adversarial Imitation Learning (GAIL)* (Ho and Ermon 2016). Given a set of demonstrations from an unknown expert policy π_E, GAIL aims to learn a policy π that minimizes the Jensen-Shannon divergence between d_π and d_E:

$$\min_\pi \mathbb{D}_{\text{JS}}[d_\pi(s,a) || d_E(s,a)] - \lambda H(\pi),$$

where d_π and d_E are the densities derived from the learning policy π and the expert policy π_E, and $H(\pi)$ is an entropy regularization term (Ziebart et al. 2008; Ziebart, Bagnell, and Dey 2010).

GAIL applies a saddle-point optimization strategy: it jointly trains a discriminator D and a policy π to optimize the following minimax objective:

$$\min_\pi \max_D \mathbb{E}_{d_\pi(s,a)}[\log(1 - D(s,a))] + \mathbb{E}_{d_E(s,a)}[\log(D(s,a))].$$

In practice, a fixed set of demonstrations from expert densities d_E are given, while samples from d_π are obtained by *on-policy* interactions with the environment.

Problem Setting

In this paper, we address the problem of learning in an MDP with *highly delayed* feedbacks. More concretely, in this MDP, an agent learns from the ***trajectory-wise*** reward r_e, which is only non-zero upon reaching an absorbing (terminal) state:

$$r_e(s_t, a_t, s_{t+1}) \neq 0 \Leftrightarrow s_t \notin \mathcal{S}_0, s_{t+1} \in \mathcal{S}_0.$$

Without losing clarity, we use $r_e(\tau)$ to denote the trajectory-wise reward obtained by a trajectory τ. For arbitrary two trajectories τ_i, τ_j, their relative ranking of r_e should align with the task objective:

Assumption 1 (Legitimacy of the Trajectory Rewards)**.** $\forall\, \tau_i, \tau_j, r_e(\tau_i) \geq r_e(\tau_j) \Longrightarrow P_{\pi^*}(\tau_i) \geq P_{\pi^*}(\tau_j)$, *where*

$$P_{\pi^*}(\tau) = \sum_{i=0}^{T} \Big(\gamma^i \log \pi^*(a_i|s_i) | \tau := \{(s_0,a_0),(s_1,a_2),\cdots,(s_T,a_T)\}\Big)$$

is the extent to which an oracle policy π^* agrees with a trajectory τ.

Our problem setting provides a generalized framework for a variety of prior arts, including preference-based RL (Fürnkranz et al. 2012; Wirth et al. 2017) and learning from human feedbacks (Mnih et al. 2015). Prior work of learning sparse-rewarded tasks with K-step feedbacks (Sun, Bagnell, and Boots 2018; Kang, Jie, and Feng 2018; Večerík et al. 2017; Jing et al. 2020) can also be reduced to our problem setting, with the advantage that their reward signals are finer-grained and more frequently provided. Compared with an elaborate reward function, trajectory-wise rewards are easier to access and more intuitive to human perception

(Christiano et al. 2017; Fürnkranz et al. 2012), which, however, makes regular RL more challenging.

To alleviate the learning difficulty, we assume that an agent learning a policy π is allowed to leverage external demonstrations $\mathcal{R}_T$ from an unknown teacher policy π_T, which are *sub-optimal* but more accessible than expert demonstrations. For the following of this paper, we use $d_\pi(s,a)$ and $d_T(s,a)$ to denote the density distribution derived from policy π and π_T, respectively. In practice, d_T is usually approximated from the demonstration data $\mathcal{R}_T$ (Ziebart et al. 2008; Fu, Luo, and Levine 2017; Ho and Ermon 2016). Moreover, the learning agent can also access its self-generated transitions cached in a replay buffer $\mathcal{R}_\mathbf{B}$, whose density distribution is $d_B(s,a)$.

Methodology

Exploration-Driven Objective

We propose an *exploration-driven* IL objective to learn from sub-optimal demonstrations, which is formulated as below:

Objective 1 (Exploration-Driven Imitation Learning)**.**

$$\max_\pi J(\pi) := \underbrace{-\mathbb{D}_{\mathbf{KL}}[d_\pi(s,a)||d_T(s,a)]}_{\textit{Imitation}} + \underbrace{\mathbb{D}_{\mathbf{KL}}[d_\pi(s,a)||d_B(s,a)]}_{\textit{Exploration}},$$

in which $\mathbb{D}_{KL}$ denotes the KL-divergence between two distributions:

$$\mathbb{D}_{\mathbf{KL}}[p||q] = \mathbb{E}_{p(x)} \log \frac{p(x)}{q(x)}.$$

Objective (1) can be interpreted as joint motivations for imitation and exploration. The first term $-\mathbb{D}_{\mathbf{KL}}[d_\pi(s,a)||d_T(s,a)]$ encourages distribution matching between $d_\pi(s,a)$ and $d_T(s,a)$. The second term $\mathbb{D}_{\mathbf{KL}}[d_\pi(s,a)||d_B(s,a)]$, though counter-intuitive at first sight, serves as an objective for self-exploration. Since $d_B(s,a)$ is the density derived from previously-learned policies, maximizing $\mathbb{D}_{\mathbf{KL}}[d_\pi(s,a)||d_B(s,a)]$ is in favor of visiting state-actions that are rarely seen by previously learned policies, which acts as a repulsive force from $d_B(s,a)$.

Specifically, the proposed objective encourages exploration, which is opposed to a conventional IL objective that solely pursues distribution matching between d_π and d_T:

$$\max_\pi J_{\text{IL}}(\pi) := -\mathbb{D}_{\mathbf{KL}}[d_\pi(s,a)||d_T(s,a)]. \tag{2}$$

An optimal solution to Eq (2) is a policy that exactly recovers the teacher's density distribution, with $d_\pi(s,a) = d_T(s,a)$ (Ziebart et al. 2008). Given this objective, π is restricted from further exploring density distributions that deviate from d_T, which impedes its potential of generating more superior trajectories. We will verify by empirical studies that optimizing Objective (1) achieves more efficient exploration compared with a pure imitation-driven objective.

Adaptive Learning Target

Following Objective (1), the learning policy has obtained the potential to yield trajectories with performance surpassing the teacher. To fully utilize this self-generated resource, *SAIL* adaptively adjusts the teacher's buffer to replace teacher demonstrations with more superior trajectories sampled from the learning agent, by leveraging the trajectory-wise feedback from the environment. This strategy dynamically improves the lower bound of the teacher's performance. As a result, the density $d_T(s,a)$ of the teacher buffer is approaching an oracle distribution:

Theorem 1. *For a deterministic policy, rewards of its generated trajectories indicate the policy' agreement with an oracle:*

$$\forall \pi_i, \pi_j,\ \mathbb{E}_{\tau\sim\pi_i}\left[r_e(\tau)\right] > \mathbb{E}_{\tau\sim\pi_j}\left[r_e(\tau)\right] \Longrightarrow \\ \mathbb{D}_{\mathbf{KL}}\left[\pi_i(a|s)||\pi^*(a|s)\right] < \mathbb{D}_{\mathbf{KL}}\left[\pi_j(a|s)||\pi^*(a|s)\right].$$

Therefore, when the teacher buffer is updated with more-superior trajectories generated by a deterministic policy over time, as in our case, the distribution derived by the teacher buffer is approaching to optimality. Unlike prior art that bundles their critic learning process with environment rewards, we leverage this delayed and coarse-grained feedback to construct a dynamic learning target with increasing superiority, which relieves the bottleneck brought by sub-optimal demonstrations.

Off-Policy Adversarial TD Learning

While our proposed approach is appealing in combining the merits of exploitation and exploration, it is challenging to directly optimize Objective (1). To make it more approachable, we draw a connection from Objective (1) to a conventional RL problem:

Remark 1. *Objective (1) can be rephrased as the following, which is equivalent to a max-return RL objective with* $\log \frac{d_T(s,a)}{d_B(s,a)}$ *in place of the environment rewards:*

$$\max_\pi\ J(\pi) := \mathbb{E}_{d_\pi(s,a)}\left[\log \frac{d_T(s,a)}{d_B(s,a)}\right], \tag{3}$$

One can consider the optimization of Equation (3) as a process of policy selection: for the *support* of (s,a) where the teacher has visited more frequently than the previously-learned policies, π is encouraged to build positive densities on those state-actions, leading to $d_\pi(s,a) > 0$ wherever $d_T(s,a) > d_B(s,a)$. Intuitively, this process implies that the agent trusts the teacher more than the previously learned policies.

Based on this insight, we can relate Objective (1) to Temporal-Difference (TD) learning, and solve it under an *actor-critic* framework. To obtain the reward function $\log \frac{d_T(s,a)}{d_B(s,a)}$, we build upon prior arts (Ho and Ermon 2016) to learn a discriminator D that optimizes the following:

$$\max_{D:\mathcal{S}\times\mathcal{A}\to(0,1)} \mathbb{E}_{d_B(s,a)}[\log(1-D(s,a))] + \mathbb{E}_{d_T(s,a)}[\log(D(s,a))]. \tag{4}$$

D aims to distinguish between the self-generated data from d_B and the teacher demonstrations from d_T. A well-learned discriminator shall satisfy the following (Goodfellow et al. 2014):

$$D^*(s,a) = \frac{d_T(s,a)}{d_T(s,a) + d_B(s,a)}.$$

The output of D with a constant shift, which we found to be more empirically effective, is used to render synthetic rewards to the agent:

$$r'(s,a) = -\log(1 - D(s,a)) \approx \log(\tfrac{d_T(s,a)}{d_B(s,a)} + 1).$$

In the initial training stage, a well-trained discriminator renders higher rewards to teacher demonstrations with $D(s,a) \to 1$, and lower rewards for self-generated samples with $D(s,a) \to 0$. The learning policy is therefore designed to confuse the discriminator by maximizing the shaped accumulated rewards.

To improve sample efficiency, we adopt an *off-policy* learning framework. Our objective is accordingly rephrased to maximize the expectation of Q-values over distributions of a behavior policy β (Silver et al. 2014; Lillicrap et al. 2016):

$$\max_\theta J_\beta(\pi_\theta) := \int_s d_\beta(s) Q(s, \pi_\theta(s)) d_s = \mathbb{E}_{d_\beta(s)}[Q(s, \pi_\theta(s))], \tag{5}$$

where $d_\beta(s)$ is the normalized stationary *state* distribution of β, analogous to the *state-action* distribution. The Q-function is a fixed point solution to the Bellman operation based on the shaped rewards:

$$Q(s,a) = r'(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{T}(s'|s,a), a' \sim \pi(s')}[Q(s', a')]. \tag{6}$$

Accordingly, the policy-gradient for the actor can be derived as (Lillicrap et al. 2016):

$$\nabla_\theta J_\beta(\pi_\theta) \approx \mathbb{E}_{s \sim d_\beta}[\nabla_\theta \pi_\theta(s) \nabla_a Q(s,a)|_{a=\pi_\theta(s)}]. \tag{7}$$

In the next section, we introduce an algorithm that realizes our objective via the abovementioned off-policy TD learning. It adopts an even more effective sampling approach that further accelerates the learning procedure.

Self-Adaptive Imitation Learning

Combing all the building blocks, we now introduce our approach, dubbed as *Self-Adaptive Imitation Learning (SAIL)*, as described in Algorithm 1. *SAIL* maintains two replay-buffers $\mathcal{R}_T$ and $\mathcal{R}_\text{B}$, for caching teacher demonstrations and self-generated transitions, respectively. It jointly learns three components: a discriminator D that serves as a reward provider, a critic Q that minimizes the Bellman error based on the shaped rewards, and an actor π that maximizes the shaped returns. During iterative training, high-quality trajectories generated by the actor are selected to refill the teacher demonstration buffer $\mathcal{R}_T$, while other trajectories are cached in the self-replay buffer $\mathcal{R}_\text{B}$. We highlight three key aspects of *SAIL*:

(1) *Leveraging delayed environment feedback to update teacher buffer $\mathcal{R}_T$*: High-quality trajectories with reward r_e above a threshold C_{d_T} are selected to update the teacher buffer $\mathcal{R}_T$. In practice, we use a window W_k to track the top K trajectory rewards of all trajectories added to the teacher buffer (*e.g.* K=5). Then we update $C_{d_T} = \min\{r_e(\tau_i) | r_e(\tau_i) \in W_k\}$, which is adaptively increasing to guarantee the improving quality of trajectories in the teacher's buffer.

Algorithm 1 Self-Adaptive Imitation Learning

Input: teacher replay buffer $\mathcal{R}_T$ with demonstrations, self-replay-buffer $\mathcal{R}_\text{B}$ with random transitions. policy π_θ, discriminator D_w, critic Q_ϕ, batch size $N > 0$, coefficient $\alpha > 0$, top-K trajectory window W_k

for $n = 1, \dots$ **do**
  sample trajectory $\tau \sim \pi_\theta$
  if $r_e(\tau) > C_{d_T}$ **then**
    $\mathcal{R}_T \leftarrow \mathcal{R}_T \cup \tau$; $\alpha \leftarrow 0$
    $C_{d_T} \leftarrow \min\{r_e(\tau_i) | r_e(\tau_i) \in W_k\}$
  else
    $\mathcal{R}_\text{B} \leftarrow \mathcal{R}_\text{B} \cup \tau$
  end if
  if $n \mod \texttt{discriminator-update} = 0$ **then**
    $\{(s_i, a_i, \cdots)\}_{i=1}^N \sim \mathcal{R}_\text{B}, \{(s_i^T, a_i^T, \cdots)\}_{i=1}^N \sim \mathcal{R}_T$
    update D_w by ascending gradient :
    $\nabla_w \frac{1}{N} \sum_{i=1}^N \big[\log D\Big(s_i^T, a_i^T\Big) + \log\Big(1 - D\big(s_i, a_i\big)\Big)\big]$
  end if
  if $n \mod \texttt{Q-update} = 0$ **then**
    $\{s_i, a_i, s_i'\}_{i=1}^N \sim \mathcal{R}_\text{B}, \{s_i^T, a_i^T, s_i^{'T}\}_{i=1}^N \sim \mathcal{R}_T$
    $y_i \leftarrow -\log(1 - D(s_i, a_i)) + \gamma \bar{Q}(s_i', \pi(s_i'))$
    $y_i^{'T} \leftarrow -\log(1 - D(s_i^T, a_i^T)) + \gamma \bar{Q}(s_i^{'T}, \pi(s_i^{'T}))$
    update Q_ϕ by minimizing critic loss:
    $J(Q_\phi) = \frac{1-\alpha}{N} \sum_i [(Q_\phi(s_i, a_i) - y_i)^2] + \frac{\alpha}{N} \sum_i [(Q_\phi(s_i^T, a_i^T) - y_i^{'T})^2]$
  end if
  if $n \mod \texttt{policy-update} = 0$ **then**
    $\{(s_i, ., ., ., .)\}_{i=1}^N \sim \mathcal{R}_\text{B}, \{(s_i^T, ., ., ., .)\}_{i=1}^N \sim \mathcal{R}_T$
    update π by sampled policy gradient:
    $\nabla_\theta J(\pi_\theta) \approx \frac{1-\alpha}{N} \sum_i [\nabla_\theta \pi_\theta(s_i) \nabla_a Q(s_i, a_i)|_{a_i = \pi_\theta(s_i)}] + \frac{\alpha}{N} \sum_i [\nabla_\theta \pi_\theta(s_i^T) \nabla_a Q(s_i^T, a_i^T)|_{a_i^T = \pi_\theta(s_i^T)}]$
  end if
end for

(2) *Realizing exploration-driven IL with an off-policy discriminator*: Prior art such as *GAIL* relies on *on-policy* training of a discriminator to estimate the ratio of $\frac{d_T(s,a)}{d_\pi(s,a)}$. On the contrary, we learn an *off-policy* discriminator D that aligns with our proposed objective and encourages efficient exploration, whose effectiveness will be elaborated in the Experiment section.

(3) *Sampling from teacher demonstrations for boosted learning efficiency*: In the initial learning step, we sample from both the teacher dataset $\mathcal{R}_T$ and the self-generated dataset $\mathcal{R}_\text{B}$ to construct a mixed density distribution, which plays the role of d_β in Eq (7). More concretely, we derive a mixture distribution: $d_\text{mix} = \alpha d_T + (1 - \alpha) d_B$, where α is the ratio of samples from teacher demonstrations. In practice, we initialize $\alpha = 0.5$. Once the learning policy generates trajectories with performance comparable to the teacher, we anneal the value of α to zero.

Reasoning of sampling from a mixture of distributions: Sampling from teacher demonstrations has been studied by other prior arts (Večerík et al. 2017; Reddy, Dragan, and Levine 2019). Along with the same spirit, the mixture sam-

pling distribution in our case accelerates the IL process by a *behavior-cloning* strategy. To see the rationale, one can rephrase the objective in Equation 5 as the following:

$$\max_\theta J_\beta(\pi_\theta) := \alpha \underbrace{\mathbb{E}_{d_T(s)}[Q(s, \pi_\theta(s))]}_{\text{Behavior Cloning}} + (1-\alpha)\mathbb{E}_{d_B(s)}\left[Q(s, \pi_\theta(s))\right].$$

In the early training stage, the discriminator will favor teacher trajectories by assigning them with highest rewards:

$$\mathbb{E}_{d_T(s)}[\max_a Q(s,a)] = \mathbb{E}_{d_T(s,a)}\left[Q(s,a)\right] \equiv \mathbb{E}_{d_T(s)}\left[Q(s, \pi_T(s))\right],$$

which encourages the learning policy to imitate π_T on teacher-visted states $d_T(s)$. We will verify by ablation study that sampling from teacher demonstrations accelerates the process of IL. Given a problem-setting with sub-optimal demonstrations, once the learning agent reaches the teacher-level performance, we relieve this behavior-cloning regularization by annealing α to zero, in order to reinforce the effects of exploration as proposed in our objective.

Related Work

Our work shares close connections with the following topics:

Imitation Learning (IL) aims to learn from expert demonstrations without accessing environment feedbacks, among which representative examples include ***GAIL*** (Ho and Ermon 2016) and its on-policy extensions (Kang, Jie, and Feng 2018; Wu et al. 2019; Fu, Luo, and Levine 2017). Later IL favors off-policy RL frameworks (Sasaki, Yohira, and Kawaguchi 2019; Kostrikov, Nachum, and Tompson 2020). Especially, ***DAC*** learns a discriminator by off-policy learning and corrects the distribution shifts by importance sampling (Kostrikov et al. 2019). In contrast to our approach, the above prior arts are motivated to exactly recover the teacher policy.

Our work also draws a subtle connection to *Self-Imitation Learning (SIL)* (Oh et al. 2018; Guo et al. 2018), in that they both utilize self-generated trajectories to build a learning target. However, SIL requires timely feedbacks from the environment to learn a delicate critic, which is in essence on-policy RL, while *SAIL* addresses a different setting by performing exploration-driven IL in an off-policy manner.

Learning from Demonstrations (LfD) facilitates RL by augmenting environment feedbacks with external demonstrations. Prior work relies on demonstrations that are sufficient and optimal (Hester et al. 2018; Večerík et al. 2017). Especially, ***DDPGfD*** leverages a DDPG framework (Lillicrap et al. 2016) to enable off-policy LfD in continuous spaces (Večerík et al. 2017). Later approaches, such as ***POfD*** (Kang, Jie, and Feng 2018), learn from *sub-optimal* demonstrations and trust the environment rewards to learn a critic, whereas demonstrations are only used as auxiliary guidance (Sun, Bagnell, and Boots 2018; Zhang et al. 2019; Gao et al. 2018). In contrast, our approach learns a critic without using environmental rewards, which is more robust especially when environment feedbacks are highly delayed. Some leverage the suboptimal guidance to enforce a policy regularization term, whose effects are gradually decayed to tackle the imperfect guidance (Jing et al. 2020). The above problem settings can be considered as relaxed versions of ours with finer-grained feedbacks.

Preference-based RL is a problem setting where the agent learns from the preference of an expert, which saves the necessity of designing elaborated numeric rewards (Weng 2011; Wirth et al. 2017). The preference relations can be over state-actions pairs (Fürnkranz et al. 2012) or over a pair of trajectories $\tau_i \succ \tau_j$ (Brown, Goo, and Niekum 2020), while the former provides more supervision information than the later. Few prior arts address IL in preference-based RL, except for (Brown, Goo, and Niekum 2020; Brown et al. 2020), which tackle IL in an MDP provided with only trajectory-ranked demonstrations but no environment feedbacks. Focusing on a different problem setting, *SAIL* utilizes the self-generated trajectories to build an increasing teacher distribution, which therefore requires fewer teacher demonstrations.

Exploration itself is an independent topic in RL. Classical exploration approaches work by involving randomness into its learning loop (Fortunato et al. 2017; Sutton 1990; Haarnoja et al. 2018). More recent approaches propose to use *intrinsic* rewards for exploration (Bellemare et al. 2016; Pathak et al. 2017). Especially, (Pathak et al. 2017) proposed ***curiosity-driven*** exploration, a model-based approach which leverages the prediction loss of a transition model as a reward bonus to encourage surprising behavior. Another exploration approach pursues a maximized *information gain* about the agent's belief of the environment (Houthooft et al. 2016). Readers are referred to (McFarlane 2018) for a comprehensive discussion on the exploration techniques in RL.

Experiments

In this section, we study how *SAIL* achieves the objective of imitation learning and exploration in an environment with delayed rewards. Extensive experiments have been conducted to answer the following key questions:

1. Is *SAIL* sample-efficient?
2. Can *SAIL* surpass the demonstration performance via off-policy exploration?
3. Which components in *SAIL* contribute to the exploration or sample efficiency?
4. Is *SAIL* robust against different sub-optimal teachers?

Setup

We built *SAIL* on a TD3 framework (Fujimoto, Van Hoof, and Meger 2018) based on *stable-baselines*[1] implementations. It is tested on 4 popular MuJoCo[2] tasks: *Walker2d-v2, Hopper-v2, HalfCheetah-v2*, and *Swimmer-v2*. For each task, we generate teacher demonstrations from a deterministic policy that was pre-trained to be sub-optimal. All experiments are conducted using one imperfect demonstration trajectory on five random seeds, with each trajectory containing no more than 1000 transitions. Models are evaluated after training using 10^6 interaction samples. We defer more details and additional experimental results to the Supplementary.

The original benchmarks are all in dense-reward settings. To construct the delayed rewarded environment as proposed

[1] https://stable-baselines.readthedocs.io/en/master/

[2] https://github.com/openai/mujoco-py

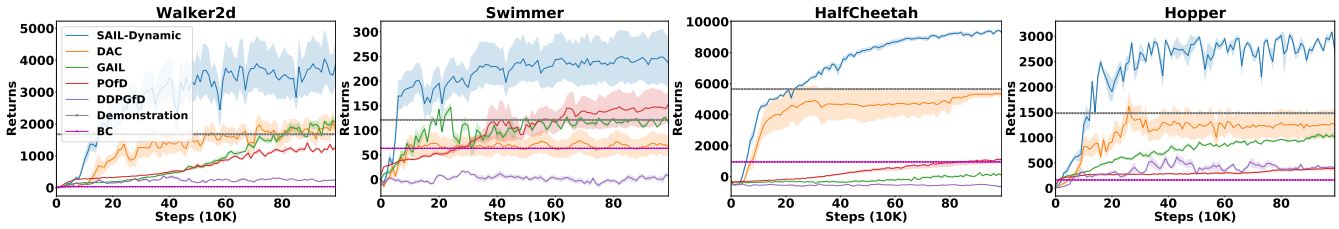

Figure 2: Learning curves of *SAIL* and other baselines using 1 suboptimal demonstration trajectory.

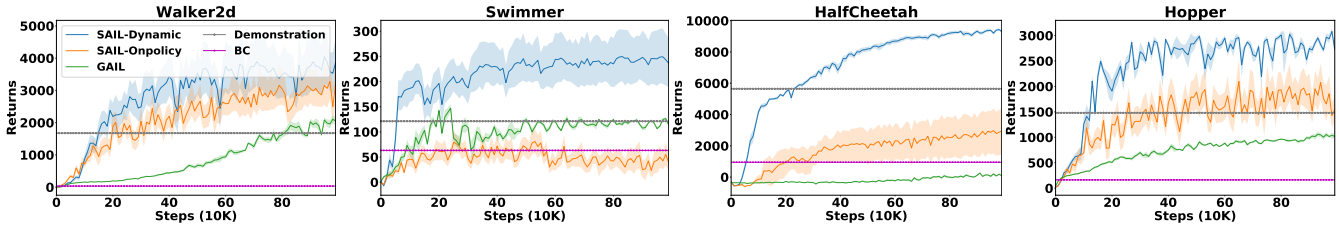

Figure 3: Comparing *SAIL* with its ***on-policy*** variant using 1 suboptimal demonstration trajectory.

Benchmark	HalfCheetah	Swimmer	Hopper	Walker
SAIL	**10660.59±105.53**	**309.47±3.0**	**3302.06±14.22**	**5868.53±108.82**
Curiosity Exploration	9043.07 ± 165.97	30.87 ± 6.52	3075.46 ± 15.61	5361.78 ± 58.24
Entropy Exploration	$8839.32 + 280.47$	65.04 ± 7.66	3079.29 ± 53.25	2792.76 ± 830.20
Teacher Demonstration	5646.71	121.16	1480.69	1675.01

Table 1: Off-policy exploration (*SAIL*) achieves higher performance than other exploration approaches.

in our paper, we omit the original rewards such that only episodic feedback is provided upon the completion of a trajectory.

To align with Assumption 1, we cache the original return of each trajectory $R(\tau) = \sum_i r(s_i, a_i)$, and downscale it to get a coarser grained supervision, with $r_e(\tau) = \lfloor 0.1 * R(\tau) \rfloor$.

We compare *SAIL* with 5 popular baselines that are mostly applicable to our problem setting: DAC, GAIL, POfD, DDPGfD, and BC, as discussed in the Related Work. For baselines that utilize environment rewards, such as POfD and DDPGfD, we provide them with modified rewards $r_e(s, a)$ upon the completion of each trajectory, instead of the original dense reward $r(s, a)$.

Performance on Continuous Action-Space Tasks

Sample efficiency: As the results are shown in Figure 2, *SAIL* is the only method that performs consistently better in all tasks in terms of both sample efficiency and asymptotic performance. At the initial stage of the learning, *SAIL* can quickly exploit the suboptimal demonstrations and approach to the demonstration's performance with significantly fewer samples.

Exploration ability: Besides sample efficiency, another advantage of *SAIL* is that it can effectively explore the environment to achieve expert-level performance, even with highly sparse rewards. We observe that prior solutions of learning from environment rewards for exploration, such as POfD and DDPGfD, cannot effectively address our proposed problem setting, as it is sample-costly to learn a meaningful critic from the delayed feedback. Unlike other imitation learning baselines whose performance is limited by the demonstrations, *SAIL* can rapidly surpass the imperfect teacher via constructing a better demonstration buffer.

Effects of Off-Policy Exploration in *SAIL*

Comparison with IL without Exploration: In order to illustrate the benefits of maximizing Objective (1) over a conventional *IL* objective, such as $\mathbb{D}_{\mathbf{KL}}[d_\pi(s, a)||d_T(s, a)]$, we conducted a comparison study where we trained the discriminator using teacher demonstrations τ_T and on-policy self-generated samples τ_π, instead of off-policy samples. This on-policy training scheme is the same as proposed in *GAIL* (Ho and Ermon 2016). In this way, the discriminator can get approximations of $\log(\frac{d_T}{d_\pi})$ instead of $\log(\frac{d_T}{d_B})$. We use the output of this on-policy discriminator to shape rewards, whereas Q and π are still updated in the same off-policy fashion as our proposed approach.

As illustrated in Figure 3, compared to *GAIL* (green) which is an on-policy baseline, *SAIL-OnPolicy* (orange) still enjoys the benefits of an off-policy learning scheme in general. However, it is less effective compared with our proposed approach. Even when π and Q are learned off-policy, *SAIL-OnPolicy* is slower to surpass the teacher demonstration (dashed gray line), due to its pure imitation-driven objective. *SAIL* enjoys fast improvement in performance not only because of an adaptive teacher demonstration buffer but also because it realizes the exploration-driven optimization.

Comparison with Other Exploration Approaches: We also compared *SAIL* with its two variants to evaluate the effects of different exploration approaches. In particular, we

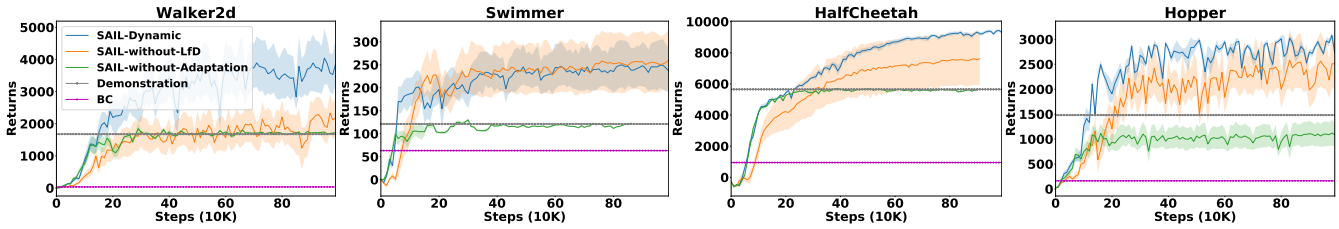

Figure 4: Ablation study by removing different algorithmic components from *SAIL*. Only one teacher trajectory is used as demonstration.

Benchmark	Evaluated Performance / Demonstration Performance		
HalfCheetah	10660.59 ± 105.53 / 5646.71	10217.00 ± 104.08 / 3598.90	9264.1 ± 163.45 / 875.39
Swimmer	309.47 ± 3.0 / 121.16	367.02 ± 1.11 / 46.82	361.17 ± 1.37 / 33.61
Hopper	3302.06 ± 14.22 / 1480.69	3814.07 ± 10.32 / 665.16	3589.92 ± 12.24 / 282.91
Walker	5868.53 ± 108.82 / 1675.01	4819.20 ± 1240.84 / 484.96	4574.61 ± 71.22 / 255.73

Table 2: Using 10^6 interaction samples, the performance of *SAIL* is robust regardless of the quality of sub-optimal teacher demonstrations.

integrated *SAIL* with a soft-actor-critic (Haarnoja et al. 2018) RL framework to enable an *entropy-based* exploration. For the other variant, we adopted the idea of random-distillation (Burda et al. 2019) to create a *curiosity* reward in addition to the reward provided by the discriminator. For both variant versions, we train the discriminator with *on-policy* samples in order to remove the effects of off-policy exploration. Comparison results in Table 1 indicate that, adopting an off-policy exploration approach (*SAIL*) is more effective given a fixed number of environment interactions. Entropy-based exploration is prone to high variance, while curiosity-based exploration, on the other hand, requires learning a forward transition model, and achieves lower final performance.

Ablation Study

We further evaluate *SAIL* by ablation and sensitivity studies to analyze the following aspects:

- ***Effects of learning from expert demonstrations***: As shown in Figure 4, we observed that sampling from a mixture of teacher data and self-generated data accelerates the learning performance in early training stages. Specifically, the *SAIL-Dynamic* (blue) refers our proposed approach, and *SAIL-without-LfD* (orange) only uses self-generated data to learn policy by setting $\alpha = 0$ constantly. We see that the *SAIL* is superior to *SAIL-without-LfD* in terms of initial performance, which is ascribed to a learning strategy resemblant to *behavior-cloning* when sampling from teacher demonstrations.
- ***Effects of updating teacher demonstration buffers***: As shown in Figure 4, *SAIL-without-Expert-Adaptation* (green) refers to a variant of *SAIL* which never update the teacher's replay buffer, even when a better trajectory is collected . We can observe that its asymptotic performance is bounded by the teacher's demonstration, which reveals the limitation of most existing IL approaches. One key insight from these results is that, instead of learning critics based on sparse rewards, leveraging the sparse guidance to improve the quality of the teacher can be much more effective in improving the ultimate performance.
- ***Robustness of SAIL on different teacher qualities:*** To evaluate the robustness of *SAIL* against different teacher performance, we pre-trained a group of teacher policies with varying qualities, ranging from near-randomness to sub-optimality, then used their generated trajectories as demonstrations. For each experiment, we only used one teacher trajectory as demonstration. Results in Table 2 show that *SAIL* can achieve robust performance no matter how sub-optimal the teacher behaves. Powered by both an exploration-driven objective and a self-adaptive learning strategy, *SAIL* can constantly explore with more superior trajectories to improve its learning target, which results in improving learning performance.

Conclusion

In this paper, we address the problem of reinforcement learning in environments with highly *delayed* rewards given *sub-optimal* demonstrations. To address this challenging problem, we propose a novel objective that encourages exploration-based imitation learning. Towards this objective, we design an effective algorithm called *Self Adaptive Imitation Learning (SAIL)*. The proposed approach is validated to (i) address sample efficiency by off-policy imitation learning, (ii) accelerate the IL learning process by fully utilizing teacher demonstration, and (iii) surpass the imperfect teacher with a large margin by iteratively performing imitation and exploration. Experimental results on challenging locomotion tasks indicate that *SAIL* significantly surpasses state-of-the-arts in terms of both sample efficiency and asymptotic performance.

Acknowledgements

This research was jointly supported by the National Institute on Aging 1RF1AG072449, the Office of Naval Research N00014-20-1-2382, and the National Science Foundation IIS-1749940.

References

Bellemare, M.; Srinivasan, S.; Ostrovski, G.; Schaul, T.; Saxton, D.; and Munos, R. 2016. Unifying count-based exploration and intrinsic motivation. In *Advances in neural information processing systems*, 1471–1479.

Brown, D. S.; Coleman, R.; Srinivasan, R.; and Niekum, S. 2020. Safe Imitation Learning via Fast Bayesian Reward Inference from Preferences. *ICML*.

Brown, D. S.; Goo, W.; and Niekum, S. 2020. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Conference on Robot Learning*, 330–359.

Burda, Y.; Edwards, H.; Storkey, A.; and Klimov, O. 2019. Exploration by random network distillation. *ICLR*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 4299–4307.

Fortunato, M.; Azar, M. G.; Piot, B.; Menick, J.; Osband, I.; Graves, A.; Mnih, V.; Munos, R.; Hassabis, D.; Pietquin, O.; et al. 2017. Noisy networks for exploration. *arXiv preprint arXiv:1706.10295*.

Fu, J.; Luo, K.; and Levine, S. 2017. Learning robust rewards with adversarial inverse reinforcement learning. *ICLR*.

Fujimoto, S.; Van Hoof, H.; and Meger, D. 2018. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*.

Fürnkranz, J.; Hüllermeier, E.; Cheng, W.; and Park, S.-H. 2012. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine learning*, 89(1-2): 123–156.

Gao, Y.; Xu, H.; Lin, J.; Yu, F.; Levine, S.; and Darrell, T. 2018. Reinforcement learning from imperfect demonstrations. *arXiv preprint arXiv:1802.05313*.

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

Guo, Y.; Oh, J.; Singh, S.; and Lee, H. 2018. Generative adversarial self-imitation learning. *arXiv preprint arXiv:1812.00950*.

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

Hester, T.; Vecerik, M.; Pietquin, O.; Lanctot, M.; Schaul, T.; Piot, B.; Horgan, D.; Quan, J.; Sendonaris, A.; Osband, I.; et al. 2018. Deep q-learning from demonstrations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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

Houthooft, R.; Chen, X.; Duan, Y.; Schulman, J.; De Turck, F.; and Abbeel, P. 2016. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, 1109–1117.

Jing, M.; Ma, X.; Huang, W.; Sun, F.; Yang, C.; Fang, B.; and Liu, H. 2020. Reinforcement learning from imperfect demonstrations under soft expert guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5109–5116.

Kang, B.; Jie, Z.; and Feng, J. 2018. Policy optimization with demonstrations. In *International Conference on Machine Learning*, 2474–2483.

Kiran, B. R.; Sobh, I.; Talpaert, V.; Mannion, P.; Sallab, A. A. A.; Yogamani, S.; and Pérez, P. 2020. Deep reinforcement learning for autonomous driving: A survey. *arXiv preprint arXiv:2002.00444*.

Kostrikov, I.; Agrawal, K. K.; Dwibedi, D.; Levine, S.; and Tompson, J. 2019. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. *ICLR*.

Kostrikov, I.; Nachum, O.; and Tompson, J. 2020. Imitation Learning via Off-Policy Distribution Matching. *ICLR*.

Kupcsik, A.; Hsu, D.; and Lee, W. S. 2018. Learning dynamic robot-to-human object handover from human feedback. In *Robotics research*, 161–176. Springer.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2016. Continuous control with deep reinforcement learning. *ICLR*.

McFarlane, R. 2018. A survey of exploration strategies in reinforcement learning. *McGill University, http://www. cs. mcgill. ca/cs526/roger. pdf, accessed: April*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529.

Oh, J.; Guo, Y.; Singh, S.; and Lee, H. 2018. Self-imitation learning. *arXiv preprint arXiv:1806.05635*.

Palan, M.; Landolfi, N. C.; Shevchuk, G.; and Sadigh, D. 2019. Learning reward functions by integrating human demonstrations and preferences. *arXiv preprint arXiv:1906.08928*.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 16–17.

Pham, H. X.; La, H. M.; Feil-Seifer, D.; and Nguyen, L. V. 2018. Autonomous uav navigation using reinforcement learning. *arXiv preprint arXiv:1801.05086*.

Reddy, S.; Dragan, A. D.; and Levine, S. 2019. SQIL: Imitation learning via reinforcement learning with sparse rewards. *arXiv preprint arXiv:1905.11108*.

Ross, S.; Gordon, G.; and Bagnell, D. 2011. 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*, 627–635.

Sasaki, F.; Yohira, T.; and Kawaguchi, A. 2019. Sample efficient imitation learning for continuous control. *ICLR*.

Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; and Riedmiller, M. 2014. Deterministic policy gradient algorithms.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *Nature*, 550(7676): 354–359.

Sun, W.; Bagnell, J. A.; and Boots, B. 2018. Truncated horizon policy search: Combining reinforcement learning & imitation learning. *ICLR*.

Sun, W.; Venkatraman, A.; Gordon, G. J.; Boots, B.; and Bagnell, J. A. 2017. Deeply aggrevated: Differentiable imitation learning for sequential prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3309–3318. JMLR. org.

Sutton, R. S. 1990. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*, 216–224. Elsevier.

Sutton, R. S.; and Barto, A. G. 2018. Reinforcement learning: An introduction. 45–47. MIT press.

Večerík, M.; Hester, T.; Scholz, J.; Wang, F.; Pietquin, O.; Piot, B.; Heess, N.; Rothörl, T.; Lampe, T.; and Riedmiller, M. 2017. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*.

Weng, P. 2011. Markov decision processes with ordinal rewards: Reference point-based preferences. In *Twenty-First International Conference on Automated Planning and Scheduling*.

Wirth, C.; Akrour, R.; Neumann, G.; and Fürnkranz, J. 2017. A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research*, 18(1): 4945–4990.

Wu, Y.-H.; Charoenphakdee, N.; Bao, H.; Tangkaratt, V.; and Sugiyama, M. 2019. Imitation Learning from Imperfect Demonstration. *ICML*.

Zhang, X.; Li, Y.; Ma, H.; and Luo, X. 2019. Pretrain Soft Q-Learning with Imperfect Demonstrations. *arXiv preprint arXiv:1905.03501*.

Ziebart, B. D.; Bagnell, J. A.; and Dey, A. K. 2010. Modeling interaction via the principle of maximum causal entropy.

Ziebart, B. D.; Maas, A. L.; Bagnell, J. A.; and Dey, A. K. 2008. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, 1433–1438. Chicago, IL, USA.