Adversarial Intrinsic Motivation for Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Learning with an objective to minimize the mismatch with a reference distribution has been shown to be useful for generative modeling and imitation learning. In this paper, we investigate whether one such objective, the Wasserstein-1 distance between a policy's state visitation distribution and a target distribution, can be utilized effectively for reinforcement learning (RL) tasks. Specifically, this paper focuses on goal-conditioned reinforcement learning where the idealized (unachievable) target distribution has full measure at the goal. We introduce a quasimetric specific to Markov Decision Processes (MDPs), and show that the policy that minimizes the Wasserstein-1 distance of its state visitation distribution to this target distribution under this quasimetric is the policy that reaches the goal in as few steps as possible. Our approach, termed Adversarial Intrinsic Motivation (AIM), estimates this Wasserstein-1 distance through its dual objective and uses it to compute a supplemental reward function. Our experiments show that this reward function changes smoothly with respect to transitions in the MDP and assists the agent in learning. Additionally, we combine AIM with Hindsight Experience Replay (HER) and show that the resulting algorithm accelerates learning significantly on several simulated robotics tasks when compared to HER with a sparse positive reward at the goal state.

1 Introduction

Reinforcement Learning (RL) [65] deals with the problem of learning a policy to accomplish a given task in an optimal manner. This task is typically communicated to the agent by means of a reward function. If the reward function is sparse [4] (e.g., most transitions yield a reward of 0), much random exploration might be needed before the agent experiences any signal relevant to learning [11, 2].

Some of the different ways to speed up reinforcement learning by modifying or augmenting the reward function are shaped rewards [47], redistributed rewards [2], intrinsic motivations [8, 60, 62, 63, 49, 52], and learned rewards [72, 49]. Unfortunately, the optimal policy under such modified rewards might sometimes be different than the optimal policy under the task reward [47, 16]. The problem of learning a reward signal that speeds up learning by communicating *what to do* but does not interfere by specifying *how to do it* is thus a useful and complex one [73].

This work considers whether a task-dependent reward function learned based on the distribution mismatch between the agent's state visitation distribution and a target task (expressed as a distribution) can guide the agent towards accomplishing this task. Adversarial methods to minimize distribution mismatch have been used with great success in generative modeling [26] and imitation learning [35, 22, 70, 67, 25]. In both these scenarios, the task is generally to minimize the mismatch with a target distribution induced by the data or expert demonstrations. Instead, we consider the task of goal-conditioned RL, where the ideal target distribution assigns full measure to a goal state. While

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

the agent can never match this idealized target distribution perfectly unless starting at the goal, intuitively, minimizing the mismatch with this distribution should lead to trajectories that maximize the proportion of the time spent at the goal, thereby prioritizing transitions essential to doing so.

The theory of optimal transport [69] gives us a way to measure the distance between two distributions (called the Wasserstein distance) even if they have disjoint support. Previous work [3, 29] has shown how a neural network approximating a potential function may be used to estimate the Wasserstein-1 distance using its dual formulation, but assumes that the metric space this distance is calculated on is Euclidean. A Euclidean metric might not be the appropriate metric to use in more general RL tasks however, such as navigating in a maze or environments where the state features change sharply with transitions in the environment.

This paper introduces a quasimetric tailored to Markov Decision Processes (MDPs), the time-step metric, to measure the Wasserstein distance between the agent's state visitation distribution and the idealized target distribution. While this time-step metric could be an informative reward on its own, estimating it is a problem as hard as policy evaluation [28]. Instead, we show that the dual objective, which maximizes difference in potentials while utilizing the structure of this quasimetric for the necessary regularization, can be optimized through sampled transitions.

We use this dual objective to estimate the Wasserstein-1 distance and propose a reward function based on this estimated distance. An agent that maximizes returns under this reward minimizes this Wasserstein-1 distance. The competing objectives of maximizing the difference in potentials for estimating the Wasserstein distance and minimizing it through reinforcement learning on the subsequent reward function leads to our algorithm, Adversarial Intrinsic Motivation (AIM).

Our analysis shows that if the above Wasserstein-1 distance is computed using the time-step metric, then minimizing it leads to a policy that reaches the goal in the minimum expected number of steps. It also shows that if the environment dynamics are deterministic, then this policy is the optimal policy.

In practice, minimizing the Wasserstein distance works well even when the environment dynamics are stochastic. Our experiments show that AIM learns a reward function that changes smoothly with transitions in the environment. We further conduct experiments on the family of goal-conditioned reinforcement learning problems [1, 56] and show that AIM when used along with hindsight experience replay (HER) greatly accelerates learning of an effective goal-conditioned policy compared to learning with HER and the sparse task reward. Further, our experiments show that this acceleration is similar to the acceleration observed by using the actual distance to the goal as a dense reward.

2 Related Work

We highlight the related work based on the various aspects of learning that this work touches, namely intrinsic motivation, goal-conditioned reinforcement learning, and adversarial imitation learning.

2.1 Intrinsic Motivation

Intrinsic motivations [8, 52, 51] are rewards presented by an agent to itself in addition to the external task-specific reward. Researchers have pointed out that such intrinsic motivations are a characteristic of naturally intelligent and curious agents [27, 5, 6]. Intrinsic motivation has been proposed as a way to encourage RL agents to learn skills [10, 9, 59, 55] that might be useful across a variety of tasks, or as a way to encourage exploration [11, 58, 7, 21]. The optimal reward framework [60, 63] and shaped rewards [47] (if generated by the agent itself) also consider intrinsic motivation as a way to assist an RL agent in learning the optimal policy for a given task. Such an intrinsically motivated reward signal has previously been learned through various methods such as evolutionary techniques [49, 57], meta-gradient approaches [62, 72, 73], and others. The Wasserstein distance has been used to present a valid reward for imitation learning [70, 17] as well as program synthesis [24].

2.2 Goal-Conditioned Reinforcement Learning

Goal-conditioned reinforcement learning can be considered a form of multi-task reinforcement learning [15] where the agent is given the goal state it needs to reach at the beginning of every episode, and the reward function is sparse with a non-zero reward only on reaching the goal state. UVFA [56], HER [1], and others [71, 18] consider this problem of reaching certain states in the

environment. Relevant to our work, Venkattaramanujam et al. [68] learns a distance between states using a random walk that is then used to shape rewards and speed up learning, but requires goals to be visited before the distance estimate is useful. DisCo RL [46] extends the idea of goal-conditioned RL to distribution-conditioned RL.

Contemporaneously, Eysenbach et al. [19, 20] has proposed a method which considers examples of success (similar to goals) and tries to predict and maximize the likelihood of seeing those examples under the current policy and trajectory. This method is a different approach to a problem similar to the one we look at, and their solution minimizes the Hellinger distance to the goal, a form of f-divergence. Theoretically, the Wasserstein distance is more informative when considering distributions that are disjoint, and does not require us to assume that the agent has already reached the goal through random exploration. AIM might in fact induce a form of directed exploration in order to reach the goal.

2.3 Adversarial Imitation Learning and Minimizing Distribution Mismatch

Adversarial imitation learning [35, 22, 70, 67, 25] has been shown to be an effective method to learn agent policies that minimize distribution mismatch between an agent's state-action visitation distribution and the state-action visitation distribution induced by an expert's trajectories. In most cases this distribution that the expert induces is achievable by the agent and hence these techniques aim to match the expert distribution exactly. In the context of goal-conditioned reinforcement learning, GoalGAIL [18] uses adversarial imitation learning with a few expert demonstrations to accelerate the learning of a goal-conditioned policy. In this work, we focus on unrealizable target distributions that cannot be completely matched by the agent, and indeed, are not induced by any trajectory distribution.

FAIRL [25] is an adversarial imitation learning technique which minimizes the Forward KL divergence and has been shown experimentally to cover some hand-specified state distributions, given a smoothness regularization as used by WGAN [29]. f-IRL [48] learns a reward function where the optimal policy matches the expert distribution under the more general family of f-divergences. Further, techniques beyond imitation learning [41, 33] have looked at matching a uniform distribution over states to guarantee efficient exploration.

3 Background

In this section we first set up the goal-conditioned reinforcement learning problem, and then give a brief overview of optimal transport.

3.1 Goal-Conditioned Reinforcement Learning

Consider a goal-conditioned MDP as the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{G}, P, \rho_0, \sigma, \gamma \rangle$ with discrete state space $\mathcal{S}$, discrete action space $\mathcal{A}$, a subset of states which is the goal set $\mathcal{G} \subseteq \mathcal{S}$, and transition dynamics $P : \mathcal{S} \times \mathcal{A} \times \mathcal{G} \longmapsto \Delta(\mathcal{S})$ ($\Delta(\cdot)$ is a distribution over a set) which might vary based on the goal (see below). $\rho_0 : \Delta(\mathcal{S})$ is the starting state distribution, and $\sigma : \Delta(\mathcal{G})$ is the distribution a goal is drawn from. $\gamma \in [0, 1)$ is the discount factor. We use discrete states and actions for ease of exposition, but our idea extends to continuous states and actions, as seen in the experiments.

At the beginning of an episode, the starting state is drawn from ρ_0 and the goal for that episode is drawn from σ. The reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{G} \longmapsto \mathbb{R}$ is deterministic, and $r(s_t, a_t, s_{t+1}|s_g) := \mathbb{I}[s_{t+1} = s_g]$. That is, there is a positive reward when an agent reaches the goal ($s_{t+1} = s_g$), and 0 everywhere else. Since the goal is given to the agent at the beginning of the episode, in goal-conditioned RL the agent knows what this task reward function is (unlike the more general RL problem). The transition dynamics are goal-conditioned as well, with an automatic transition to an absorbing state $\bar{s}$ on reaching the goal s_g and then staying in that state with no rewards thereafter ($P(\bar{s}|s_g, a, s_g) = 1 \ \forall \ a \in \mathcal{A}$ and $P(\bar{s}|\bar{s}, a, s_g) = 1 \ \forall \ a \in \mathcal{A}$). In short, the episode terminates on reaching the goal state.

The agent takes actions in this environment based on a policy $\pi \in \Pi : \mathcal{S} \times \mathcal{G} \longmapsto \Delta(\mathcal{A})$. The return H_g for an episode with goal s_g is the discounted cumulative reward over that episode $H_g = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}|s_g)$, where $s_0 \sim \rho_0$, $a_t \sim \pi(\cdot|s_t, s_g)$, and $s_{t+1} \sim P(\cdot|s_t, a_t, s_g)$. The agent aims to find the policy $\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{g \in \mathcal{G}} \mathbb{E}_{s_0 \sim \rho_0} \mathbb{E}_{\pi}[H_g]$ that maximizes the expected returns in this goal-conditioned MDP. For a policy π, the agent's goal-conditioned state distribution

$\rho_\pi(s|s_g) = \mathbb{E}_{s_0 \sim \rho_0}[(1-\gamma)\sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi, s_g)]$. Overloading the terminology a bit, we also define the goal-conditioned target distribution $\rho_g(s|s_g) = \delta(s_g)$, a Dirac measure at the goal state s_g.

While learning using traditional RL paradigms is possible in goal-conditioned RL, there has also been previous work (Section 2.2) on leveraging the structure of the problem across goals. Hindsight Experience Replay (HER) [1] attempts to speed up learning in this sparse reward setting by taking episodes of agent interactions, where they might not have reached the goal specified for that episode, and relabeling the transitions with the goals that *were* achieved during the episode. Off-policy learning algorithms are then used to learn from this relabeled experience.

3.2 Optimal Transport and Wasserstein-1 Distance

The theory of optimal transport [69, 13] considers the question of how much work must be done to transport one distribution to another optimally. More concretely, suppose we have a metric space ($\mathcal{M}$, d) where $\mathcal{M}$ is a set and d is a metric on $\mathcal{M}$ (Definitions in Appendix A). For two distributions μ and ν with finite moments on the set $\mathcal{M}$, the Wasserstein-p distance is denoted by:

$$W_p(\mu, \nu) := \inf_{\zeta \in Z(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\zeta}[d(X,Y)^p]^{1/p} \tag{1}$$

where Z is the space of all possible couplings, i.e., distributions $\zeta \in \Delta(\mathcal{M} \times \mathcal{M})$ whose marginals are μ and ν respectively. Finding this optimal coupling tells us what is the least amount of work, as measured by d, that needs to be done to convert μ to ν. This Wasserstein-p distance can then be used as a cost function (negative reward) by an RL agent to match a given target distribution [70, 17, 24].

Finding the ideal coupling above is generally considered intractable. However, if what we need is an accurate estimate of the Wasserstein distance and not the optimal transport plan we can turn our attention to the dual form of this distance. The Kantorovich-Rubinstein duality [69, 53] for the Wasserstein-1 distance (which we refer to simply as the Wasserstein distance hereafter) on a ground metric d is of particular interest and gives us the following:

$$W_1(\mu, \nu) = \sup_{\text{Lip}(f) \leq 1} \mathbb{E}_{y\sim\nu}[f(y)] - \mathbb{E}_{x\sim\mu}[f(x)] \tag{2}$$

where the supremum is over all 1-Lipschitz functions $f : \mathcal{M} \longmapsto \mathbb{R}$ in the metric space. Importantly, Jevtić [37] has recently shown that this dual formulation extends to quasimetric spaces as well. More details, such as the definition of the Lipschitz constant of a function and special cases used in WGAN [3, 29, 25] are elaborated in Appendix B.

4 Time-Step Metric

The choice of the ground metric d is important when computing the Wasserstein distance between two distributions. This ground metric should specify how much work is needed to transport a unit of measure from one point to another. That is, if we want the Wasserstein distance to give an estimate of the work needed to transport the agent's state visitation distribution to the goal state, the underlying metric should incorporate a notion of this work.

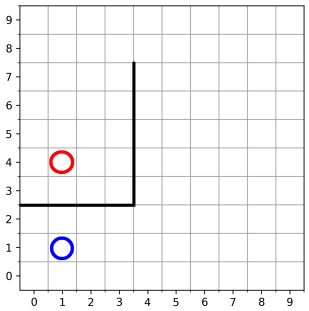


Figure 1: Grid world example

Consider the grid world shown in Figure 1, where a wall (bold line) marks an impassable barrier in part of the state space. If the states are specified by their Cartesian coordinates on the grid, the Euclidean distance between the states specified by the blue and red circles is not a good measure of the number of steps it will take an agent to go from one to the other. This mismatch would lead to an underestimation of the work involved if the two distributions we wanted to compare were concentrated at those two circles. Similarly, we will get errors in estimating the Wasserstein distance if the grid world is toroidal (where an agent is transported to the opposite side of the grid if it walks off one side) or if the transitions are asymmetric (windy grid world [65]).

To estimate the work needed to transport measure in an MDP when executing a policy π, we consider a *quasimetric* – a metric that does not need to be symmetric – dependent on the transitions experienced when executing that policy and the cost of those transitions.

Definition 1. *A transition cost metric d_c^π in an MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, transition function P, agent policy π, and a per transition cost function $c : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \longmapsto \mathbb{R}^+ \cup \{\infty\}$ is a quasimetric where the distance from state $s \in \mathcal{S}$ to state $s_g \in \mathcal{S}$ is based on the expected transitions under policy π and the cost c of those transitions.*

$$d_c^\pi(s, s_g) := \mathop{\mathbb{E}}_{\pi, P}\left[\sum_{t=0}^{\infty} c(s_t, a_t, s_{t+1}|s_g)|s_0 = s\right]$$

and the cost of transitioning from the target state s_g to the absorbing state $\bar{s}$, $c(s_g, a, \bar{s}|s_g) = 0 \; \forall \; a \in \mathcal{A}$, and from the absorbing state $\bar{s}$ to itself, $c(\bar{s}, a, \bar{s}|s_g) = 0 \; \forall \; a \in \mathcal{A}$.

An example of such a quasimetric is the cost of moving a robotic arm from one position to another where the action at each step has a variable cost based on the current required by the motors. In this paper we use the expected number of steps to reach some state $s_g \in \mathcal{S}$ from some other state $s \in \mathcal{S}$, which is defined as the **time-step metric** d_T^π.

$$d_T^\pi(s, s_g) := \mathbb{E}\;\left[T(s_g|\pi, s)\right]$$

where $T(s_g|\pi, s)$ is the random variable for the first time-step that state s_g is encountered by the agent after starting in state s and following policy π. This quasimetric has the property that the per step cost is uniformly 1 for all transitions except ones from the goal to the absorbing state (and the absorbing state to itself), which are 0. Thus, it can be written recursively as:

$$d_T^\pi(s, s_g) = \begin{cases} 0 & \text{if } s = s_g \\ 1 + \mathbb{E}_{a\sim\pi(\cdot|s,s_g)}\, \mathbb{E}_{s'\sim P(\cdot|s,a,s_g)}\left[d_T^\pi(s', s_g)\right] & \text{otherwise} \end{cases} \tag{3}$$

Recall that the potential function $f : \mathcal{S} \longmapsto \mathbb{R}$ needs to be 1-Lipschitz with respect to this quasimetric d_T^π so that we can estimate the Wasserstein distance using the dual (Equation 2). In Appendix C we prove that we can ensure L-Lipschitz continuity by enforcing that the difference in values of f on expected transitions from every state are bounded by L, implying

$$\text{Lip}(f) \leq \sup_{s\in\mathcal{S}}\left\{\mathop{\mathbb{E}}_{a\sim\pi(\cdot|s,s_g)}\;\mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a,s_g)}\left[|f(s) - f(s')|\right]\right\}. \tag{4}$$

We remark that finding a proper way to enforce the Lipschitz constraint in adversarial methods remains an open problem [44]. However, for the time-step metric considered here, equation 4 is one elegant way of doing so. By ensuring that the Kantorovich potentials do not drift too far from each other on expected transitions under agent policy π in the MDP, we can ensure that the potential function maintains the conditions necessary to estimate the Wasserstein distance [69, 3]. Finally, the minimum distance $d_T^\blacklozenge$ from state s to a given goal state s_g (corresponding to policy $\pi^\blacklozenge$) is defined by the Bellman optimality condition (Equation 16 in Appendix D).

Consider how the time-step distance to the goal and the value function for goal-conditioned RL relate to each other. When the reward is 0 everywhere except for transitions to the goal state, the value becomes $V^\pi(s|s_g) = \mathbb{E}\left[\gamma^{T(s_g|\pi,s)}\right]$. $d_T^\pi(s_0, s_g)$ and $V(s_0|s_g)$ are related as follows.

Proposition 1. *A lower bound on the value of any state under a policy π can be expressed in terms of the time-step distance from that state to the goal: $V(s_0|s_g) \geq \gamma^{d_T^\pi(s_0,s_g)}$.*

The proofs for all theoretical results are in Appendix D. The Jensen gap $\Delta_{\text{Jensen}}^\pi(s) := V^\pi(s|s_g) - \gamma^{d_T^\pi(s,s_g)}$ describes the sharpness of the lower bound in the proposition above and it is zero if and only if $\text{Var}(T(s_g|\pi, s)) = 0$ [42]. From this line of reasoning, we can deduce the following proposition:

Proposition 2. *If the transition dynamics are deterministic, the policy that maximizes expected return is the policy that minimizes the time-step metric ($\pi^* = \pi^\blacklozenge$).*

5 Wasserstein-1 Distance for Goal-Conditioned Reinforcement Learning

In this section we consider the problem of goal-conditioned reinforcement learning. In Section 5.1 we analyze the Wasserstein distance computed under the time-step metric d_T^π. Section 5.2 proposes an algorithm, Adversarial Intrinsic Motivation (AIM), to learn the potential function for the Kantorovich-Rubinstein duality used to estimate the Wasserstein distance, and giving an intrinsic reward function used to update the agent policy in tandem.

5.1 Wasserstein-1 Distance under the Time-Step Metric

From Sections 3.2 and 4 the Wasserstein distance under the time-step metric d_T^π of an agent policy π with visitation measure ρ_π to a particular goal s_g and its distribution ρ_g can be expressed as:

$$W_1^\pi(\rho_\pi, \rho_g) = \sum_{s \in \mathcal{S}} \rho_\pi(s|s_g) d_T^\pi(s, s_g) \tag{5}$$

where W_1^π refers to the Wasserstein distance with the ground metric d_T^π.

The following proposition shows that the Wasserstein distance decreases as $d_T^\pi(s, s_g)$ decreases, while also revealing a surprising connection with the Jensen gap.

Proposition 3. *For a given policy π, the Wasserstein distance of the state visitation measure of that policy from the goal state distribution ρ_g under the ground metric d_T^π can be written as*

$$W_1^\pi(\rho_\pi, \rho_g) = \mathop{\mathbb{E}}_{s_0 \sim \rho_0} \left[h(d_T^\pi(s_0, s_g)) + \frac{\gamma}{1-\gamma} (\Delta_{Jensen}^\pi(s_0) - 1) \right] \tag{6}$$

where h is an increasing function of d_T^π.

The first component in the above analytical expression shows that the Wasserstein distance depends on the expected number of steps, decreasing if the expected distance decreases. The second component shows the risk-averse nature of the Wasserstein distance. Concretely, the bounds for the Jensen inequality given by Liao and Berg [42] imply that there are non-negative constants $C_1 = C_1(d_T^\pi, \gamma)$ and $C_2 = C_2(d_T^\pi, \gamma)$ depending only on the expected distance and discount factor such that

$$C_1 \mathrm{Var}(T(s_g|\pi, s)) \leq \Delta_{\text{Jensen}}^\pi(s) \leq C_2 \mathrm{Var}(T(s_g|\pi, s)).$$

From the above, we can deduce that a policy with lower variance will have lower Wasserstein distance when compared to a policy with same expected distance from the start but higher variance. The relation between the optimal policy in goal-conditioned RL and the Wasserstein distance can be made concrete if we consider deterministic dynamics.

Theorem 1. *If the transition dynamics are deterministic, the policy that minimizes the Wasserstein distance over the time-step metrics in a goal-conditioned MDP (see equation 5) is the optimal policy.*

5.2 Adversarial Intrinsic Motivation to minimize Wasserstein-1 Distance

The above section makes it clear that minimizing the Wasserstein distance to the goal will lead to a policy that reaches the goal in as few steps as possible in expectation. If the dynamics of the MDP are deterministic, this policy will also be optimal. We now consider how to go about minimizing this distance. Note that the dual form (Equation 2) can be used to estimate the distance, *even if the ground metric d_T^π is not known*. The potential function f in the dual needs to be 1-Lipschitz ($\mathrm{Lip}(f) \leq 1$), which can be ensured using the constraint we get from Equation 4. This constraint needs to be enforced for all states and subsequent transitions expected under the agent policy.

We now consider the full problem. The reinforcement learning algorithm aims to learn a goal-conditioned policy with parameters $\theta \in \Theta$ whose state visitation distribution ρ_θ minimizes the Wasserstein distance to a goal-conditioned target distribution ρ_g for a given goal $s_g \sim \sigma$. We leverage the presence of the set of goals that the agent should be capable of reaching with a goal-conditioned potential function $f_\phi : \mathcal{S} \times \mathcal{G} \longmapsto \mathbb{R}$ with parameters $\phi \in \Phi$. These objectives of the potential function and the agent can be expressed together using the following adversarial objective:

$$\min_{\theta \in \Theta} \max_{\phi \in \Phi} \mathop{\mathbb{E}}_{s_g \sim \sigma} \left[f_\phi(s_g, s_g) - \mathop{\mathbb{E}}_{s \sim \rho_\theta} [f_\phi(s, s_g)] \right] \tag{7}$$

where the potential function f_ϕ is 1-Lipschitz over the state space. Combining the objectives in Equations 7 and 4, the loss for the potential function f_ϕ then becomes:

$$\begin{aligned} L_f := & \mathop{\mathbb{E}}_{s_g \sim \sigma} \left[-f_\phi(s_g, s_g) + \mathop{\mathbb{E}}_{s \sim \rho_\theta} [f_\phi(s, s_g)] \right] + \\ & \lambda \mathop{\mathbb{E}}_{(s,a,s',s_g) \sim \mathcal{D}} \left[(\max(|f_\phi(s, s_g) - f_\phi(s', s_g)| - 1, 0))^2 \right] \end{aligned} \tag{8}$$

Where the distribution $\mathcal{D}$ should ideally contain all states in $\mathcal{S}$, expected goals in $\mathcal{G}$, and the transitions according to the agent policy π_θ and transition function P. Such a distribution is difficult to

obtain directly. We approximate it with a small replay buffer of transitions from recent episodes experienced by the agent, and relabel these episodes with achieved goals (similar to HER [1]). Such an approximation does not respect the discounted measure of states later on in an episode, but is consistent with how other approaches in deep reinforcement learning tend to approximate the state visitation distribution, especially for policy gradient approaches [50]. While it does not include all states and all goals, we see empirically that the above approximation works well.

Now we turn to the reward function that should be presented to the agent such that maximizing the return will minimize the Wasserstein distance. The Wasserstein discriminator is a potential function [47] (its value depends on the state). It can thus be used to create a shaped reward $\hat{r}(s, a, s', s_g) = r(s, a, s'|s_g) + \gamma f_\phi(s', s_g) - f_\phi(s, s_g)$ without risk of changing the optimal policy. Alternatively, we can explicitly minimize samples of the Wasserstein distance: $\hat{r}(s, a, s', s_g) = f_\phi(s', s_g) - f_\phi(s_g, s_g)$. Finally, instead of the second term $f_\phi(s_g, s_g)$, we can just use a constant bias term. In practice, all these choices work well, and we use the latter (with $b = \max_{s \in \mathcal{S}} f_\phi(s, s_g)$) to reduce variance in $\hat{r}$.

$$\hat{r}(s, a, s', s_g) = f_\phi(s', s_g) - b \tag{9}$$

We lay out the basic procedure to learn and use adversarial intrinsic motivation (AIM) in Algorithm 1, and also include how to use this algorithm in conjunction with HER. If not using HER, Line 23 where hindsight goals are added to the replay buffer can be skipped.

6 Experiments

In our experiments, we evaluate the extent to which the reward learned through AIM is useful as a proxy for the environment reward signal, or in tandem with the environment reward signal. In particular, we ask the following questions:

- Does AIM speed up learning of a policy to get to a single goal compared to learning with a sparse reward?
- Does the learned reward function qualitatively guide the agent to the goal?
- Does AIM work well with stochastic transition dynamics or sharp changes in the state features?
- Does AIM generalize to a large set of goals and continuous state and action spaces?

Our experiments suggest that the answer to all 4 questions is "yes", with the first three questions tested in the grid world presented in Figure 1 where the goal is within a room, and the agent has to go around the room from its start state to reach the goal. Goal-conditioned tasks in the established Fetch robot domain show that AIM also accelerates learning across multiple goals in continuous state and action spaces.

Grid World In this task, the goal is inside a room and the agent's starting position is such that it needs to navigate around the room to find the doorway and be able to reach the goal. The agent can move in the 4 cardinal directions unless blocked by a wall or the edge of the grid. The agent policy is learned using maximum entropy Q-learning [30] with no hindsight goals used for this experiment.

The agent's state visitation distribution after 100 training iterations when using AIM-learned rewards is shown in Figure 2a and the learned rewards for each state are plotted in Figure 2b. The state visitation after 500 training iterations when learning with the true task reward shows that the agent is unable to learn a policy that takes it to the goal (Figure 2c). These figures show that AIM enables the agent to reach the goal and learn the required policy quickly, while learning with the sparse task reward fails to do so, even with five times more training.

While the theoretical results presented in Section 5 can only guarantee that the optimal policy is the minimizer of the Wasserstein distance for the case of deterministic dynamics, additional experiments in the grid world given in Appendix F show that AIM also speeds up learning in the presence of stochastic transitions. Further, they confirm that the AIM learned reward function is smooth with respect to transitions even if the state features change sharply. This property is in contrast to the gradient penalty based regularization popular in Wasserstein GANs [29] and adversarial imitation learning [25] approaches, which cannot handle transitions where the state features change sharply.

Fetch Robot The generalization capability of AIM across multiple goals in goal-conditioned RL tasks with continuous states and actions is tested in the MuJoCo simulator [66], on the Fetch robot

tasks from OpenAI gym [14] which have been used for evaluation previously [1, 71]. Descriptions of these tasks and their goal space is in Appendix H. We soften the Dirac target distribution for continuous states to instead be a Gaussian with variance of 0.01 of the range of each feature.

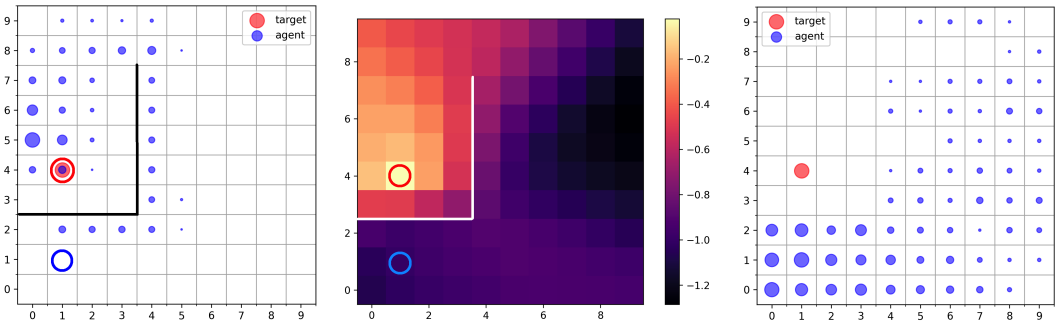
The goals in this setting are not the full state, but rather the dimensions of factored states relevant to the given goal. The task wrapper additionally returns the goal achieved by the agent, and so we can use this reduced goal space to learn our reward function, rather than the full state space. It is unclear how this smaller goal space might affect AIM. While the smaller goal space might make learning easier for potential function f_ϕ, the partially observable nature of the goals might lead to a less informative reward.

We compare HER (refer subsection 3.1) with the sparse environment reward (HER + R), HER with AIM rewards (HER + AIM), and HER with AIM rewards and environment rewards (HER + AIM + R). The negative Euclidean distance between the achieved and target goals is an appropriate dense and informative reward for these tasks. We therefore additionally train an agent using this dense reward (HER + dense). We used the HER implementation using Twin Delayed DDPG (TD3) [23] as the underlying RL algorithm from the stable baselines repository [34]. We did an extensive sweep of the hyperparameters for the baseline HER + R (laid out in Appendix H), with an additional coarser search on relevant hyperparameters for AIM.

Figure 3 shows that using the AIM-learned reward speeds up learning in three of the four Fetch domains, even without the environment reward. This improvement is very close to what we would see if we used the dense reward (based on the actual distance function in this domain). Results are averaged across the 6 different seeds, with shaded regions showing standard error across runs. Statistical analysis using a mixed effects ANOVA and a Tukey test at a significance level of 95% (more detail in Appendix G) show that in three of the four environments AIM and AIM+ R have similar odds of reaching the goal as the dense shaped reward, and in all four environments AIM and AIM+ R have higher odds of reaching the goal as the sparse reward.

FAIRL [25] (which has been shown to learn policies that cover hand-specified state distributions) was also applied on these 4 domains but it failed to learn at all. Interestingly, scaling the reward such that it is always negative led to similar performance to (but not better than) AIM. We hypothesize that FAIRL, as defined and presented, fails in these domains because the environments are episodic, and the episode ends earlier if the goal is reached. Since the FAIRL reward is positive closer to the target distribution, the agent can get close to the target, but refrain from reaching it (and ending the episode) to collect additional positive reward.

The domain where AIM does not seem to have a large advantage (Slide) is one where the agent strikes an object initially and that object has to come to rest near the goal. Since the credit assignment has to propagate back to the single point when the agent strikes the object regardless of how dense the subsequent reward is, AIM does not add much value to this task.



(a) Agent state visitation learning with AIM rewards (100 iterations)

(b) Learned Reward (100 training iterations)

(c) Agent state visitation learning with task reward (500 iterations)

Figure 2: Grid world experiments. Agent's undiscounted state visitation (2a, 2c): Blue circle indicates agent's start state. Red circle is the goal. Blue bubbles indicate relative time agent's policy causes it to spend in respective states. Learned reward function (2b): AIM reward at each state of the grid world. Black (or white) lines indicate walls the agent cannot transition through.

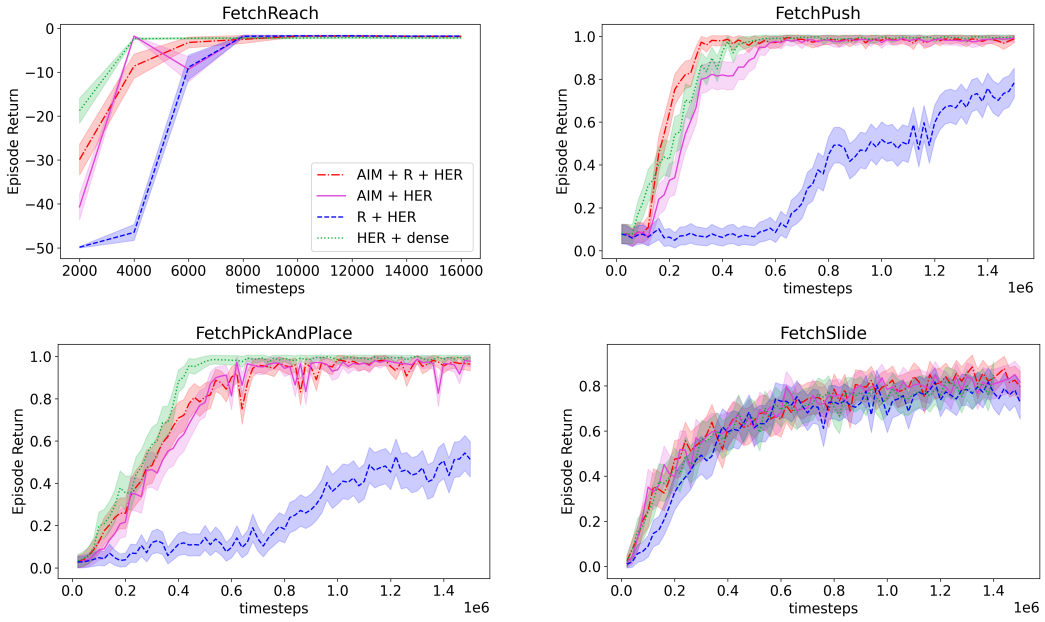


Figure 3: Evaluating AIM with HER on some goal-conditioned RL tasks. AIM learns the reward function in tandem with the policy updates. The "dense" reward is the ideal negative distance to goal.

7 Discussion and Future Work

Approaches for estimating the Wasserstein distance to a target distribution by considering the dual of the Kantorovich relaxation have been previously proposed [3, 29, 70]. We improve upon them by choosing a metric space more suited to the MDP and notions of optimality in the MDP. This choice allows us to leverage the structure introduced by the dynamics of the MDP to regularize the Kantorovich potential using a novel objective.

Previous work [12] has pointed out that the gradients from sample estimates of the Wasserstein distance might be biased. This issue is mitigated in our implementation through multiple updates of the discriminator, which they found to be empirically useful in reducing the bias. Additionally, recent work has pointed out that the discriminator in WGAN might be bad at estimating the Wasserstein distance [64]. While our experiments indicate that the potential function in AIM is learned appropriately, future work could look more deeply to verify possible inefficiencies in this estimation.

The process of learning the Wasserstein distance through samples of the environment while simultaneously estimating the cost of the full path is reminiscent of the A^* algorithm [31], where the optimistic heuristic encourages the agent to explore in a directed manner, and adjusts its estimates based on these explorations.

In the goal-conditioned RL setting, AIM seems to be an effective intrinsic reward that balances exploration and exploitation for the task at hand. The next step is to consider whether the Wasserstein distance can be estimated similarly for more general tasks, and whether minimizing this distance in those tasks leads to the optimal policy. A different potential avenue for future work is the problem of more general exploration [33, 41] by specifying a uniform distribution as the target, or using this directed exploration as an intermediate step for efficient exploration [38].

Finally, reward design is an important aspect of practical reinforcement learning. Not only do properly shaped reward speed up learning [47], but reward design can also subtly influence the kinds of behaviors deemed acceptable for the RL agent [40] and could be a potential safety issue keeping reinforcement learning from being deployed on real world problems. Learning-based approaches that can assist in specifying reward functions safely given alternative approaches for communicating the task could be of value in such a process of reward design, and an avenue for future research.

References

[1] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight Experience Replay. *arXiv:1707.01495 [cs]*, February 2018. URL http://arxiv.org/abs/1707.01495. arXiv: 1707.01495.

[2] Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. In *Advances in Neural Information Processing Systems*, pages 13566–13577, 2019.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning*, pages 214–223, July 2017. URL http://proceedings.mlr.press/v70/arjovsky17a.html. ISSN: 2640-3498 Section: Machine Learning.

[4] Dilip Arumugam, P. Henderson, and P. Bacon. An information-theoretic perspective on credit assignment in reinforcement learning. *ArXiv*, abs/2103.06224, 2021.

[5] Gianluca Baldassarre. What are intrinsic motivations? a biological perspective. In *2011 IEEE international conference on development and learning (ICDL)*, volume 2, pages 1–8. IEEE, 2011.

[6] Gianluca Baldassarre, Tom Stafford, Marco Mirolli, Peter Redgrave, Richard M Ryan, and Andrew Barto. Intrinsic motivations and open-ended development in animals, humans, and robots: an overview. *Frontiers in psychology*, 5:985, 2014.

[7] Adrien Baranes and Pierre-Yves Oudeyer. R-iac: Robust intrinsically motivated exploration and active learning. *IEEE Transactions on Autonomous Mental Development*, 1(3):155–169, 2009.

[8] Andrew G. Barto. Intrinsic Motivation and Reinforcement Learning. In Gianluca Baldassarre and Marco Mirolli, editors, *Intrinsically Motivated Learning in Natural and Artificial Systems*, pages 17–47. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-32374-4 978-3-642-32375-1. doi: 10.1007/978-3-642-32375-1_2. URL http://link.springer.com/10.1007/978-3-642-32375-1_2.

[9] Andrew G Barto and Ozgür Simsek. Intrinsic motivation for reinforcement learning systems. In *Proceedings of the Thirteenth Yale Workshop on Adaptive and Learning Systems*, pages 113–118, 2005.

[10] Andrew G Barto, Satinder Singh, and Nuttapong Chentanez. Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd International Conference on Development and Learning*, pages 112–19. Cambridge, MA, 2004.

[11] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in neural information processing systems*, pages 1471–1479, 2016.

[12] Marc G. Bellemare, Ivo Danihelka, Will Dabney, S. Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and R. Munos. The cramer distance as a solution to biased wasserstein gradients. *ArXiv*, abs/1705.10743, 2017. URL https://arxiv.org/abs/1705.10743.

[13] Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. From optimal transport to generative modeling: the VEGAN cookbook. *arXiv:1705.07642 [stat]*, May 2017. URL http://arxiv.org/abs/1705.07642. arXiv: 1705.07642.

[14] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[15] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[16] Jack Clark and Dario Amodei. Faulty reward functions in the wild, Dec 2016. URL https://openai.com/blog/faulty-reward-functions/.

[17] Robert Dadashi, Léonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal Wasserstein Imitation Learning. *arXiv:2006.04678 [cs, stat]*, June 2020. URL http://arxiv.org/abs/2006.04678. arXiv: 2006.04678.

[18] Y. Ding, Carlos Florensa, Mariano Phielipp, and P. Abbeel. Goal-conditioned imitation learning. In *NeurIPS*, 2019.

[19] Benjamin Eysenbach, Sergey Levine, and Ruslan Salakhutdinov. Replacing rewards with examples: Example-based policy search via recursive classification. *arXiv preprint arXiv:2103.12656*, 2021.

[20] Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. C-learning: Learning to achieve goals via recursive classification. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tc5qisoB-C.

[21] Sébastien Forestier, Rémy Portelas, Yoan Mollard, and Pierre-Yves Oudeyer. Intrinsically motivated goal exploration processes with automatic curriculum learning. *arXiv preprint arXiv:1708.02190*, 2017.

[22] Justin Fu, Katie Luo, and Sergey Levine. Learning Robust Rewards with Adversarial Inverse Reinforcement Learning. *arXiv:1710.11248 [cs]*, August 2018. URL http://arxiv.org/abs/1710.11248. arXiv: 1710.11248.

[23] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pages 1587–1596. PMLR, 2018.

[24] Yaroslav Ganin, Tejas Kulkarni, Igor Babuschkin, SM Ali Eslami, and Oriol Vinyals. Synthesizing programs for images using reinforced adversarial learning. In *International Conference on Machine Learning*, pages 1666–1675. PMLR, 2018.

[25] Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, pages 1259–1277. PMLR, 2020.

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[27] Jacqueline Gottlieb, Pierre-Yves Oudeyer, Manuel Lopes, and Adrien Baranes. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17(11):585–593, 2013.

[28] Matthieu Guillot and Gautier Stauffer. The stochastic shortest path problem: a polyhedral combinatorics perspective. *European Journal of Operational Research*, 285(1):148–158, 2020.

[29] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved Training of Wasserstein GANs. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5767–5777. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans.pdf.

[30] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1352–1361, 2017.

[31] Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

[32] Matthew Hausknecht and Peter Stone. Deep reinforcement learning in parameterized action space. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.04143.

[33] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.

[34] Ashley Hill, Antonin Raffin, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, Rene Traore, Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, and Yuhuai Wu. Stable baselines. https://github.com/hill-a/stable-baselines, 2018.

[35] Jonathan Ho and Stefano Ermon. Generative Adversarial Imitation Learning. *arXiv:1606.03476 [cs]*, June 2016. URL http://arxiv.org/abs/1606.03476. arXiv: 1606.03476.

[36] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010. ISSN 1533-7928. URL http://jmlr.org/papers/v11/jaksch10a.html.

[37] Filip Jevtić. *Combinatorial Structure of Finite Metric Spaces*. PhD thesis, The University of Texas at Dallas, August 2018.

[38] Yuu Jinnai, Jee Won Park, Marlos C. Machado, and George Konidaris. Exploration in reinforcement learning with deep covering options. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeIyaVtwB.

[39] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

[40] W Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis) design for autonomous driving. *arXiv preprint arXiv:2104.13906*, 2021.

[41] Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

[42] JG Liao and Arthur Berg. Sharpening jensen's inequality. *The American Statistician*, 2018.

[43] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

[44] Kanglin Liu and Guoping Qiu. Lipschitz constrained gans via boundedness and continuity. *Neural Computing and Applications*, pages 1–13, 2020.

[45] Douglas C Montgomery. *Design and analysis of experiments*. John wiley & sons, 2017.

[46] Soroush Nasiriany. Disco rl: Distribution-conditioned reinforcement learning for general-purpose policies. Master's thesis, EECS Department, University of California, Berkeley, Aug 2020. URL http://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-151.html.

[47] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287, 1999.

[48] Tianwei Ni, Harshit Sikchi, Yufei Wang, Tejus Gupta, Lisa Lee, and Benjamin Eysenbach. F-irl: Inverse reinforcement learning via state marginal matching. In *Conference on Robot Learning*, 2020.

[49] Scott Niekum. Evolved intrinsic reward functions for reinforcement learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 1955–1956, 2010.

[50] Chris Nota and Philip S Thomas. Is the policy gradient a gradient? In *International Conference on Autonomous Agents and Multi-Agent Systems*, pages 939–947, 2020.

[51] Pierre-Yves Oudeyer and Frederic Kaplan. How can we define intrinsic motivation? In *the 8th International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*. Lund University Cognitive Studies, Lund: LUCS, Brighton, 2008.

[52] Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2009.

[53] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. URL http://arxiv.org/abs/1803.00567.

[54] Antonin Raffin. Rl baselines zoo. https://github.com/araffin/rl-baselines-zoo, 2018.

[55] Vieri Giuliano Santucci, Gianluca Baldassarre, and Marco Mirolli. Which is the best intrinsic motivation signal for learning multiple skills? *Frontiers in neurorobotics*, 7:22, 2013.

[56] T. Schaul, Daniel Horgan, K. Gregor, and D. Silver. Universal value function approximators. In *ICML*, 2015.

[57] Massimiliano Schembri, Marco Mirolli, and Gianluca Baldassarre. Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot. In *2007 IEEE 6th International Conference on Development and Learning*, pages 282–287. IEEE, 2007.

[58] Özgür Şimşek and Andrew G Barto. An intrinsic reward mechanism for efficient exploration. In *Proceedings of the 23rd international conference on Machine learning*, pages 833–840, 2006.

[59] Satinder Singh, Andrew G Barto, and Nuttapong Chentanez. Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pages 1281–1288, 2005.

[60] Satinder Singh, Richard L. Lewis, Andrew G. Barto, and Jonathan Sorg. Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):70–82, June 2010. ISSN 1943-0612. doi: 10.1109/TAMD.2010.2051031. Conference Name: IEEE Transactions on Autonomous Mental Development.

[61] M. Smyth. Quasi uniformities: Reconciling domains with metric spaces. In *MFPS*, 1987.

[62] Jonathan Sorg, Richard L Lewis, and Satinder P Singh. Reward design via online gradient ascent. In *Advances in Neural Information Processing Systems*, pages 2190–2198, 2010.

[63] Jonathan Sorg, Satinder P Singh, and Richard L Lewis. Internal rewards mitigate agent boundedness. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 1007–1014, 2010.

[64] Jan Stanczuk, Christian Etmann, L. Kreusser, and C. Schönlieb. Wasserstein gans work because they fail (to approximate the wasserstein distance). *ArXiv*, abs/2103.01678, 2021.

[65] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[66] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[67] Faraz Torabi, Garrett Warnell, and Peter Stone. Generative Adversarial Imitation from Observation. *arXiv:1807.06158 [cs, stat]*, June 2019. URL http://arxiv.org/abs/1807.06158. arXiv: 1807.06158.

[68] Srinivas Venkattaramanujam, E. Crawford, T. Doan, and Doina Precup. Self-supervised learning of distance functions for goal-conditioned reinforcement learning. *ArXiv*, abs/1907.02998, 2019.

[69] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[70] Huang Xiao, Michael Herman, Joerg Wagner, Sebastian Ziesche, Jalal Etesami, and Thai Hong Linh. Wasserstein Adversarial Imitation Learning. *arXiv:1906.08113 [cs, stat]*, June 2019. URL http://arxiv.org/abs/1906.08113. arXiv: 1906.08113.

[71] Yunzhi Zhang, Pieter Abbeel, and Lerrel Pinto. Automatic Curriculum Learning through Value Disagreement. *arXiv:2006.09641 [cs, stat]*, June 2020. URL http://arxiv.org/abs/2006.09641. arXiv: 2006.09641.

[72] Zeyu Zheng, Junhyuk Oh, and Satinder Singh. On Learning Intrinsic Rewards for Policy Gradient Methods. *arXiv:1804.06459 [cs, stat]*, June 2018. URL http://arxiv.org/abs/1804.06459. arXiv: 1804.06459.

[73] Zeyu Zheng, Junhyuk Oh, Matteo Hessel, Zhongwen Xu, Manuel Kroiss, Hado van Hasselt, David Silver, and Satinder Singh. What Can Learned Intrinsic Rewards Capture? *arXiv:1912.05500 [cs]*, July 2020. URL http://arxiv.org/abs/1912.05500. arXiv: 1912.05500.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description.

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We do so in the subsection 5.2 where we justify design choices
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] In our discussion section 7
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Refer to Section 5.1
   (b) Did you include complete proofs of all theoretical results? [Yes] Section 5.1 and Appendix D
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In section 6, Appendix F, and Appendix H
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Refer to Appendix F and Appendix H
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Section 6
   (b) Did you mention the license of the assets? [Yes] Appendix H
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]