CROSS-DOMAIN IMITATION LEARNING VIA OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-domain imitation learning studies how to leverage expert demonstrations of one agent to train an imitation agent with a different embodiment or morphology. Comparing trajectories and stationary distributions between the expert and imitation agents is challenging because they live on different systems that may not even have the same dimensionality. We propose *Gromov-Wasserstein Imitation Learning (GWIL)*, a method for cross-domain imitation that uses the Gromov-Wasserstein distance to align and compare states between the different spaces of the agents. Our theory formally characterizes the scenarios where GWIL preserves optimality, revealing its possibilities and limitations. We demonstrate the effectiveness of GWIL in non-trivial continuous control domains ranging from simple rigid transformation of the expert domain to arbitrary transformation of the state-action space. [1]

1 INTRODUCTION

Reinforcement learning (RL) methods have attained impressive results across a number of domains, e.g., Berner et al. (2019); Kober et al. (2013); Levine et al. (2016); Vinyals et al. (2019). However, the effectiveness of current RL method is heavily correlated to the quality of the training reward. Yet for many real-world tasks, designing dense and informative rewards require significant engineering effort. To alleviate this effort, imitation learning (IL) proposes to learn directly from expert demonstrations. Most current IL approaches can be applied solely to the simplest setting where the expert and the agent share the same embodiment and transition dynamics that live in the same state and action spaces. In particular, these approaches require expert demonstrations from the agent domain. Therefore, we might reconsider the utility of IL as it seems to only move the problem, from designing informative rewards to providing expert demonstrations, rather than solving it. However, if we relax the constraining setting of current IL methods, then natural imitation scenarios that genuinely alleviate engineering effort appear. Indeed, not requiring the same dynamics would enable agents to imitate humans and robots with different morphologies, hence widely enlarging the applicability of IL and alleviating the need for in-domain expert demonstrations.

This relaxed setting where the expert demonstrations comes from another domain has emerged as a budding area with more realistic assumptions (Gupta et al., 2017; Liu et al., 2019; Sermanet et al., 2018; Kim et al., 2020; Raychaudhuri et al., 2021) that we will refer to as *Cross-Domain Imitation Learning*. A common strategy of these works is to learn a mapping between the expert and agent domains. To do so, they require access to proxy tasks where both the expert and the agent act optimally in there respective domains. Under some structural assumptions, the learned map enables to transform a trajectory in the expert domain into the agent domain while preserving the optimality. Although these methods indeed relax the typical setting of IL, requiring proxy tasks heavily restrict the applicability of Cross-Domain IL. For example, it rules out imitating an expert never seen before as well as transferring to a new robot.

In this paper, we relax the assumptions of Cross-Domain IL and propose a benchmark and method that do not need access to proxy tasks. To do so, we depart from the point of view taken by previous work and formalize Cross-Domain IL as an optimal transport problem. We propose a method, that we call *Gromov Wasserstein Imitation Learning (GWIL)*, that uses the Gromov-Wasserstein distance

[1] Project site with videos: https://sites.google.com/view/gwil/home

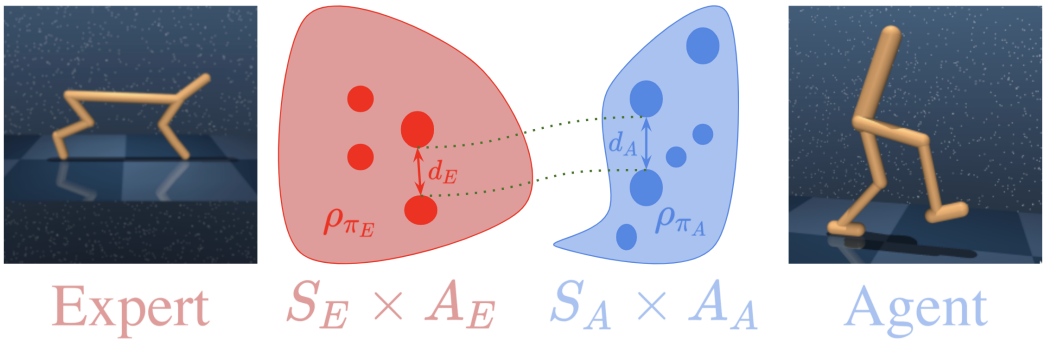


Figure 1: The Gromov-Wasserstein distance enables us to compare the stationary state-action distributions of two agents with different dynamics and state-action spaces. We use it as a pseudo-reward for cross-domain imitation learning.

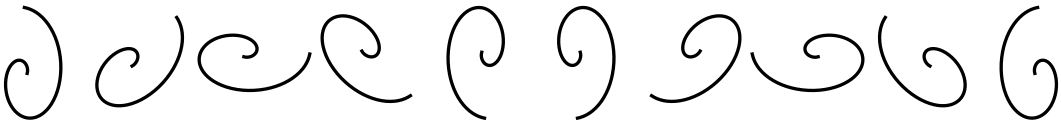


Figure 2: Isometric policies (definition 2) have the same pairwise distances within the state-action space of the stationary distributions. In Euclidean spaces, isometric transformations preserve these pairwise distances and include rotations, translations, and reflections.

to solve the benchmark. We formally characterize the scenario where GWIL preserves optimality (theorem 1), revealing the possibilities and limitations. The construction of our proxy rewards to optimize optimal transport quantities using RL generalizes previous work that assumes uniform occupancy measures (Dadashi et al., 2020; Papagiannis & Li, 2020) and is of independent interest. Our experiments show that GWIL learns optimal behaviors with a single demonstration from another domain without any proxy tasks in non-trivial continuous control settings.

2 Related Work

Imitation learning. An early approach to IL is Behavioral Cloning (Pomerleau, 1988; 1991) which amounts to training a classifier or regressor via supervised learning to replicate the expert's demonstration. Another key approach is Inverse Reinforcement Learning (Ng & Russell, 2000; Abbeel & Ng, 2004; Abbeel et al., 2010), which aims at learning a reward function under which the observed demonstration is optimal and can then be used to train a agent via RL. To bypass the need to learn the expert's reward function, Ho & Ermon (2016) show that IRL is a dual of an occupancy measure matching problem and propose an adversarial objective whose optimization approximately recover the expert's state-action occupancy measure, and a practical algorithm that uses a generative adversarial network (Goodfellow et al., 2014). While a number of recent work aims at improving this algorithm relative to the training instability caused by the minimax optimization, Primal Wasserstein Imitation Learning (PWIL) (Dadashi et al., 2020) and Sinkhorn Imitation Learning (SIL) (Papagiannis & Li, 2020) view IL as an optimal transport problem between occupancy measures to completely eliminate the minimax objective and outperforms adversarial methods in terms of sample efficiency. Heess et al. (2017); Peng et al. (2018); Zhu et al. (2018); Aytar et al. (2018) scale imitation learning to complex human-like locomotion and game behavior in non-trivial settings. Our work is an extension of Dadashi et al. (2020); Papagiannis & Li (2020) from the Wasserstein to the Gromov-Wasserstein setting. This takes us beyond limitation that the expert and imitator are in the same domain and into the cross-domain setting between agents that live in different spaces.

Transfer learning across domains and morphologies. Work transferring knowledge between different domains in RL typically learns a mapping between the state and action spaces. Ammar et al. (2015) use unsupervised manifold alignment to find a linear map between states that have similar local geometry but assume access to hand-crafted features. More recent work in transfer learning

across viewpoint and embodiment mismatch learn a state mapping without handcrafted features but assume access to paired and time-aligned demonstration from both domains (Gupta et al., 2017; Liu et al., 2018; Sermanet et al., 2018). Furthermore, Kim et al. (2020); Raychaudhuri et al. (2021) propose methods to learn a state mapping from unpaired and unaligned tasks. All these methods require proxy tasks, *i.e*. a set of pairs of expert demonstrations from both domains, which limit the applicability of these methods to real-world settings. Stadie et al. (2017) have proposed to combine adversarial learning and domain confusion to learn a policy in the agent's domain without proxy tasks but their method only works in the case of small viewpoint mismatch. Zakka et al. (2021) take a goal-driven perspective that seeks to imitate task progress rather than match fine-grained structural details to transfer between physical robots. In contrast, our method does not rely on learning an explicit cross-domain latent space between the agents, nor does it rely on proxy tasks. The Gromov-Wasserstein distance enables us to directly compare the different spaces without a shared space. The existing benchmark tasks we are aware of assume access to a set of demonstrations from *both* agents whereas the experiments in our paper *only* assume access to expert demonstrations.

3 Preliminaries

Metric Markov Decision Process. An infinite-horizon discounted *Markov decision Process (MDP)* is a tuple $(S, A, R, P, p_0, \gamma)$ where S and A are state and action spaces, $P : S \times A \to \Delta(S)$ is the transition function, $R : S \times A \to \mathbb{R}$ is the reward function, $p_0 \in \Delta(S)$ is the initial state distribution and γ is the discount factor. We equip MDPs with a distance $d : S \times A \to \mathbb{R}^+$ and call the tuple $(S, A, R, P, p_0, \gamma, d)$ a *metric MDP*.

Gromov-Wasserstein distance. Let $(\mathcal{X}, d_\mathcal{X}, \mu_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y}, \mu_\mathcal{Y})$ be two metric measure spaces, where $d_\mathcal{X}$ and $d_\mathcal{Y}$ are distances, and $\mu_\mathcal{X}$ and $\mu_\mathcal{Y}$ are measures on their respective spaces[2]. The *Gromov-Wasserstein distance* (Mémoli, 2011) extends the Wasserstein distance from optimal transportation (Villani, 2009) to these spaces and is defined as

$$\mathcal{GW}((\mathcal{X}, d_\mathcal{X}, \mu_\mathcal{X}), (\mathcal{Y}, d_\mathcal{Y}, \mu_\mathcal{Y}))^2 = \min_{u \in \mathcal{U}(\mu_\mathcal{X}, \mu_\mathcal{Y})} \sum_{\mathcal{X}^2 \times \mathcal{Y}^2} |d_\mathcal{X}(x, x') - d_\mathcal{Y}(y, y')|^2 u_{x,y} u_{x',y'}, \quad (1)$$

where $\mathcal{U}(\mu_\mathcal{X}, \mu_\mathcal{Y})$ is the set of couplings between the atoms of the measures defined by

$$\mathcal{U}(\mu_\mathcal{X}, \mu_\mathcal{Y}) = \left\{ u \in \mathbb{R}^{\mathcal{X} \times \mathcal{Y}} \;\middle|\; \forall x \in \mathcal{X}, \sum_{y \in \mathcal{Y}} u_{x,y} = \mu_\mathcal{X}(x), \forall y \in \mathcal{Y}, \sum_{x \in \mathcal{X}} u_{x,y} = \mu_\mathcal{Y}(y) \right\}.$$

$\mathcal{GW}$ compares the structure of two metric measure spaces by comparing the pairwise distances within each space to find the best isometry between the spaces. Figure 1 illustrates this distance in the case of the metric measure spaces $(S_E \times A_E, d_E, \rho_{\pi_E})$ and $(S_A \times A_A, d_A, \rho_{\pi_A})$.

4 Cross-Domain Imitation Learning via Optimal Transport

4.1 Comparing policies from arbitrarily different MDPs

For a stationary policy π acting on a metric MDP (S, A, R, P, γ, d), the *occupancy measure* is:

$$\rho_\pi : S \times A \to \mathbb{R} \qquad \rho(s, a) = \pi(a|s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi).$$

We compare policies from arbitrarily different MDPs in terms of their occupancy measures.

Definition 1 (Gromov-Wasserstein distance between policies)**.** *Given an expert policy π_E and an agent policy π_A acting, respectively, on*

$$M_E = (S_E, A_E, R_E, P_E, T_E, d_E) \quad \text{and} \quad M_A = (S_A, A_A, R_A, P_A, T_A, d_A).$$

We define the Gromov-Wasserstein distance between π_E and π_A as the Gromov-Wasserstein distance between the metric measure spaces $(S_E \times A_E, d_E, \rho_{\pi_E})$ and $(S_A \times A_A, d_A, \rho_{\pi_A})$:

$$\mathcal{GW}(\pi, \pi') = \mathcal{GW}((S_E \times A_E, d_E, \rho_{\pi_E}), (S_A \times A_A, d_A, \rho_{\pi_A})). \quad (2)$$

[2]We use discrete spaces for readability but show empirical results in continuous spaces.

We now define an isometry between policies by comparing the distances between the state-action spaces and show that $\mathcal{GW}$ defines a distance up to an isometry between the policies. Figure 2 illustrates examples of simple isometric policies.

Definition 2 (Isometric policies)**.** *Two policies π_E and π_A are isometric if there exists a bijection $\phi : \text{supp}[\rho_{\pi_E}] \to \text{supp}[\rho_{\pi_A}]$ that satisfies for all $(s_E, a_E), (s_E{}', a_E{}') \in \text{supp}[\rho_{\pi_E}]^2$:*

$$d_E\left((s_E, a_E), (s_E{}', a_E{}')\right) = d_A\left(\phi(s_E, a_E), \phi(s_E{}', a_E{}')\right)$$

In other words, ϕ is an isometry between $(\text{supp}[\rho_{\pi_E}], d_E)$ and $(\text{supp}[\rho_{\pi_A}], d_A)$.

Proposition 1. *$\mathcal{GW}$ defines a metric on the collection of all isometry classes of policies.*

Proof. By definition 1, $\mathcal{GW}(\pi_E, \pi_A) = 0$ if and only if $\mathcal{GW}((S_E, d_E, \rho_{\pi_E}), (S_A, d_A, \rho_{\pi_A})) = 0$. By Mémoli (2011, Theorem 5.1), this is true if and only if there is an isometry that maps $\text{supp}[\rho_{\phi_E}]$ to $\text{supp}[\rho_{\phi_A}]$. By definition 2, this is true if and only if π_A and π_E are isometric. The symmetry and triangle inequality follow from Mémoli (2011, Theorem 5.1). □

The next theorem[3] gives a sufficient condition to recover, by minimizing $\mathcal{GW}$, an optimal policy[4] in the agent's domain up to an isometry.

Theorem 1. *Consider two MDPs*

$$M_E = (S_E, A_E, R_E, P_E, p_E, \gamma) \quad \text{and} \quad M_A = (S_A, A_A, R_A, P_A, p_A, \gamma).$$

Suppose that there exists four distances $d_E^S, d_E^A, d_A^S, d_A^A$ defined on S_E, A_E, S_A and A_E respectively, and two isometries $\phi : (S_E, d_E^S) \to (S_A, d_A^S)$ and $\psi : (A_E, d_E^S) \to (A_S, d_A^S)$ such that for all $(s_E, a_E, s'_E) \in S_E \times A_E \times S_E$ the three following conditions hold:

$$R(s_E, a_E) = R_A(\phi(s_E), \psi(a_E)) \tag{3}$$

$$P_{E_{s_E, a_E}}(s'_E) = P_{A_{\phi(s_E)\psi(a_E)}}(\phi(s'_E)) \tag{4}$$

$$p_E(s_E) = p_A(\phi(s_E)). \tag{5}$$

Consider an optimal policy π_E^* in M_E. Suppose that π_{GW} minimizes $\mathcal{GW}(\pi_E^*, \pi_{GW})$ with

$$d_E : (s_E, a_E) \mapsto d_E^S(s_E) + d_E^A(a_E) \quad \text{and} \quad d_A : (s_A, a_A) \mapsto d_A^S(s_A) + d_A^A(a_A).$$

Then π_{GW} is isometric to an optimal policy in M_A.

Proof. Consider the occupancy measure $\rho_A^* : S_A \times A_A \to \mathbb{R}$ given by

$$(s_A, a_A) \mapsto \rho_{\pi_E^*}(\phi^{-1}(s_A), \psi^{-1}(a_A)).$$

We first show that ρ_A^* is feasible in M_A, *i.e.* there exists a policy π_A^* acting in M_A with occupancy measure ρ_A^* (a). Then we show that π_A^* is optimal in M_A (b) and is isometric to π_E^* (c). Finally we show that π_{GW} is isometric to π_A^*, which concludes the proof (d).

(a) Consider $s_A \in S_A$. By definition of ρ_A^*,

$$\sum_{a_A \in A_A} \rho_A^*(s_A) = \sum_{a_A \in A_A} \rho_{\pi_E^*}(\phi^{-1}(s_A), \psi^{-1}(a_A)) = \sum_{a_E \in A_E} \rho_{\pi_E^*}(\phi^{-1}(s_A), a_E).$$

Since $\rho_{\pi_E^*}$ is feasible in M, it follows from Puterman (2014, Theorem 6.9.1) that

$$\sum_{a_E \in A_E} \rho_{\pi_E^*}(\phi^{-1}(s_A), a_E) = p_E(\phi^{-1}(s_A)) + \gamma \sum_{s_E \in S_E, a_E \in A_E} P_{E_{s_E, a_E}}(\phi^{-1}(s_A)) + \rho_{\pi_E^*}(s_E, a_E).$$

By conditions 4 and 5 and by definition of ρ_A^*,

$$\begin{aligned}
&p_E(\phi^{-1}(s_A)) + \gamma \sum_{s_E \in S_E, a_E \in A_E} P_{E_{s_E, a_E}}(\phi^{-1}(s_A)) + \rho_{\pi_E^*}(s_E, a_E) \\
&= p_A(s_A) + \gamma \sum_{s_E \in S_E, a_E \in A_E} P_{A_{\phi(s_E), \psi(a_E)}}(s_A) + \rho_A^*(\phi(s_E), \psi(a_E)) \\
&= p_A(s_A) + \gamma \sum_{s'_A \in S_A, a_A \in A_A} P_{A_{s'_A, a_A}}(s_A) + \rho_A^*(s'_A, a_A).
\end{aligned}$$

[3] Our proof is in finite state-action spaces for readability and can be directly extended to infinite spaces.

[4] A policy is optimal in the MDP (S, A, R, P, γ, d) if it maximizes the expected return $\mathbb{E} \sum_{t=0}^{\infty} R(s_t, a_t)$.

It follows that

$$\sum_{a_A \in A_A} \rho_A^*(s_A) = p_A(s_A) + \gamma \sum_{s'_A \in S_A, a_A \in A_A} P_{A_{s'_A, a_A}}(s_A) + \rho_A^*(s'_A, a_A).$$

Therefore, by Puterman (2014, Theorem 6.9.1), ρ_A^* is feasible in M_A, *i.e.* there exists a policy π_A^* acting in M_A with occupancy measure ρ_A^*.

(b) By condition 5 and definition of ρ_A^*, the expected return of π_A^* in M_A is then

$$\begin{aligned}
&\sum_{s_A \in S_A, a_A \in A_A} \rho_A^*(s_A, a_A) R_A(s_A, a_A) \\
&\quad = \sum_{s_A \in S_A, a_A \in A_A} \rho_E^*(\phi^{-1}(s_A), \psi^{-1}(a_A)) R_E(\phi^{-1}(s_A), \psi^{-1}(a_A)) \\
&\quad = \sum_{s_E \in S_E, a_E \in A_E} \rho_E^*(s_E, a_E) R_E(s_E, a_E)
\end{aligned}$$

Consider any policy π_A in M'. By condition 5, the expected return of π_A is

$$\sum_{s_A \in S_A, a_A \in A_A} \rho_{\pi_A}(s_A, a_A) R_A(s_A, a_A) = \sum_{s_E \in S_E, a_E \in A_E} \rho_{\pi_A}(\phi(s_E), \psi(a_E)) R_E(s_E, a_E).$$

Using the same arguments that we used to show that ρ_A^* is feasible in M', we can show that

$$(s_E, a_E) \mapsto \rho_{\pi_A}(\phi(s_E), \psi(a_E))$$

is feasible in M. It follows by optimality of π_E^* in M that

$$\begin{aligned}
\sum_{s_E \in S_E, a_E \in A_E} \rho_{\pi_A}(\phi(s_E), \psi(a_E)) R_E(s_E, a_E) &\leq \sum_{s_E \in S_E, a_E \in A_E} \rho_{\pi_E^*}(\phi(s_E), \psi(a_E)) R_E(s_E, a_E) \\
&= \sum_{s_A \in S_A, a_A \in A_A} \rho_A^*(s_A, a_A) R_A(s_A, a_A).
\end{aligned}$$

It follows that π_A^* is optimal in M'.

(c) Notice that

$$\xi : (s_E, a_E) \mapsto (\phi(s_E), \psi(a_E))$$

is an isometry between $(S_E \times A_E, d_E)$ and $(S_A \times A_A, d_A)$, where d_E and d_A and given, resp., by

$$(s_E, a_E) \mapsto d_E^S(s_E) + d_E^A(a_E) \quad \text{and} \quad (s_A, a_A) \mapsto d_A^S(s_A) + d_A^A(a_A).$$

Therefore by definition of ρ_A^*, π_A^* is isometric to π_E^*.

(d) Recall from the statement of the theorem that π_{GW} is a minimizer of $\mathcal{GW}(\pi_E^*, \pi_{GW})$. Since π_A^* is isometric to π_E^*, it follows from prop. 1 that $\mathcal{GW}(\pi_E^*, \pi_A^*) = 0$. Therefore $\mathcal{GW}(\pi_E^*, \pi_{GW})$ must be 0. By prop. 1, it follows that there exists an isometry

$$\chi : (\text{supp}[\rho_E^*], d_E) \to (\text{supp}[\rho_{\pi_{GW}}], d_A).$$

Notice that $\chi \circ \xi^{-1}|_{\text{supp}[\rho_A^*]}$ is an isometry from $(\text{supp}[\rho_A^*], d_A)$ to $(\text{supp}[\rho_{\pi_{GW}}], d_A)$. It follows that π_{GW} is isometric to π_A^*, an optimal policy in M_A, which concludes the proof. □

Remark 1. *Theorem 1 shows the possibilities and limitations of our method. It shows that our method can recover optimal policies even though arbitrary isometries are applied to the state and action spaces of the expert's domain. Importantly, we don't need to know the isometries, hence our method is applicable to a wide range of settings. We will show empirically that our method produces strong results in other settings where the environment are not isometric and don't even have the same dimension. However, a limitation of our method is that it recovers optimal policy only up to isometries. We will see that in practice, running our method on different seeds enables to find an optimal policy in the agent's domain.*

Algorithm 1 Gromov-Wasserstein imitation learning from a single expert demonstration.

Inputs: expert demonstration τ, metrics on the expert (d_E) and agent (d_A) space
Initialize the imitation agent's policy π_θ and value estimates V_θ
while Unconverged **do**
  Collect an episode τ'
  Compute $\mathcal{GW}(\tau, \tau')$
  Set pseudo-rewards r with eq. (7)
  Update π_θ and V_θ to optimize the pseudo-rewards
end while

4.2 Gromov-Wasserstein Imitation Learning

Minimizing $\mathcal{GW}$ between an expert and agent requires derivatives through the transition dynamics, which we typically don't have access to. We introduce a reward proxy suitable for training an agent's policy that minimizes $\mathcal{GW}$ via RL. Figure 1 illustrates the method. For readability, we combine expert state and action variables (s_E, a_E) into single variables z_E, and similarly for agent state-action pairs. Also, we define $Z_E = S_E \times A_E$ and $Z_A = S_A \times A_A$.

Definition 3. *Given an expert policy π_E and an agent policy π_A, the Gromov-Wasserstein reward of the agent is defined as $r_{\mathcal{GW}} : \text{supp}[\rho_{\pi_A}] \to \mathbb{R}$ given by*

$$r_{\mathcal{GW}}(z_A) = -\frac{1}{\rho_\pi(z_A)} \sum_{\substack{z_E \in Z_E \\ z'_E \in Z_E \\ z'_A \in Z_A}} |d_E(z_E, z'_E)) - d_A(z_A, z'_A)|^2 u^\star_{z_E, z_A} u^\star_{z'_E, z'_A}$$

where $u^\star$ is the coupling minimizing objective 1.

Proposition 2. *The agent's policy π_A trained with $r_{\mathcal{GW}}$ minimizes $\mathcal{GW}(\pi_E, \pi_A)$.*

Proof. Suppose that π_A maximizes $\mathbb{E}(\sum_{t=0}^{\infty} \gamma^t r_{\mathcal{GW}}(s_t^A, a_t^A))$ and denote by ρ_{π_A} its occupancy measure. By Puterman (2014, Theorem 6.9.4), π_A maximizes the following objective:

$$\begin{aligned}
\mathop{\mathbb{E}}_{z_A \sim \rho_{\pi_A}} r_{\mathcal{GW}}(z_A) &= -\sum_{z_A \in \text{supp}[\rho_{\pi_A}]} \frac{\rho_{\pi_A}(z_A)}{\rho_{\pi_A}(z_A)} \sum_{\substack{z_E \in Z_E \\ z'_E \in Z_E \\ z'_A \in Z_A}} |d_E(z_E, z'_E) - d_A(z_A, z'_A)|^2 u^\star_{z_A, z_E} u^\star_{z'_A, z'_E} \\
&= -\sum_{\substack{z_E \in Z_E \\ z'_E \in Z_E \\ z_A \in Z_A \\ z'_A \in Z_A}} |d_E(z_E, z'_E) - d_A(z_A, z'_A)|^2 u^\star_{z_A, z_E} u^\star_{z'_A, z'_E} \\
&= -\mathcal{GW}^2(\pi_E, \pi_A)
\end{aligned}$$

□

In practice we approximate the occupancy measures of π by $\hat{\rho}_\pi(s, a) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(s = s_t \wedge a = a_t)$ where $\tau = (s_1, a_1, .., s_T, a_T)$ is a finite trajectory collected with π. Assuming that all state-action pairs in the trajectory are different[5], $\hat{\rho}$ is a uniform distribution. Given an expert trajectory τ_E and an agent trajectory τ_A [6], the (squared) Gromov-Wasserstein distance between the empirical occupancy measures is

$$\mathcal{GW}^2(\tau_E, \tau_A) = \min_{\theta \in \Theta^{T_E \times T_A}} \sum_{\substack{1 \le i, i' \le T_E \\ 1 \le j, j' \le T_A}} |d_E((s_i^E, a_i^E), (s_{i'}^E, a_{i'}^E)) - d_A((s_j^A, s_j^A), (s_{j'}^A, a_{j'}^A))|^2 \theta_{i,j} \theta_{i',j'} \tag{6}$$

[5] We can add the time step to the state to distinguish between two identical state-action pairs in the trajectory.

[6] Note that the Gromov-Wasserstein distance defined in equ. (6) does not depend on the temporal ordering of the trajectories.

where Θ is the set of is the set of couplings between the atoms of the uniform measures defined by

$$\Theta^{T\times T'} = \left\{ \theta \in \mathbb{R}^{T\times T'} \;\middle|\; \forall i \in [T], \sum_{j\in[T']} \theta_{i,j} = 1/T, \forall j \in [T'], \sum_{i\in[T]} \theta_{i,j} = 1/T' \right\}.$$

In this case the reward is given for every state-action pairs in the trajectory by:

$$r(s_j^A, s_j^A) = -T_A \sum_{\substack{1\leq i,i'\leq T_E \\ 1\leq j'\leq T_A}} |d_E((s_i^E, a_i^E), (s_{i'}^E, a_{i'}^E)) - d_A((s_j^A, s_j^A), (s_{j'}^A, a_{j'}^A))|^2 \theta^\star_{i,j}\theta^\star_{i',j'} \tag{7}$$

where $\theta^\star$ is the coupling minimizing objective 6.

In practice we drop the factor T_A because it is the same for every state-action pairs in the trajectory.

Remark 2. *The construction of our reward proxy is defined for any occupancy measure and extends to previous work optimizing optimal transport quantities via RL that assumes uniform occupancy measure in the form of a trajectory to bypass the need for derivatives through the transition dynamics (Dadashi et al., 2020; Papagiannis & Li, 2020).*

Optimizing the pseudo-rewards. The pseudo-rewards we obtain from $\mathcal{GW}$ for the imitation agent enable us to turn the imitation learning problem into a reinforcement learning problem (Sutton & Barto, 2018) to find the optimal policy for the Markov decision process induced by the pseudo-rewards. We consider agents with continuous state-action spaces and thus do policy optimization with the soft actor-critic algorithm (Haarnoja et al., 2018). Algorithm 1 sums up GWIL in the case where a single expert trajectory is given to approximate the expert occupancy measure.

5 Experiments

We propose a benchmark set for cross-domain IL methods consisting of 3 tasks and aiming at answering the following questions:

1. *Does GWIL recover optimal behaviors when the agent domain is a rigid transformation of the expert domain?* Yes, we demonstrate this with the maze in sect. 5.1.
2. *Can GWIL recover optimal behaviors when the agent has different state and action spaces than the expert?* Yes, we show in sect. 5.2 for *slightly* different state-action spaces between the cartpole and pendulum, and in sect. 5.3 for *significantly* different spaces between a walker and cheetah.

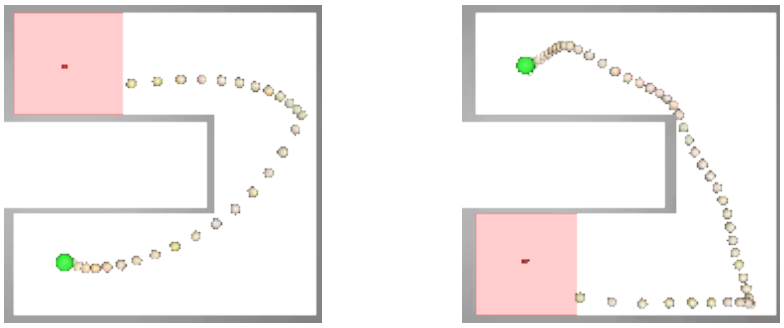
To answer these three questions, we use simulated continuous control tasks implemented in Mujoco (Todorov et al., 2012) and the DeepMind control suite (Tassa et al., 2018). In all settings we use the Euclidean metric within the expert and agent spaces for d_E and d_A.

5.1 Agent domain is a rigid transformation of the expert domain

We evaluate the capacity of IL methods to transfer to rigid transformation of the expert domain by using the PointMass Maze environment from Hejna et al. (2020). The agent's domain is obtained by applying a reflection to the expert's maze. This task satisfies the condition of theorem 1 with ϕ being the reflection through the central horizontal plan and ψ being the reflection through the x-axis in the action space. Therefore by theorem 1, the agent's optimal policy should be isometric to the policy trained using GWIL. By looking at the geometry of the maze, it is clear that every policy in the isometry class of an optimal policy is optimal. Therefore we expect GWIL to recover an optimal policy in the agent's domain. Figure 3 shows that GWIL indeed recovers an optimal policy.

5.2 Agent and the expert have slightly different state and action spaces

We evaluate here the capacity of IL methods to transfer to transformation that does not have to be rigid but description map should still be apparent by looking at the domains. A good example of such transformation is the one between the pendulum and cartpole. The pendulum is our expert's domain



(a) Expert (b) Agent

Figure 3: Given a single expert trajectory in the expert's domain (a), GWIL recovers an optimal policy in the agent's domain (b) without any external reward, as predicted by theorem 1. The green dot represents the initial state position and the episode ends when the agent reaches the goal represented by the red square.

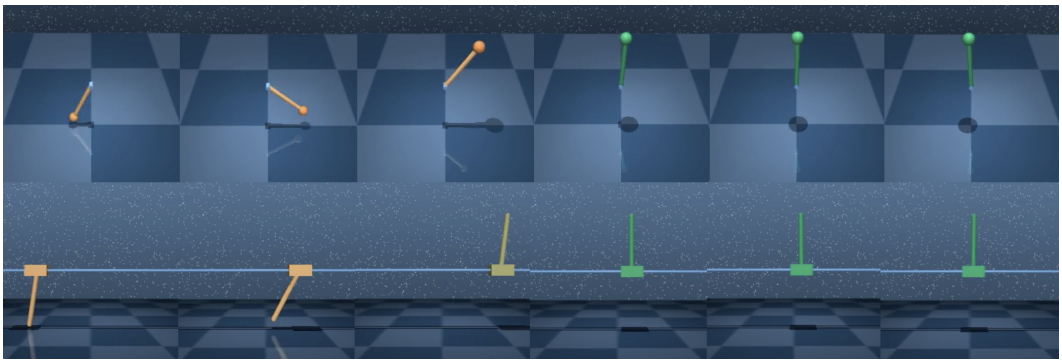


Figure 4: Given a single expert trajectory in the pendulum's domain (above), GWIL recovers the optimal behavior in the agent's domain (cartpole, below) without any external reward.

while cartpole constitutes our agent's domain. The expert is trained on the swingup task. Even though the transformation is not rigid, GWIL is able to recover the optimal behavior in the agent's domain as shown in fig. 4. Notice that pendulum and cartpole do not have the same state-action space dimension: The pendulum has 3 dimensions while the cartpole has 5 dimensions. Therefore GWIL can indeed be applied to transfer between problems with different dimension.

5.3 Agent and the expert have significantly different state and action spaces

We evaluate here the capacity of IL methods to transfer to non-trivial transformation between domains. A good example of such transformation is two arbitrarily different morphologies from the DeepMind Control Suite such as the cheetah and walker. The cheetah constitutes our expert's domain while the walker constitutes our agent's domain. The expert is trained on the run task.

Although the mapping between these two domains is not trivial, minimizing the Gromov-Wasserstein solely enables the walker to interestingly learn to move backward and forward by imitating a cheetah. Since the isometry class of the optimal policy – moving forward– of the cheetah and walker contains a suboptimal element –moving backward–, we expect GWIL to recover one of these two trajectories. Indeed, depending on the seed used, GWIL produces a cheetah-imitating walker moving forward or a cheetah-imitating walker moving backward, as shown in fig. 5.

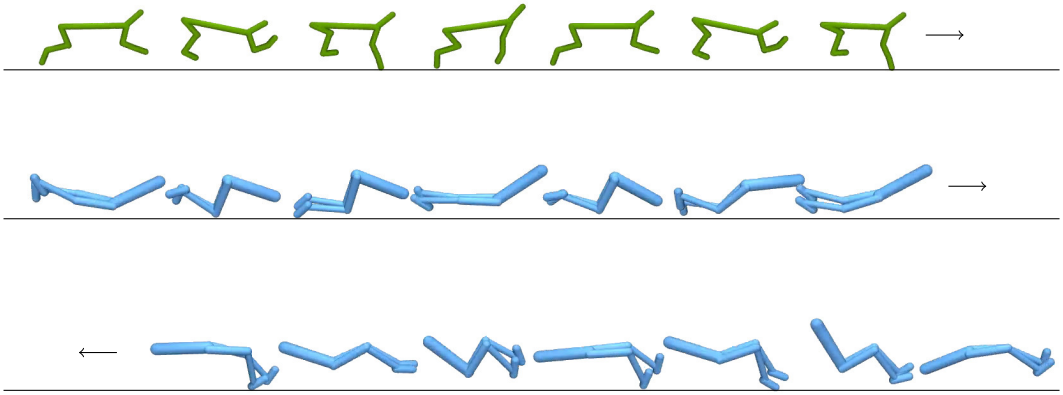


Figure 5: Given a single expert trajectory in the cheetah's domain (above), GWIL recovers the two elements of the optimal policy's isometry class in the agent's domain (walker), moving forward which is optimal (middle) and moving backward which is suboptimal (below). Interestingly, the resulting walker behaves like a cheetah.

6 Conclusion

Our work demonstrates that optimal transport distances are a useful foundational tool for cross-domain imitation across incomparable spaces. Future directions include exploring:

1. **Scaling to more complex environments and agents** towards the goal of transferring the structure of many high-dimensional demonstrations of complex tasks into an agent.
2. The use of $\mathcal{GW}$ to help agents **explore in extremely sparse-reward environments** when we have expert demonstrations available from other agents.
3. How $\mathcal{GW}$ compares to **other optimal transport distances** that work apply between two metric MDPs, such as Alvarez-Melis et al. (2019), that have more flexibility over how the spaces are connected and what invariances the coupling has.
4. **Metrics aware of the MDP's temporal structure** such as Zhou & Torre (2009); Cohen et al. (2021) that build on dynamic time warping (Müller, 2007). The Gromov-Wasserstein ignores the temporal information and ordering present within the trajectories.

REFERENCES

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

Pieter Abbeel, Adam Coates, and Andrew Y Ng. Autonomous helicopter aerobatics through apprenticeship learning. *The International Journal of Robotics Research*, 29(13):1608–1639, 2010.

David Alvarez-Melis, Stefanie Jegelka, and Tommi S. Jaakkola. Towards optimal transport with global invariances. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1870–1879. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/alvarez-melis19a.html.

Haitham Bou Ammar, Eric Eaton, Paul Ruvolo, and Matthew E Taylor. Unsupervised cross-domain transfer in policy gradient reinforcement learning via manifold alignment. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Yusuf Aytar, Tobias Pfaff, David Budden, Tom Le Paine, Ziyu Wang, and Nando de Freitas. Playing hard exploration games by watching youtube. *arXiv preprint arXiv:1805.11592*, 2018.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Samuel Cohen, Giulia Luise, Alexander Terenin, Brandon Amos, and Marc Deisenroth. Aligning time series on incomparable spaces. In *International Conference on Artificial Intelligence and Statistics*, pp. 1036–1044. PMLR, 2021.

Robert Dadashi, Léonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal wasserstein imitation learning. *arXiv preprint arXiv:2006.04678*, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Abhishek Gupta, Coline Devin, YuXuan Liu, Pieter Abbeel, and Sergey Levine. Learning invariant feature spaces to transfer skills with reinforcement learning. *arXiv preprint arXiv:1703.02949*, 2017.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Nicolas Heess, Dhruva TB, Srinivasan Sriram, Jay Lemmon, Josh Merel, Greg Wayne, Yuval Tassa, Tom Erez, Ziyu Wang, SM Eslami, et al. Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*, 2017.

Donald Hejna, Lerrel Pinto, and Pieter Abbeel. Hierarchically decoupled imitation for morphological transfer. In *International Conference on Machine Learning*, pp. 4159–4171. PMLR, 2020.

Jonathan Ho and S. Ermon. Generative adversarial imitation learning. In *NIPS*, 2016.

Kuno Kim, Yihong Gu, Jiaming Song, Shengjia Zhao, and Stefano Ermon. Domain adaptive imitation learning. In *International Conference on Machine Learning*, pp. 5286–5295. PMLR, 2020.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Fangchen Liu, Zhan Ling, Tongzhou Mu, and Hao Su. State alignment-based imitation learning. *arXiv preprint arXiv:1911.10947*, 2019.

YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1118–1125. IEEE, 2018.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pp. 69–84, 2007.

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Georgios Papagiannis and Yunpeng Li. Imitation learning with sinkhorn distances. *arXiv preprint arXiv:2008.09167*, 2020.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

D. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *NIPS*, 1988.

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

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Dripta S Raychaudhuri, Sujoy Paul, Jeroen van Baar, and Amit K Roy-Chowdhury. Cross-domain imitation from observations. *arXiv preprint arXiv:2105.10037*, 2021.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 1134–1141. IEEE, 2018.

Bradly C Stadie, Pieter Abbeel, and Ilya Sutskever. Third-person imitation learning. *arXiv preprint arXiv:1703.01703*, 2017.

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

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

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*, pp. 5026–5033. IEEE, 2012.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Kevin Zakka, Andy Zeng, Pete Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. Xirl: Cross-embodiment inverse reinforcement learning. *arXiv preprint arXiv:2106.03911*, 2021.

Feng Zhou and Fernando Torre. Canonical time warping for alignment of human behavior. *Advances in neural information processing systems*, 22:2286–2294, 2009.

Yuke Zhu, Ziyu Wang, Josh Merel, Andrei Rusu, Tom Erez, Serkan Cabi, Saran Tunyasuvunakool, János Kramár, Raia Hadsell, Nando de Freitas, et al. Reinforcement and imitation learning for diverse visuomotor skills. *arXiv preprint arXiv:1802.09564*, 2018.

A OPTIMIZATION OF THE PROXY REWARD

In this section we show that the proxy reward introduced in equation 7 constitutes a learning signal that is easy to optimize using standards RL algorithms. Figure 6 shows proxy reward curves across 5 different seeds for the 3 environments. We observe that in each environment the SAC learner converges quickly and consistently to the asymptotic episodic return. Thus there is reason to think that the proxy reward introduced in equation 7 will be similarly easy to optimize in other cross-domain imitation settings.

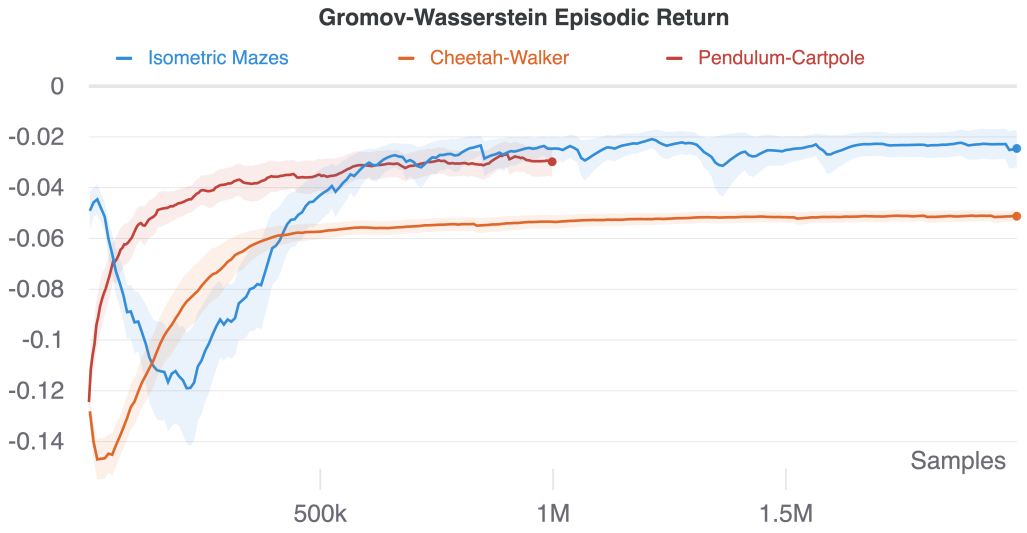


Figure 6: The proxy reward introduced in equation 7 gives a learning signal that is easily optimized using a standard RL algorithm.

B TRANSFER TO SPARSE-REWARD ENVIRONMENTS

In this section we show that GWIL can be used to facilitate learning in sparse-reward environments when the learner has only access to one expert demonstration from another domain. We compare GWIL to a baseline learner having access to a single demonstration from the same domain and minimizing the Wasserstein distance, as done in Dadashi et al. (2020). In these experiments, both agents are given a sparse reward signal in addition to their respective optimal transport proxy reward. We perform experiments in two sparse-reward environment. In the first environment, the agent controls a point mass in a maze and obtain a non-zero reward only if it reaches the end of the maze. In the second environment, which is a sparse version of cartpole, the agent controls a cartpole and obtains a non-zero reward only if he can maintain the cartpole up for 10 consecutive time steps. Note that a SAC agent fails to learn any meaningful behavior in both environments. Figure 7 shows that GWIL is competitive with the baseline learner in the sparse maze environment even though GWIL has only access to a demonstration from another domain, while the baseline learner has access to a demonstration from the same domain. Thus there is reason to think that GWIL efficiently and reliably extracts useful information from the expert domain and hence should work well in other cross-domain imitation settings.

Figure 7: In sparse-reward environments, GWIL obtains similar performance than a baseline learner minimizing the Wasserstein distance to an expert in the same domain.

C Scalability of GWIL

In this section we show that our implementation of GWIL offers good performance in terms of wall-clock time. Note that the bottleneck of our method is in the computation of the optimal coupling which only depends on the number of time steps in the trajectories, and not on the dimension of the expert and the agent. Hence our method naturally scales with the dimension of the problems. Furthermore, while we have not used any entropy regularizer in our experiments, entropy regularized methods have been introduced to enable Gromov-Wasserstein to scale to demanding machine learning tasks and can be easily incorporated into our code to further improve the scalability. Figure compares the time taken by GWIL in the maze with the time taken by the baseline learner introduced in the previous section. It shows that imitating with Gromov-Wasserstein requires the same order of time than imitating with Wasserstein. Figure compares the wall-clock time taken by a walker imitating a cheetah using GWIL to reach a walking speed (i.e., a horizontal velocity of 1) and the wall-clock time taken by a SAC walker trained to run. It shows that a GWIL walker imitating a cheetah reaches a walking speed faster than a SAC agent trained to run. Thus there is reason to think that GWIL offers a scalable approach to cross-domain imitation learning and can be applicable to a wide range of domains.

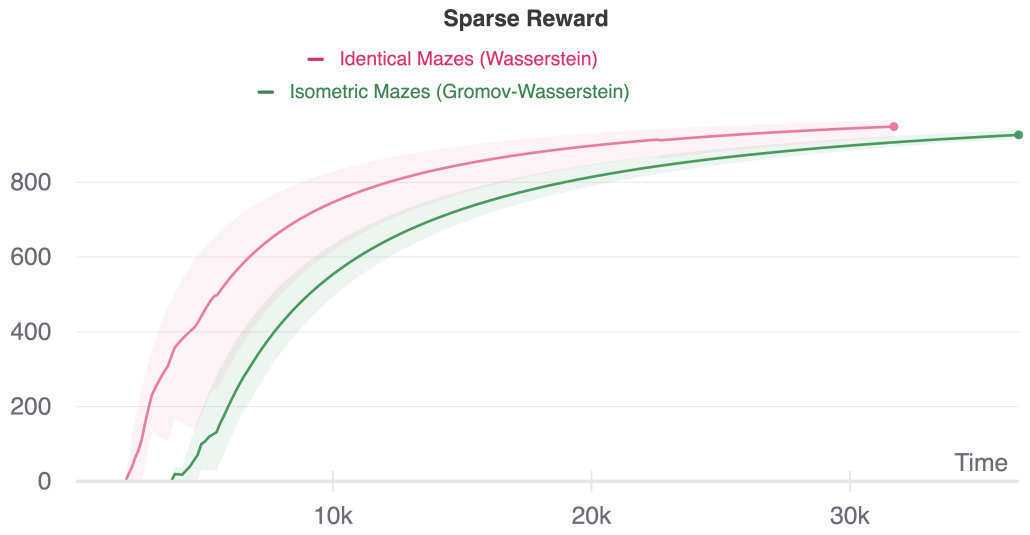


Figure 8: In the sparse maze environment, GWIL requieres the same order of wall-clock time than a baseline learner minimizing the Wasserstein distance to an expert in the same domain.

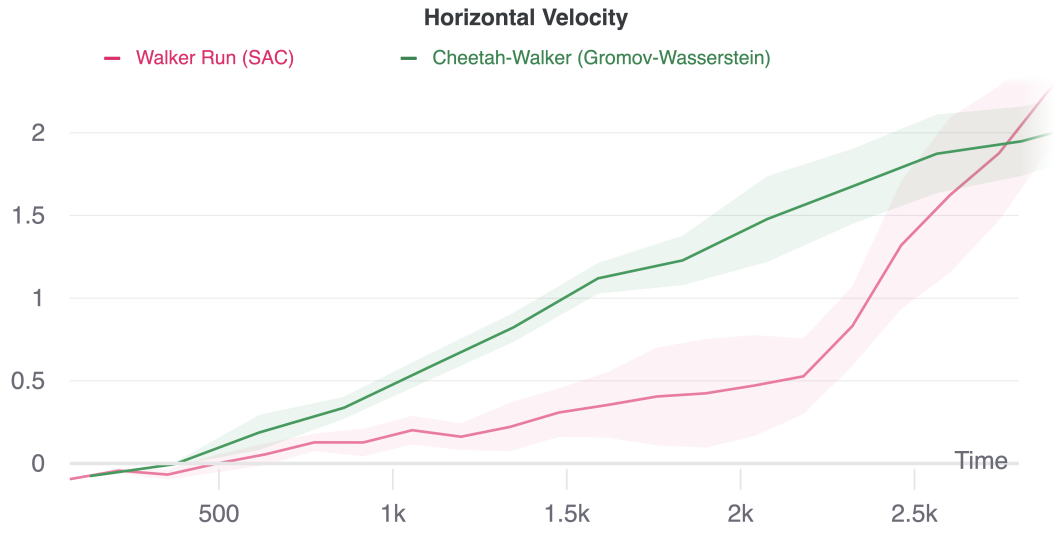


Figure 9: A GWIL walker imitating a cheetah reaches a walking speed faster than a SAC walker trained to run in terms of wall-clock time.