TASK ADAPTATION FROM SKILLS: INFORMATION GEOMETRY, DISENTANGLEMENT, AND NEW OBJECTIVES FOR UNSUPERVISED REINFORCEMENT LEARNING

Yucheng Yang[1], **Tianyi Zhou**[2], **Qiang He**[3], **Lei Han**[4], **Mykola Pechenizkiy**[1], **Meng Fang**[5]
[1]Eindhoven University of Technology, [2]University of Maryland, College Park,
[3]Ruhr University Bochum, [4]Tencent Robotics X, [5]University of Liverpool
{y.yang, m.pechenizkiy}@tue.nl, tianyi@umd.edu, mfang@liverpool.ac.uk

ABSTRACT

Unsupervised reinforcement learning (URL) aims to learn general skills for unseen downstream tasks. Mutual Information Skill Learning (MISL) addresses URL by maximizing the mutual information between states and skills but lacks sufficient theoretical analysis, e.g., how well its learned skills can initialize a downstream task's policy. Our new theoretical analysis in this paper shows that the diversity and separability of learned skills are fundamentally critical to downstream task adaptation but MISL does not necessarily guarantee these properties. To complement MISL, we propose a novel disentanglement metric LSEPIN. Moreover, we build an information-geometric connection between LSEPIN and downstream task adaptation cost. For better geometric properties, we investigate a new strategy that replaces the KL divergence in information geometry with Wasserstein distance. We extend the geometric analysis to it, which leads to a novel skill-learning objective WSEP. It is theoretically justified to be helpful to downstream task adaptation and it is capable of discovering more initial policies for downstream tasks than MISL. We finally propose another Wasserstein distance-based algorithm PWSEP that can theoretically discover all optimal initial policies.

1 INTRODUCTION

Reinforcement learning (RL) has drawn growing attention by its success in autonomous control (Kiumarsi et al., 2017), Go (Silver et al., 2016) and video games (Mnih et al., 2013; Vinyals et al., 2019). However, a primary limitation of the current RL is its high sample complexity. Inspired by the successful pretrain-finetune paradigm in other deep learning fields like natural language processing (Radford et al., 2019; Devlin et al., 2019) and computer vision (Henaff, 2020; He et al., 2020), there has been growing work studying the pretraining of RL. RL agent receives no task-related reward during pretraining and learns by its intrinsic motivations (Oudeyer & Kaplan, 2009). Some of these intrinsic motivations can help the agent to learn representations of the observations (Schwarzer et al., 2021) and some learn the dynamics model (Ha & Schmidhuber, 2018; Sekar et al., 2020). In this work, we focus on Unsupervised RL (URL) that learns a set of skills without external reward and the learned skills are expected to be quickly adapted to unseen downstream tasks.

A common approach for skill discovery of URL is Mutual Information Skill Learning (**MISL**) (Eysenbach et al., 2022) that maximizes the mutual information between state and skill latent (Eysenbach et al., 2019; Florensa et al., 2017; Hansen et al., 2020; Liu & Abbeel, 2021b). The intuition is that by maximizing this mutual information the choice of skills can effectively affect where the states are distributed so that these skills could be potentially used for downstream tasks. There are more algorithms using objectives modified on this mutual information. For example, Lee et al. (2019); Liu & Abbeel (2021b) added additional terms for better exploration, and Sharma et al. (2020); Park et al. (2022a) focus on modified input structure to prepare the agent for specific kinds of downstream tasks.

Despite the popularity of MISL, there has been little theoretical analysis of how well the MISL-learned skills can be applied as downstream task initializations. Previous work Eysenbach et al. (2022) has tried to analyze MISL but they consider an impractical downstream task adaptation procedure

that uses the average state distribution of all learned skills as initialization instead of directly using the learned skills. Therefore, it is still unclear how well the MISL-learned skills can be applied as downstream task initializations.

In this work, we theoretically analyze the connection between the properties of learned skills and their downstream task performance. Our results show that the diversity and separability of learned skills are fundamentally critical to downstream task adaptation. Separability, or the distinctiveness of skill distributions, is key for diverse skills. Without it, even a large number of skills may cover only a limited range resulting in limited diversity. The importance of diversity is empirically demonstrated in previous works (Eysenbach et al., 2019; Kim et al., 2021; He et al., 2022; Laskin et al., 2022). Our results also show that MISL alone does not necessarily guarantee these properties. To complement MISL, we propose a novel disentanglement metric that is able to measure the diversity and separability of learned skills. Our theoretical analysis relates the disentanglement metric to downstream task adaptation.

In particular, we introduce a novel disentanglemen metric "**L**east **SEP**arability and **IN**formativeness (**LSEPIN**)", which is directly related to the task adaptation cost from learned skills and complementary to the widely adopted mutual information objective of MISL. LSEPIN captures both the informativeness, diversity, and separability of the learned skills, which are critical to downstream tasks and can be used to design better URL objectives. We relate LSEPIN to **W**orst-case **A**daptation **C**ost (**WAC**), which measures the largest possible distance between a downstream task's optimal feasible state distribution and its closest learned skill's state distribution. Our results show increasing LSEPIN could potentially result in lower WAC.

In addition, we show that optimizing MISL and LSEPIN are essentially maximizing distances measured by KL divergences between state distributions. However, a well-known issue is that KL divergence is not a true metric, i.e., it is not symmetric and does not satisfy the triangle inequality. This motivates us to investigate whether an alternative choice of distance can overcome the limitations of MISL. Wasserstein distance is a symmetric metric satisfying the triangle inequality and has been feasibly applied for deep learning implementations (Arjovsky et al., 2017; Dadashi et al., 2020), so we investigate a new strategy that replaces the KL divergence in MISL with Wasserstein distance and exploits its better geometric properties for theoretical analysis. This leads to new skill learning objectives for URL and our results show that the objective built upon Wasserstein distance, "**W**asserstein **SEP**aratibility (**WSEP**)", is able to discover more potentially optimal skills than MISL. Furthermore, we propose and analyze an unsupervised skill-learning algorithm "**P**rojected **SEP**" (**PWSEP**) that has the favored theoretical property to discover all potentially optimal skills and is able to solve the open question of "vertex discovery" from Eysenbach et al. (2022).

Analysis of LSEPIN is complement to prior work to extend the theoretical analysis of MISL to practical downstream task adaptation, while the analysis of WSEP and PWSEP opens up a new unsupervised skill learning approach. Our results also answer the fundamental question of URL about what properties of the learned skills lead to better downstream task adaptation and what metrics can measure these properties.

Our main contributions can be summarized in the following:

1. We theoretically study a novel but practical task adaptation cost (i.e., WAC) for MISL, which measures how well the MISL-learned skills can be applied as downstream task initializations.
2. We propose a novel disentanglement metric (i.e., LSEPIN) that captures both the informativeness and separability of skills. LSEPIN is theoretically related to WAC and can be used to develop URL objectives.
3. We propose a new URL formulation based on Wasserstein distance and extend the above theoretical analysis to it, resulting in novel URL objectives for skill learning. Besides also promoting separability, they could discover more skills than existing MISL that are potentially optimal for downstream tasks.

Although our contribution is mainly theoretical, in appendices H and I we show the feasibility of practical algorithm design with our proposed metrics and empirical examples to validate our results. A summary of our proposed metrics and algorithm is in appendix A and frequently asked questions are answered in appendix B.

2 PRELIMINARIES

We consider infinite-horizon MDPs $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, p_0, \gamma)$ *without external rewards* with discrete states $\mathcal{S}$ and actions $\mathcal{A}$, dynamics $P(s_{t+1}|, s_t, a_t)$, initial state distribution $p_0(s_0)$, and discount factor $\gamma \in [0, 1]$. A policy $\pi(a|s)$ has its discounted state occupancy measure as $p^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t P_t^\pi(s)$, where $P_t^\pi(s)$ is the probability that policy π visits state $\mathbf{s}$ at time t. There can be downstream tasks that define extrinsic reward as a state-dependent function $r(s)$, where action-dependent reward functions can be handled by modifying the state to include the previous action. The cumulative reward of the corresponding downstream task is $\mathbb{E}_{p^\pi(s)}[r(s)]$.

We formulate the problem of unsupervised skill discovery as learning a skill-conditioned policy $\pi(a_t|s_t, z)$ where $z \in \mathcal{Z}$ represents the latent skill and $\mathcal{Z}$ is a discrete set. $H(\cdot)$ and $I(\cdot;\cdot)$ denote entropy and mutual information, respectively. $W(\cdot,\cdot)$ denotes Wasserstein distance. We use upper-case letters for random variables and lower-case letters for samples, eg. $s \sim p(S)$.

2.1 MUTUAL INFORMATION SKILL LEARNING

Unsupervised skill learning algorithms aim to learn a policy $\pi(A|S, Z)$ conditioned on a latent skill z. Their optimization objective is usually the mutual information $I(S; Z)$ and they differ on the prior or approximation of this objective (Gregor et al., 2017; Eysenbach et al., 2019; Achiam et al., 2018; Hansen et al., 2020).

In practical algorithms, the policy is generally denoted as $\pi_\theta(A|S, z_{\text{input}})$ with parameters θ and conditioned on an skill latent $z_{\text{input}} \sim p(Z_{\text{input}})$. Let $p^{\pi_\theta}(S|z_{\text{input}})$ denote the state distribution of policy. The practical objective of MISL could be:

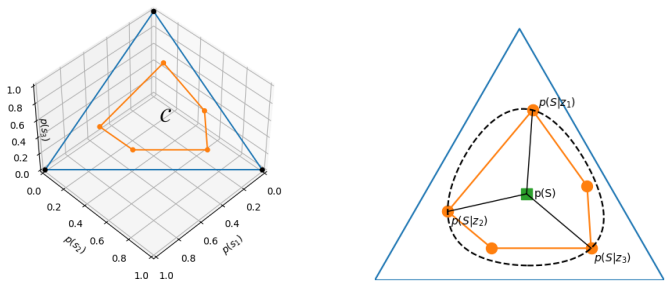
$$\max_{\theta, p(Z_{\text{input}})} I(S; Z_{\text{input}}) = \mathbb{E}_{p(Z_{\text{input}})}[D_{\text{KL}}(p^{\pi_\theta}(S|z_{\text{input}}) \parallel p^{\pi_\theta}(S))], \tag{1}$$

Policy parameters θ and the latent variable Z_{input} can be composed into a single representation, $z = (\theta, z_{\text{input}})$, then $\pi_\theta(A|S, z_{\text{input}}) = \pi(A|S, z)$. We call representation z "skill" in the following paper. Then, MISL is learned by finding an optimal $p(Z)$ that solves

$$\max_{p(Z)} I(S; Z) = \mathbb{E}_{p(Z)}[D_{\text{KL}}(p(S|z) \parallel p(S))]., \tag{2}$$

where $p(S) = \mathbb{E}_{p(Z)}[p(S|z)]$, is the average state distribution of discovered skills.

2.2 INFORMATION GEOMETRY OF MISL



(a) (b)

Figure 1: Visualized examples: (a) $\mathcal{C}$ is the feasible state distribution set, and the blue simplex is the probability simplex for state distribution with $|\mathcal{S}| = 3$. (b) MISL discovers 3 skills at the vertices $\{z_1, z_2, z_3\}$ on the "circle" with maximum "radius" centered in their average state distribution $p(S)$.

Prior work Eysenbach et al. (2022) shows that the set $\mathcal{C}$ of state distributions feasible under the dynamics of the MDP constitutes a convex polytope lying on a probability simplex of state distributions, where every point in the polytope $\mathcal{C}$ is represented by a skill latent z and its state distribution is $p(S|z)$. For any downstream task defined by a reward function $r : \mathcal{S} \to \mathcal{R}$, because of the linearity of $\mathbb{E}_{p(s)}[r(s)]$ and convexity of $\mathcal{C}$, the state distribution that maximizes the cumulative reward $\mathbb{E}_{p(s)}[r(s)]$

lies at one of the vertices of $\mathcal{C}$ (Boyd & Vandenberghe, 2014). Equation (2) shows that MISL learns a skill distribution $z \sim p(Z)$ to put weight on skills that have maximum KL divergence to the average state distribution. It can be considered as finding skills that lie on the unique (uniqueness proved in appendix E) "circle" with maximum "radius" inside the polytope $\mathcal{C}$, thus the discovered skills lie at the vertices of polytope $\mathcal{C}$, as shown in Lemma 6.5 of Eysenbach et al. (2022) by Theorem 13.11 of Cover & Thomas (2006). So the skills discovered by MISL are optimal for some downstream tasks. An intuitive example of the skills discovered by MISL is shown in fig. 1b.

3 Theoretical Results

Although MISL discovers some vertices that are potentially optimal for certain downstream tasks, when the downstream task favors target state distributions at the undiscovered vertices, which often happens in practice that the learned skills are not optimal for downstream tasks, there exists a "distance" from discovered vertices to the target vertex, and the "distance" from the initial skill for adaptation to the target state distribution can be considered as the adaptation cost. The prior work only analyzes the adaptation cost from the average state distribution of skills $p(S) = \mathbb{E}_z[p(S|z)]$ to the target state distribution. Because most practical MISL algorithms initialize the adaptation procedure from one of the learned skills (Lee et al., 2019; Eysenbach et al., 2019; Liu & Abbeel, 2021b; Laskin et al., 2021) instead of the average $p(S)$, the prior analysis provides little insight on why these practical algorithms work. The fundamental question for unsupervised skill learning remains unanswered: How the learned skills can be used for downstream task adaptation and what properties of the learned skills are desired for better downstream task adaption?

We have answered this question with theoretical analysis in this section, empirical validation of the theories is in appendix I. Our informal results are as follows:

1. In order to have a low adaptation cost when initializing from one of the learned skills, the learned skills need to be diverse and separate from each other. Separability means the discriminability between states inferred by different skills.
2. MISL alone does not necessarily guarantee diversity and separability. We propose a disentanglement metric LSEPIN to complement MISL for diverse and separable skills.
3. MISL discovers limited vertices, we propose WSEP metric based on Wasserstein distance that can promote diversity and separability as well as discover more vertices than MISL. One Wasserstein distance-based algorithm PWSEP can even discover all vertices.

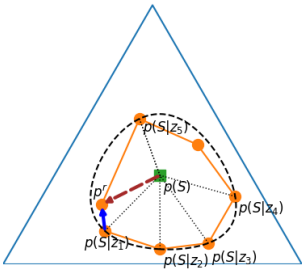


Figure 2: Example of concyclic vertices: z_1, z_2, z_3, z_4, z_5 can all be vertices optimal for MISL, $p(S)$ can be identified by $|\mathcal{S}| = 3$ of them, so eq. (2) can be solved by any 3 of these vertices. The blue arrow is to adapt from learned skills to target distribution for downstream task (our setting) and the brown arrow is to adapt from the average state distribution (setting in Eysenbach et al. (2022))

The first point is intuitive that the diverse and separable skills are likely to cover more potentially useful skills, as shown by empirical results in Eysenbach et al. (2019); Park et al. (2022b); Laskin et al. (2022). The second point claims MISL alone does not guarantee diversity and separability, and this can be seen from the example in fig. 2. In this case, there are two sets of $|\mathcal{S}| = 3$ skills $\mathcal{Z}_a : \{z_1, z_4, z_5\}$ and $\mathcal{Z}_b : \{z_2, z_3, z_5\}$ both on the maximum "circle" solving MISL. Because z_2 and z_3 have close state distributions, skills of $\mathcal{Z}_b$ are less diverse and less separable. There can be more than $|S|$ vertices on the maximum "circle" in the case of fig. 2 because, unlike prior work Eysenbach et al. (2022), we do not take into account the "non-concyclic" assumption that limits the number of vertices on the same "circle" to be $|S|$. Our proposed disentanglement metric LSEPIN would favor $\mathcal{Z}_a$ over $\mathcal{Z}_b$, and theoretical analysis of LSEPIN is conducted in section 3.2 to show its relation to downstream task adaptation cost. The downstream task procedure we consider is initialized from one of the learned tasks, for the case in fig. 2, when the target state distribution is p^r, we consider adapting from the skill in $\mathcal{Z}_a$ that is closest to p^r (blue arrow), which is z_1, while the prior work adapts from $p(S)$ (brown arrow).

The advantages of Wasserstein distance are that it is a true metric satisfying symmetry and triangle inequality. We can use it to measure distances that can not be measured by KL divergences. Optimizing these distances also promotes diversity and separability as well as results in better vertex discovery, even capable of discovering all vertices and solving the open question of "vertex discovery" from Eysenbach et al. (2022). Details about Wasserstein distance skill learning are shown in section 3.3. A summary of all proposed metrics and algorithm is in appendix A.

3.1 How to measure diversity and separatability of learned skills

Many previous MISL algorithms (Eysenbach et al., 2019; Gregor et al., 2017; Sharma et al., 2020) emphasized the importance of diversity and tried to promote diversity by using uniform $p(Z_{\text{input}})$ for eq. (1). However, uniform $p(Z_{\text{input}})$ for objective eq. (1) does not ensure diverse z for $p(Z)$ in eq. (2) since $z = (\theta, z_{\text{input}})$ also depends on the learned parameter θ. We show an example in **??** when maximizing $I(S; Z)$ with uniform $p(Z_{\text{input}})$ results in inseparable skills. Empirical discussions in Park et al. (2022b); Laskin et al. (2022) also mentioned that the learned skills of these MISL methods often lack enough diversity and separability. Furthermore, as mentioned previously by the example in fig. 2, even when $I(S; Z)$ in eq. (2) is maximized, the learned skills could still lack diversity and separability of the skills. To complement MISL, we propose a novel metric to explicitly measure the diversity and separability of learned skills.

We consider $I(S; \mathbf{1}_z)$ ($\mathbf{1}_z$ is the binary indicator function of $Z = z$) to measure the informativeness and separability of an individual skill z. In the context of unsupervised skill learning, informativeness should refer to the information shared between a skill and its inferred states. As mentioned, separability means the states inferred by different skills should be discriminable. We analyze the minimum of $I(S; \mathbf{1}_z)$ over learned skills. We name it Least SEParability and INformativeness (LSEPIN)

$$\text{LSEPIN} = \min_z I(S; \mathbf{1}_z). \tag{3}$$

$I(S; \mathbf{1}_z)$ is related to how much states inferred by skill z and states not inferred by z are discriminable from each other, so it covers not only informativeness but also separability of skills. In the context of representation learning, the metrics capturing informativeness and separability are called the disentanglement metrics (Do & Tran, 2019b; Kim et al., 2021), so we also call LSEPIN as a disentanglement metric for unsupervised skill learning. More details about the difference between disentanglement for representation learning and disentanglement for our skill learning setting are in appendix F.

3.2 How disentanglement affects downstream task adaptation

We provide a theoretical justification for the proposed disentanglement metric, showing that it can be a complement of $I(S; Z)$ to evaluate how well the URL agent is prepared for downstream tasks by the following theorems.

Definition 3.1 (Worst-case Adaptation Cost)**.** Worst-case Adaptation Cost (WAC) is defined as

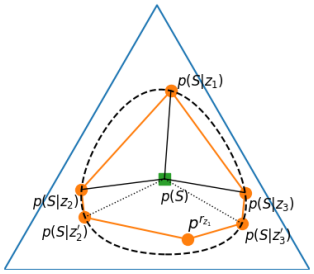
$$\text{WAC} = \max_r \min_{z \in \mathcal{Z}^*} D_{\text{KL}}(p(S|z) \parallel p^r), \tag{4}$$

where p^r is the optimal feasible state marginal distribution for the downstream task defined by r, and $\mathcal{Z}^*$ is the set of learned skills.

The following theoretical results show how the LSEPIN metric is related to the WAC in definition 3.1.

Theorem 3.1. *When learned skill sets $\mathcal{Z}_i, i = 1, 2, \ldots$ with $N \leq |\mathcal{S}|$ skills (N skills have $p(z) > 0$) sharing the same skill z are all MISL solutions, The skill set with the higher $I(S; \mathbf{1}_z)$ will have higher $p(z)$ and lower adaptation cost for all r_z in the set $\mathcal{R}_z$, where $\mathcal{R}_z$ is the set of downstream tasks always satisfying $\forall i, \forall r \in \mathcal{R}_z$, $z = \arg\max_{z' \in \mathcal{Z}_i} D_{\text{KL}}(p(S|z') \parallel p^{r_z})$. And the maximum of this adaptation cost has the following formulation:*

$$IC_z = \max_{r \in \mathcal{R}_z} \frac{C_z(r) - p(z) D_z(r)}{1 - p(z)}, \tag{5}$$



the MISL objective $I(S; Z)$ is maximized by solutions

Figure 3: Example of two MISL solutions: $\mathcal{Z}_1^* : \{z_1, z_2, z_3\}$ and $\mathcal{Z}_2^* : \{z_1, z_2', z_3'\}$, $\mathcal{Z}_2^*$ has higher $I(S; \mathbf{1}_{z_1})$

where

$$C_z(r) = I(S;Z) + D_{\mathrm{KL}}(p(S) \parallel p^r), \tag{6}$$
$$D_z(r) = D_{\mathrm{KL}}(p(S|z) \parallel p^r). \tag{7}$$

Theorem 3.1 provides a correlation between our proposed metric $I(S;\mathbf{1}_z)$ the adaptation cost IC_z from a skill in $\mathcal{Z} \setminus \{z\}$ that is closest to the downstream task optimal distribution and its detailed proof is in appendix C.1. To better understand the claim of this theorem, we can look at the intuitive example shown in fig. 3. In this case $|\mathcal{S}| = 3$. When MISL is maximized by 3 skills, the skill combinations as MISL solutions could be $\mathcal{Z}_1^* : \{z_1, z_2, z_3\}$ and $\mathcal{Z}_2^* : \{z_1, z_2', z_3'\}$. $\mathcal{Z}_2^*$ has higher $I(S;\mathbf{1}_{z_1})$ than $\mathcal{Z}_1^*$. By theorem 3.1, solution $\mathcal{Z}_2^*$ should have lower cost to adapt to the optimal distribution $p^{r_{z_1}}$ of the downstream task r_{z_1}.

Corollary 3.1.1. *When the MISL objective $I(S,Z)$ is maximized by $N \leq |\mathcal{S}|$ skills, WAC is bounded of a solution $\mathcal{Z}^*$ by*

$$WAC \leq \max_{z \in \mathcal{Z}^*} IC_z = \max_{z \in \mathcal{Z}^*} \max_{r \in \mathcal{R}_z} \frac{C_z(r) - p(z)D_z(r)}{1 - p(z)}. \tag{8}$$

WAC is the worst-case adaptation cost defined in definition 3.1, C_z and D_z are as defined in eqs. (6) and (7). $\mathcal{R}_z$ here needs to satisfy $\forall r \in \mathcal{R}_z$, $z = \arg\max_{z' \in \mathcal{Z}^*} D_{\mathrm{KL}}(p(S|z') \parallel p^r)$.

Corollary 3.1.1 provides an upper bound for WAC. The proof is deferred to appendix C.2. The results in Corollary 3.1.1 and theorem 3.1 considered situations when MISL is solved and $I(S;Z)$ is maximized, we also discussed how $I(S;\mathbf{1}_{z_1})$ and LSEPIN affects learned skills and adaptation cost when $I(S;Z)$ is not maximized in appendix C.4.

By theorem 3.1 we know that higher $I(S;\mathbf{1}_z)$ implies lower IC_z, but how much IC_z associated with an individual skill z contribute to the overall WAC can not be known in prior and it depends on specific C_z and D_z. Moreover, specific C_z and D_z depend on the "shape" of the undiscovered parts of $\mathcal{C}$ and can not be known before the discovery of all vertices. Therefore, in practice, like existing work (Durugkar et al., 2021; He et al., 2022) treating the desired properties of each skill equally in practical algorithms, we could treat every IC_z equally. We have the following theorem showing under which assumptions we can treat every IC_z equally for WAC.

Theorem 3.2. *When 1. the optimal state distribution for the downstream task is far from $p(S)$ and 2. The state space is large, i.e. $|\mathcal{S}|$ is large. IC_z of all learned skills can be considered equally contribute to WAC.*

Both assumptions for this theorem are practical and can commonly happen in complex and high-dimensional environments. When every IC_z is treated equally for WAC, optimizing LSEPIN could lead to lower WAC. It is formally analyzed and proven in appendix C.3.

In summary, we have provided theoretical insight on how $I(S;\mathbf{1}_z)$ affects downstream task adaptation and how optimizing LSEPIN could lower WAC under practical assumptions. We do not assume "non-concyclic" vertices and we consider the practical approach of directly adapting from learned skills instead of the average state distribution. LSEPIN is a complement to the mutual information objective $I(S;Z)$. Compared to $I(S;Z)$, it provides a better metric to evaluate the effectiveness of learned MISL skills for potential downstream tasks. Our results have shown the diversity and separability of the learned skills measured by $I(S;\mathbf{1}_z)$ and LSEPIN are desired for better downstream task adaptation.

Remark 3.2.1. *One limitation with MISL even with LSEPIN is that even without the limitation of the number of skills to have $p(z) > 0$, it still can not discover vertices v such that*

$$D_{\mathrm{KL}}(p(S|v) \parallel p(S)) < \max_{p(z)} \mathbb{E}_{p(z)} \left[D_{\mathrm{KL}}(p(S|z) \parallel p(S)) \right]$$

Vertex $p^{r_{z_1}}$ in fig. 3 belongs to such vertices.

3.3 Skill Learning with Wasserstein Distance

In this subsection, we analyze a new strategy that replaces the KL divergence in information geometry with Wasserstein distance for better geometric properties to overcome the limitation of MISL shown in remark 3.2.1.

Maximizing $I(S; Z)$ and LSEPIN are essentially maximizing distances measured by KL divergences between points in a polytope. KL divergence is not symmetric and does not satisfy the triangle inequality, so KL divergences between points of the polytope could be incomparable when two KL divergences don't share a same point. We study the strategy that replaces the KL divergence in MISL with Wasserstein distance since Wasserstein distance is a true metric. Then we conduct further theoretical analysis to exploit its better geometrical properties such as symmetry and triangle inequality.

In this section, we will introduce the learning objectives as well as evaluation metrics for **W**asserstein **D**istance **S**kill **L**earning (WDSL), analyze what kind of skills these objectives can learn, where the learned skills lie in the polytope, and how these learned skills contribute to downstream task adaptation. Theoretically, the favored property of WDSL is that it discovers more vertices in $\mathcal{C}$ that are potentially optimal for downstream tasks than MISL, and one WDSL algorithm can discover all vertices.

3.3.1 Objectives for Wasserstein Distance skill learning

First of all, we can trivially replace the KL divergences in the MISL objective eq. (2) with Wasserstein distance and obtain a basic WDSL objective

$$\max_{p(z)} \mathbb{E}_{p(z)} \left[W(p(S|z), p(S)) \right]. \tag{9}$$

We name it **A**verage **W**asserstein skill learning **D**istance (AWD), similar to the MISL objective in eq. (2), this objective also learns skills that lie on a hyper ball with a maximum radius. Because this objective is not our main proposition and also suffers from the limitation of remark 3.2.1, we put the analysis of this objective in appendix G.1.

We mainly analyze this objective for WDSL:

$$\text{WSEP} = \sum_{z_i \in \mathcal{Z}} W(p(S|z_i), \sum_{z_j \in \mathcal{Z} \setminus \{z_i\}} \frac{1}{|\mathcal{Z}| - 1} p(S|z_j))), \tag{10}$$

where $\mathcal{Z}$ is the set of skills with $p(z) > 0$. We call this objective **W**asserstein **SEP**aratibility (WSEP), it can be considered as a disentanglement for WDSL as it measures the Wasserstein distance between learned skills. Recall that separability for MISL is defined as how discriminable the state is, Wasserstein distances between skills can not only represent discriminability but also can express the distance between trajectories when there are no overlappings.

3.3.2 Geometry of learned skills

As mentioned before in section 2.2, the skills that are potentially optimal for downstream tasks lie at the vertices of the polytope $\mathcal{C}$ of feasible state distributions. By the following lemma, we show that optimizing WSEP will push the learned skills to the vertices of the polytope.

Lemma 3.3. *When WSEP is maximized, all learned skills with $p(z) > 0$ must lie at the vertices of the polytope.*

Proof of this lemma is in appendix G.2.

The previous theoretical results of disentanglement metric LSEPIN depend on the maximization of $I(S; Z)$, so as mentioned in remark 3.2.1, it still only discover vertices with maximum "distances" to the average distribution $p(S)$. However, WSEP does not depend on the maximization of other objectives, e.g., eq. (9), so there is no distance restriction on the vertices discovered by WSEP. Therefore, it is possible for WSEP to discover all vertices of the feasible polytope $\mathcal{C}$, thus discovering all optimal skills for potential downstream tasks. For example, in an environment with a polytope shown in fig. 1b, MISL only discovers 3 vertices on the "circle" with maximum "radius" while WSEP is able to discover all 5 vertices.

Remark 3.3.1. *When there is no limitation on the number of skills with positive probability, Maximizing WSEP could discover more vertices than MISL in some cases, and even potentially discover all vertices, as shown in the example appendix G.6.*

3.3.3 How WSEP affects downstream task adaptation

Then, we propose a theorem about how the WSEP metric is related to downstream task adaptation when there is a limitation on the quantity of learned skills.

Definition 3.2. *Mean Adaptation Cost (MAC): mean of the Wasserstein distances between the undiscovered vertices and the learned skills closest to them.*

$$MAC = \frac{1}{|\mathcal{V} \setminus \mathcal{Z}^*|} \sum_{z' \in \mathcal{V} \setminus \mathcal{Z}^*} \min_{z \in \mathcal{Z}^*} W(p(S|z'), p(S|z)) \tag{11}$$

$\mathcal{V}$ is the set of all skills that have their conditional state distribution at vertices of the MDP's feasible state distribution polytope, and $\mathcal{Z}^*$ is all learned skills with $p(z) > 0$.

Theorem 3.4. *When WSEP is maximized by $|\mathcal{Z}^*|$ skills, the MAC can be upper-bounded:*

$$MAC \leq \frac{\sum_{z \in \mathcal{Z}^*} L_{\mathcal{V}}^z - (|\mathcal{Z}^*| - 1) WSEP}{|\mathcal{V} \setminus \mathcal{Z}^*||\mathcal{Z}^*|} \tag{12}$$

$$MAC \leq \frac{\sum_{z \in \mathcal{Z}^*} L_{\mathcal{V}} - (|\mathcal{Z}^*| - 1) WSEP}{|\mathcal{V} \setminus \mathcal{Z}^*||\mathcal{Z}^*|}, \tag{13}$$

where

$$\begin{aligned} L_{\mathcal{V}}^z &= \sum_{v \in \mathcal{V}} W(p(S|v), p(S|z)) \\ L_{\mathcal{V}} &= \max_{v' \in \mathcal{V}} \sum_{v \in \mathcal{V}} W(p(S|v), p(S|v')) \end{aligned} \tag{14}$$

Theorem 3.4 shows the relation between WSEP and the upper bounds of adaptation cost MAC in the practical setting, where the number of skills to be learned is limited. The proof is in appendix G.3.

In a stationary MDP, the polytope is fixed, so the edge lengths $L_{\mathcal{V}}^z$ and $L_{\mathcal{V}}$ are constant. Larger WSEP seems to tighten the bounds, but different WSEP also means a different $\mathcal{Z}^*$ set of learned skills, thus a different $\sum_{z \in \mathcal{Z}^*} L_{\mathcal{V}}^z$. Therefore, increasing WSEP only tightens the bound in eq. (13) but not necessarily the bound in eq. (12).

Remark 3.4.1. *More distance is not always good: WSEP as a disentanglement metric promotes the distances between learned skills and These two bounds of MAC show that maximizing WSEP can indeed help with downstream task adaptation, but this does not mean that learned skills with more WSEP will always result in lower MAC. An illustrative example is shown in appendix G.5, where more distant skills with higher WSEP do not have lower adaptation costs*

Remark 3.4.2. *If we replace the Wasserstein distances in WSEP with KL divergences, we get a symmetric formulation of $KLSEP = \sum_{z_i \in \mathcal{Z}} \sum_{z_j \in \mathcal{Z}, i \neq j} D_{\mathrm{KL}}(p(S|z_i) \parallel p(S|z_j))$. It is symmetric, but it does not promote diversity and separability because KL divergence does not satisfy the triangle inequality. More details are analyzed in appendix G.8*

WSEP does not suffer from the limitation of remark 3.2.1 because it does not try to find skills on a maximum "circle". Although WSEP can potentially discover more vertices than MISL, we find that it may not be able to discover all vertices of the feasible state distribution polytope $\mathcal{C}$ in appendix G.7.

3.3.4 Solving the vertex discovery problem

The following theorem shows a learning procedure based on Wasserstance distance capable of discovering all vertices of feasible state distribution polytope $\mathcal{C}$.

Theorem 3.5. *When $\mathcal{V}$ is the set of all vertices of the feasible state distribution polytope $\mathcal{C}$, all $|\mathcal{V}|$ vertices can be discovered by $|\mathcal{V}|$ iterations of maximizing*

$$PWSEP(i) : \min_{\lambda} W\Big(p(S|z_i), \sum_{z_j \in \mathcal{Z}_i} \lambda^j p(S|z_j)\Big), \tag{15}$$

where $\mathcal{Z}_i$ is the set of skills discovered from iteration 0 to $i-1$ and z_i is the skill being learned at ith iteration. λ is a convex coeffcient of dimension $i-1$ that every element $\lambda^j \geq 0, \forall j \in \{0, 1, .., i-1\}$ and $\sum_{j \in \{0,1,..,i-1\}} \lambda^j = 1$.

In the initial iteration when $\mathcal{Z}_i = \emptyset$, PWSEP$(0)$ can be $W(p(S|z_0), p(S|z_{rand}))$ with z_{rand} to be a randomly initialized skill.

PWSEP(i) can be considered as a projection to the convex hull of $\mathcal{Z}_i$, so we call it Projected WSEP and this learning procedure the PWSEP algorithm. It can discover all $|\mathcal{V}|$ vertices with only $|\mathcal{V}|$ skills. Although lemma 3.3 shows that maximizing WSEP also discovers vertices, the discovered vertices could be duplicated (shown in appendix G.7). Maximizing projected distance PWSEP(i) could ensure the vertex learned at each new iteration is not discovered before. Proof and more analysis of the vertex discovery problem can be found in appendix G.4.

4 Related Work

MISL is widely implemented and has been the backbone of many URL algorithms (Achiam et al., 2018; Florensa et al., 2017; Hansen et al., 2020). Prior work Eysenbach et al. (2022) tried to provide theoretical justification for the empirical prevalence of MISL from an information geometric (Amari & Nagaoka, 2000), but their analysis mainly considered an unpractical downstream task adaptation procedure. Works like Eysenbach et al. (2019); Park et al. (2022b; 2023); He et al. (2022); Laskin et al. (2022) showed the empirical advantages of favored properties such as diversity and separability of learned skills. Our theoretically justified these properties and showed they benefit practical adaptation.

In Kim et al. (2021) the concept of disentanglement was mentioned. They used the SEPIN@k and WSEPIN metrics from representation learning (Do & Tran, 2019b) to promote the informativeness and separability between different dimensions of the skill latent. However, properties of latent representations could be ensured by optimization only in the representation space, so they do not explicitly regulate the state distributions of learned skills like our proposed LSEPIN and WSEP do. Appendix F discussed more details.

Recent practical unsupervised skill learning algorithms (He et al., 2022; Durugkar et al., 2021) maximize a lower bound of WSEP, so our analysis on WSEP provides theoretical insight on why these Wasserstein distance-based unsupervised skill learning algorithms work empirically. Their empirical results showed the feasibility and usefulness of skill discovery with Wasserstein distance.

Successor feature (SF) method SFOLS (Alegre et al., 2022) can also discover all vertices but learns an over-complete set of skills, which our PWSEP algorithm efficiently avoids. In appendix G.4.2, the difference between the SF setting and our skill learning setting is discussed in detail, as well as the comparison of theoretical properties between our proposed PWSEP and SFOLS. Other methods like Hansen et al. (2020); Liu & Abbeel (2021b) combined MISL with SF for URL, and they are shown to accelerate downstream task adaptation. Since they are MISL methods adapting from one of the learned skills, our theoretical results also apply to them.

5 Conclusion

We investigated the geometry of task adaptation from skills learned by unsupervised reinforcement learning. We proposed a disentanglement metric LSEPIN for mutual information skill learning to capture the diversity and separability of learned skills, which are critical to task adaptation. Unlike the prior analysis, we are able to build a theoretical connection between the metric and the cost of downstream task adaptation. We further proposed a novel strategy that replaces KL divergence with Wasserstein distance and extended the geometric analysis to it, which leads to novel objective WSEP and algorithm PWSEP for unsupervised skill learning. Our theoretical result shows why they should work, what could be done, and what limitations they have. Specifically, we found that optimizing the proposed WSEP objective can discover more optimal policies for potential downstream tasks than previous methods maximizing the mutual information objective $I(S; Z)$. Moreover, the proposed PWSEP algorithm based on Wasserstein distance can theoretically discover all optimal policies for potential downstream tasks.

Our theoretical results could inspire new algorithms using LSEPIN or Wasserstein distance for unsupervised skill learning. For Wasserstein distance, the choice of transport cost is important, which may require strong prior knowledge. Our future work will develop practical algorithms that learn deep representations such that common transport costs such as L2 distance in the representation space can accurately reflect the difficulty of traveling from one state to the other.

REFERENCES

Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. Variational option discovery algorithms. *arXiv preprint arXiv:1807.10299*, 2018.

Lucas Nunes Alegre, Ana L. C. Bazzan, and Bruno C. da Silva. Optimistic linear support and successor features as a basis for optimal policy transfer. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 394–413. PMLR, 2022. URL https://proceedings.mlr.press/v162/alegre22a.html.

Shun-ichi Amari and Hiroshi Nagaoka. Methods of information geometry. 2000.

Brandon Amos and J. Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 136–145. PMLR, 2017. URL http://proceedings.mlr.press/v70/amos17a.html.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2014. ISBN 978-0-521-83378-3. doi: 10.1017/CBO9780511804441. URL https://web.stanford.edu/%7Eboyd/cvxbook/.

Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=H1lJJnR5Ym.

Victor Campos, Alexander Trott, Caiming Xiong, Richard Socher, Xavier Giró-i-Nieto, and Jordi Torres. Explore, discover and learn: Unsupervised discovery of state-covering skills. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1317–1327. PMLR, 2020. URL http://proceedings.mlr.press/v119/campos20a.html.

Eric Carlen and Wilfrid Gangbo. Constrained steepest ascent in the 2-wasserstein metric. *Annals of Mathematics*, 157:807–846, 05 2003. doi: 10.4007/annals.2003.157.807.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations, 2019a. URL https://arxiv.org/abs/1908.09961.

Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. *arXiv preprint arXiv:1908.09961*, 2019b.

Ishan Durugkar, Steven Hansen, Stephen Spencer, and Volodymyr Mnih. Wasserstein distance maximizing intrinsic control. *CoRR*, abs/2110.15331, 2021. URL https://arxiv.org/abs/2110.15331.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *ICLR*, 2019.

Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. The information geometry of unsupervised reinforcement learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=3wU2UX0voE.

Carlos Florensa, Yan Duan, and Pieter Abbeel. Stochastic neural networks for hierarchical reinforcement learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=B1oK8aoxe.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. In *International Conference on Learning Representations*, 2017.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evol. Comput.*, 9(2):159–195, 2001. doi: 10.1162/106365601750190398. URL https://doi.org/10.1162/106365601750190398.

Steven Hansen, Will Dabney, André Barreto, David Warde-Farley, Tom Van de Wiele, and Volodymyr Mnih. Fast task inference with variational intrinsic successor features. In *International Conference on Learning Representations*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Shuncheng He, Yuhang Jiang, Hongchang Zhang, Jianzhun Shao, and Xiangyang Ji. Wasserstein unsupervised reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6884–6892, 2022.

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pp. 4182–4192. PMLR, 2020.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55.

Jaekyeom Kim, Seohong Park, and Gunhee Kim. Unsupervised skill discovery with bottleneck option learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5572–5582. PMLR, 2021. URL http://proceedings.mlr.press/v139/kim21j.html.

Bahare Kiumarsi, Kyriakos G Vamvoudakis, Hamidreza Modares, and Frank L Lewis. Optimal and autonomous control using reinforcement learning: A survey. *IEEE transactions on neural networks and learning systems*, 29(6):2042–2062, 2017.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

Yehuda Koren. On spectral graph drawing. In Tandy J. Warnow and Binhai Zhu (eds.), *Computing and Combinatorics, 9th Annual International Conference, COCOON 2003, Big Sky, MT, USA, July 25-28, 2003, Proceedings*, volume 2697 of *Lecture Notes in Computer Science*, pp. 496–508. Springer, 2003. doi: 10.1007/3-540-45071-8_50. URL https://doi.org/10.1007/3-540-45071-8_50.

Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. URLB: unsupervised reinforcement learning benchmark. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/091d584fced301b442654dd8c23b3fc9-Abstract-round2.html.

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. CIC: contrastive intrinsic control for unsupervised skill discovery. *CoRR*, abs/2202.00161, 2022. URL https://arxiv.org/abs/2202.00161.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric P. Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *CoRR*, abs/1906.05274, 2019.

Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 18459–18473, 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/99bf3d153d4bf67d640051a1af322505-Abstract.html.

Hao Liu and Pieter Abbeel. APS: active pretraining with successor features. In *International Conference on Machine Learning*, 2021b.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

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

Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. Lipschitz-constrained unsupervised skill discovery. *ArXiv*, abs/2202.00914, 2022a.

Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. Lipschitz-constrained unsupervised skill discovery, 2022b. URL https://arxiv.org/abs/2202.00914.

Seohong Park, Kimin Lee, Youngwoon Lee, and Pieter Abbeel. Controllability-aware unsupervised skill discovery. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27225–27245. PMLR, 2023. URL https://proceedings.mlr.press/v202/park23h.html.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, 11(5-6):355–607, 2019. doi: 10.1561/2200000073. URL https://doi.org/10.1561/2200000073.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Mark Rowland, Jiri Hron, Yunhao Tang, Krzysztof Choromanski, Tamas Sarlos, and Adrian Weller. Orthogonal estimation of wasserstein distances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 186–195. PMLR, 2019.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations*, 2021.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*, pp. 8583–8592. PMLR, 2020.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. In *International Conference on Learning Representations*, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Harshinder Singh, Neeraj Misra, Vladimir Hnizdo, Adam Fedorowicz, and Eugene Demchuk. Nearest neighbor estimates of entropy. *American Journal of Mathematical and Management Sciences*, 23 (3-4):301–321, 2003a. doi: 10.1080/01966324.2003.10737616. URL https://doi.org/10.1080/01966324.2003.10737616.

Harshinder Singh, Neeraj Misra, Vladimir Hnizdo, Adam Fedorowicz, and Eugene Demchuk. Nearest neighbor estimates of entropy. *American Journal of Mathematical and Management Sciences*, 23 (3-4):301–321, 2003b.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems*, 2012.

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.

Yifan Wu, George Tucker, and Ofir Nachum. The laplacian in RL: learning representations with efficient approximations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJlNpoA5YQ.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, 2021.