Wasserstein Distance Maximizing Intrinsic Control

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper deals with the problem of learning a skill-conditioned policy that acts meaningfully in the absence of a reward signal. Mutual information based objectives have shown some success in learning skills that reach a diverse set of states in this setting. These objectives include a KL-divergence term, which is maximized by visiting distinct states even if those states are not far apart in the MDP. This paper presents an approach that rewards the agent for learning skills that maximize the Wasserstein distance of their state visitation from the start state of the skill. It shows that such an objective leads to a policy that covers more distance in the MDP than diversity based objectives, and validates the results on a variety of Atari environments.

1 Introduction

This paper considers the unsupervised reinforcement learning problem of learning a set of skill-conditioned policies that act meaningfully in an environment in the absence of an extrinsic reward signal. Some previous works [16, 13, 6] approached this problem by using a mutual information objective to maximize the empowerment of the skill-conditioned policies. In essence, such a mutual information objective is maximized by learning goal-conditioned policy and a discriminator such that the discriminator can infer which skill was executed by considering the states visited by the policy conditioned on that skill.

This type of objective has been shown to learn diverse skills which can be useful for exploration and heirarchical reinforcement learning (HRL) [13, 6]. However, one potential issue with mutual information-based objectives is that they can learn skills that are discriminable but do not move far from the agent's starting state [9].

This paper instead presents an approach which considers the Wasserstein distance between the state visitation distribution of the agent's skill-conditioned policy and its start state distribution and trains the agent to maximize this distance, an approach we term Wasserstein distance maximizing Intrinsic Control (WIC). WIC also encourages the learning of diverse skills by constructing the reward function to prefer each skill maximizing the Wasserstein distance in a unique direction. We hypothesize that maximizing the Wasserstein distance will lead to policies that cover more distance in the underlying environment. This hypothesis is validated on two grid world environments where the policy learned using WIC maximizes the number of states that it visits, whereas VIC and related techniques are content to reach states that are discernible from each other.

Finally, we end with some preliminary results on the Atari benchmark that suggest that WIC is a promising approach to unsupervised intrinsic control.

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

2 Related Work

Intrinsic motivations [3, 26, 25] are rewards presented by an agent to itself in addition to the external task-specific reward. Intrinsic motivation has been proposed as a way to encourage RL agents to learn skills [5, 4, 32, 29] that might be useful across a variety of tasks, or as a way to encourage exploration [7, 31, 2, 14]. The optimal reward framework [33, 35] and shaped rewards [23] (if generated by the agent itself) also consider intrinsic motivation as a way to assist an RL agent in learning the optimal policy for a given task. Such an intrinsically motivated reward signal has previously been learned through various methods such as evolutionary techniques [24, 30], meta-gradient approaches [34, 39, 40], and others. The Wasserstein distance, in particular, has been used to present a valid reward for speeding up learning of goal-conditioned policies [11], imitation learning [37, 10, 38], as well as program synthesis [15].

Mutual information based objectives have been used to learn skill-conditioned policies that act meaningfully in the absence of an external reward [16, 13, 9, 6]. This paper considers the same problem but uses the Wasserstein distance as the objective the agent seeks to maximize.

3 Background and Setup

In this section we set up the problem and go over some of the concepts relevant to the setting.

3.1 Problem Setting

Our environment is a special case of Markov decision processes without a reward function denoted by the tuple $\mathcal{M} : \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mu \rangle$. $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \longmapsto \triangle(\mathcal{S})$ is the conditional distribution denoting the transition dynamics when taking action $a \in \mathcal{A}$ from state $s \in \mathcal{S}$ ($\triangle$ denotes a distribution over the set given as the argument), and $\mu : \triangle(\mathcal{S})$ is the initial state distribution.

Agents interact with the environment with a skill conditioned policy $\pi_\theta : \mathcal{S} \times \Omega \longmapsto \triangle(\mathcal{A})$ where Ω is the space of skills, and θ denotes the parameters of the policy. We assume that skills are sampled with some probability $P(\omega)$ which we assume to be fixed as a uniform distribution over a discrete set of skills in this paper. In particular, we assume skills are sampled uniformly from set Ω and are then followed for a fixed number of T time steps. A skill-episode starts when a skill to be followed $\omega \in \Omega$ is sampled from the skill distribution, and continues for T time steps. The trajectory of states and actions that are obtained while the agent executing skill ω interacts with the environment is denoted by $s_0, a_0, s_1, \ldots, s_{T-1}, a_{T-1}, s_T$. s_T at the end of one skill episode acts as the first state s_0 for the next skill episode. We will refer to s_0 and s_T as the start state and end state of a skill episode respectively.

Finally we define the state visitation distribution of the policy π_θ conditioned on skill ω and starting at state s_0 as:

$$\rho_\theta(s|s_0, \omega) = \frac{1}{T} \sum_{t=1}^{T} P(s_t = s|\pi_\theta, s_0, \omega) \tag{1}$$

3.2 Intrinsic Control by maximizing Mutual Information

Variational Intrinsic Control (VIC) [16] takes the above setting and sets up the problem of learning the skill-conditioned policy as one of maximizing the mutual information between the random variable ω denoting the skill and the states S_T reached after executing the skill conditioned policy π_θ conditioned on ω. Practically, this approach is implemented by learning a discriminator $D_\phi(\omega|s_T, s_0)$ with parameters ϕ that tries to predict the skill the agent policy was conditioned on given the start and end states of the skill-episode.

The output of this discriminator is then used as the reward signal to train the skill-conditioned policy π_θ. This approach encourages the learning of skills that can be distinguished by the end states of their trajectories. However, it does not encourage the learning of skills that travel as far as possible in the environment.

Other works that utilize this mutual information objective use a similar setup, but encourage the discriminator to predict the skill based on the relative direction of the states reached compared to the start state (RVIC, Baumli et al. [6]) or try to discriminate the entire trajectory (DIAYN, Eysenbach et al. [13]).

3.3 Wasserstein Distance and Optimal Transport

The field of optimal transport [27] considers the question of how to transport one distribution to another while minimizing the amount of effort expended. The Wasserstein distance estimates the amount of work that needs to be done to convert one probability distribution to the other, as measured by the ground metric d. More concretely, consider a metric space $(\mathcal{M}, d)$ where $\mathcal{M}$ is a set and d is a metric on $\mathcal{M}$. The Wasserstein-p distance between two distributions μ and ν on $\mathcal{M}$ with finite moments can be defined as:

$$W_d^p(\mu, \nu) := \min_{\zeta \in Z} \mathbb{E}_{x,y\sim\zeta} \left[d(x, y)^p\right]^{1/p} \tag{2}$$

where Z is the space of joint distributions $\zeta \in \triangle(\mathcal{M} \times \mathcal{M})$ whose marginals are μ and ν respectively.

While prior work on using the Wasserstein distance in reinforcement learning has used Euclidean distance in the state space as the ground metric [15], it may not be appropriate since it does not reflect the structure of the reinforcement learning problem. In MDPs, if we consider the metric space on the set of states $\mathcal{S}$, then the number of time-steps it would take to go from state x to state y under the current agent policy π is a quasimetric (metric that might not be symmetric between x and y) that could be considered more appropriate for measuring this work [11, 18] since it reflects distance in the MDP instead of distance in observation space.

If we restrict our attention to the Wasserstein-1 metric, the Kantorovich-Rubinstein duality allows us to express the Wasserstein-1 distance (which we refer to simply as the Wasserstein distance hereafter) in the following manner:

$$W_d^1(\mu, \nu) = \sup_{Lip(f)\leq 1} \mathbb{E}_{y\sim\nu} \left[f(y)\right] - \mathbb{E}_{x\sim\mu} \left[f(x)\right] \tag{3}$$

where the supremum is over all 1-Lipschitz functions $f : \mathcal{M} \longmapsto \mathbb{R}$ in the metric space. If this metric space is based on the time-step metric alluded to above, then the Lipschitz constraint can be enforced using the following equation for Lipschitz smoothness based on transitions experienced by the policy [11]:

$$Lip(f) = \max_{s\in\mathcal{S}} \mathbb{E}_{s'\sim\pi,\mathcal{P}} \left[|f(s') - f(s)|\right] \tag{4}$$

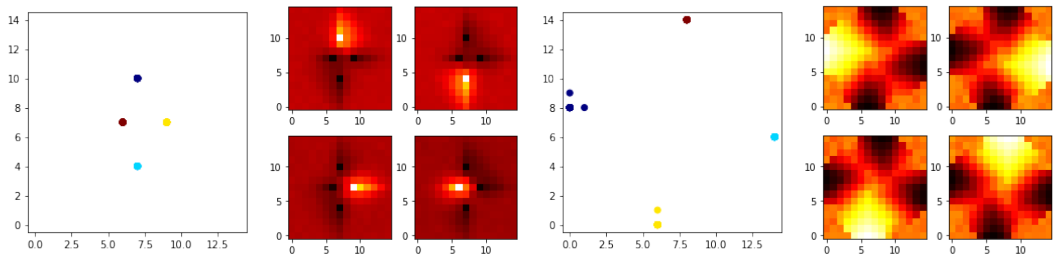
The Wasserstein distance between two distributions can then be estimated by means of a function approximator such as a neural network [1, 17, 11] by solving for equation 3 and ensuring smoothness according to Equation 4. In Section 4, we lay out the exact objective to train such a function approximator.

4 Wasserstein Distance Maximizing Intrinsic Control

This section describes how the Wasserstein distance can be used to learn a skill conditioned policy, an approach we term Wasserstein distance maximizing Intrinsic Control (WIC). At a high level, it proposes a method to learn a skill conditioned policy that attempts to get as far away from the skill's start state as measured through transitions in the MDP, and attempt to go in unique directions for each skill. That is, WIC will train a policy to maximize $\mathbb{E}_{\omega\sim P(\cdot|\Omega)} \left[W_d^1(\delta(s_0), \rho_\theta(s|s_0, \omega))\right]$, and penalize this policy for maximizing $\mathbb{E}_{\omega'\neq\omega} \left[W_d^1(\delta(s_0), \rho_\theta(s|s_0, \omega'))\right]$ for any other skill $\omega' \neq \omega$.

For a particular skill $\omega \in \Omega$ that starts executing at a state s_0, the above Wasserstein distances are estimated between a Dirac distribution at the skill's start state $\delta(s_0)$ and the skill's state visitation distribution $\rho_\theta(s|s_0, \omega)$ (Equation 1) with a potential function $f_\phi : \mathcal{S} \times \mathcal{S} \times \Omega \longmapsto \mathbb{R}$ with parameters ϕ. The potential function f_ϕ is trained by minimizing the following objective:

$$L_f(s_0, \rho_\theta, \omega) = f_\phi(s_0, s_0, \omega) - \mathbb{E}_{s\sim\rho_\theta(\cdot|s_0,\omega)} \left[f_\phi(s, s_0, \omega)\right] \tag{5}$$



(a) End points of skills learned by VIC

(b) Rewards learned by VIC

(c) End points of skills learned by WIC

(d) Rewards learned by WIC

Figure 1: The endpoints reached after executing all the skills multiple times starting in the middle of the room, and reward functions learned for VIC and WIC respectively in a 15×15 grid world where the features are a one-hot encoding. The agent starts executing each skill from the center of the grid world.

while also enforcing that the potential function f_ϕ is 1-Lipschitz with the following objective:

$$L_c(s_0, \rho_\theta, \omega) = \mathbb{E}_{s,s' \sim \rho_\theta(\cdot|s_0,\omega)} \text{maximum}(\|f_\phi(s', s_0, \omega) - f_\phi(s, s_0, \omega)\|^2 - 1, 0) \tag{6}$$

where s is a state drawn according to the state visitation distribution ρ_θ and s' is a sample of the next state the agent would visit if following policy π_θ conditioned on skill ω in that state. Maximizing the Wasserstein distance thus estimated will lead to a skill that attempts to get as far away from the start state as possible.

In order to maximize this distance, the agent is trained with rewards that encourage it to move its state visitation distribution to regions of higher potential, and thus increase the Wasserstein distance of the state visitation distribution. Since the potential function is state-based, it is enough for the reward to be a difference in potentials [23]. Further, WIC also includes a term to encourage diverse skills, meaning ones that move in unique directions in the state space. This diversity is encouraged by including a penalty term for overlapping with the positive potential gradient of any other skill. Consequently, the reward we present to the agent is as follows:

$$\begin{aligned} r(s_t, a_t, s_{t+1}, s_0, \omega) := [f(s_{t+1}, s_0, \omega) - f(s_t, s_0, \omega)] - \\ \eta \max_{\omega' \neq \omega} [f(s_{t+1}, s_0, \omega') - f(s_t, s_0, \omega')] \end{aligned} \tag{7}$$

where $\eta \in [0, 1]$ specifies how much of a penalty skills get for encouraging a direction that overlaps with other skills.

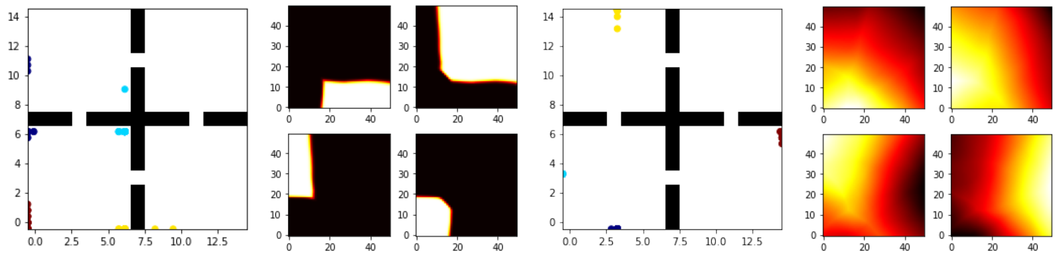
5 Experiments

We compare WIC with VIC in domains with increasing order of complexity in order to probe the difference between the skills learned by maximizing the Wasserstein distance to its start state as opposed to maximizing a mutual information objective with respect to a fixed skill sampling distribution.

5.1 Tabular

First, we evaluate WIC and VIC on a tabular grid world domain. The grid is 15×15 giving us 225 distinct states with 5 possible actions (up, down, left, right, and no-op), and the agent starts off in the center of the grid. Since this domain is tabular, the agent's state is communicated as a one-hot vector.

There are no obstacles, and $|\Omega| = 4$ skills which are sampled uniformly. Once the skill ω is sampled, the agent executes its skill-conditioned policy for $T = 10$ time steps. After executing this policy for T time steps, the state is reset back to the center, a new skill is sampled, and the policy executes again for T time steps. We compare two methods to learn and present the intrinsic reward, VIC and WIC. Both the agent policy and the discriminator (VIC) or potential function (WIC) are linear functions of the features. For WIC, a penalty $\eta = 0.9$ is used to avoid learning skills that overlap in their state visitation.



(a) End points of skills learned by VIC (b) Rewards learned by VIC (c) End points of skills learned by WIC (d) Rewards learned by WIC

Figure 2: Visualizing the endpoints after executing all the skills multiple times starting in the middle of the bottom-left room, and the reward functions used to learn them by VIC and WIC respectively in the Four Rooms environment. The features of the state here are the (x, y) coordinates of the point at which the agent is. The agent starts executing each skill from the center of the bottom left room.

The agent policy is trained using REINFORCE [36] with a state-conditioned baseline. The policy is additionally regularized with an entropy loss weighted by 0.01 to prevent premature convergence. Both the discriminator and the policy are trained using stochastic gradient descent [28, 19] with a learning rate of 0.003.

Figure 1 shows the states reached after executing each skill multiple times from the same start state at the middle of the room (Figure 1a) and the reward function used to train the policy (Figure 1b) for VIC. Figures 1c and 1d show the same respectively for WIC. As hypothesized, VIC is satisfied with reaching states that are distinct enough for the discriminator to tell apart, and does not necessarily attempt to learn a policy that travels far in the environment. WIC on the other hand, learns a reward function and a policy that attempt to travel as far away from the start state as possible.

5.2 Four Rooms

Next, we evaluate how WIC and VIC behave differently when the features are not one-hot vectors. We use a four room domain with the location of the agent communicated as its (x, y) coordinate, and each feature scaled to $[-1, 1]$. The agent starts in the center of the bottom left room and each skill is allowed $T = 40$ time steps to execute. This duration is enough for an agent to make it to the room diagonally opposite if the skill-conditioned policy is deterministic.

The number of skills $|\Omega| = 4$ and they are sampled uniformly. After the agent finishes executing one skill it samples a new skill and begins executing it from the state that was reached. The agent is reset to the middle of the bottom left room after sampling and executing skills 17 times.

Both the policy and the potential function (WIC) or discriminator (VIC) are instantiated as multi-layer perceptrons with 2 hidden layers of 128 units each, and ReLU [22] activation functions internally. The other training details remain the same as in the tabular case, except for the use of the Adam optimizer [20] with its default learning rate of 0.001.

The end states reached after executing the skill-conditioned policy for each skill multiple times from the middle of the bottom-left room are shown in Figure 2. Here again, we see that the discriminative objective of VIC is satisfied with learning skills that go to the corners of the bottom left room, whereas WIC learns a policy that travels deep into the adjoining two rooms. The reward functions visualized in both these domains makes it clear that the Wasserstein distance maximizing approach pushes the agent to go as far from the start state as it can.

5.3 Atari

So far, we have validated that WIC encourages the learning of skill-conditioned policies that try to go as far away from the state that the skill was invoked. We now apply WIC to the Atari domain [8, 21] and evaluate how well this approach scales to image based inputs and deeper function approximators. WIC is compared to VIC and RVIC, and the metrics we use for comparison are average episodic coverage, average lifetime coverage, and average episodic return.

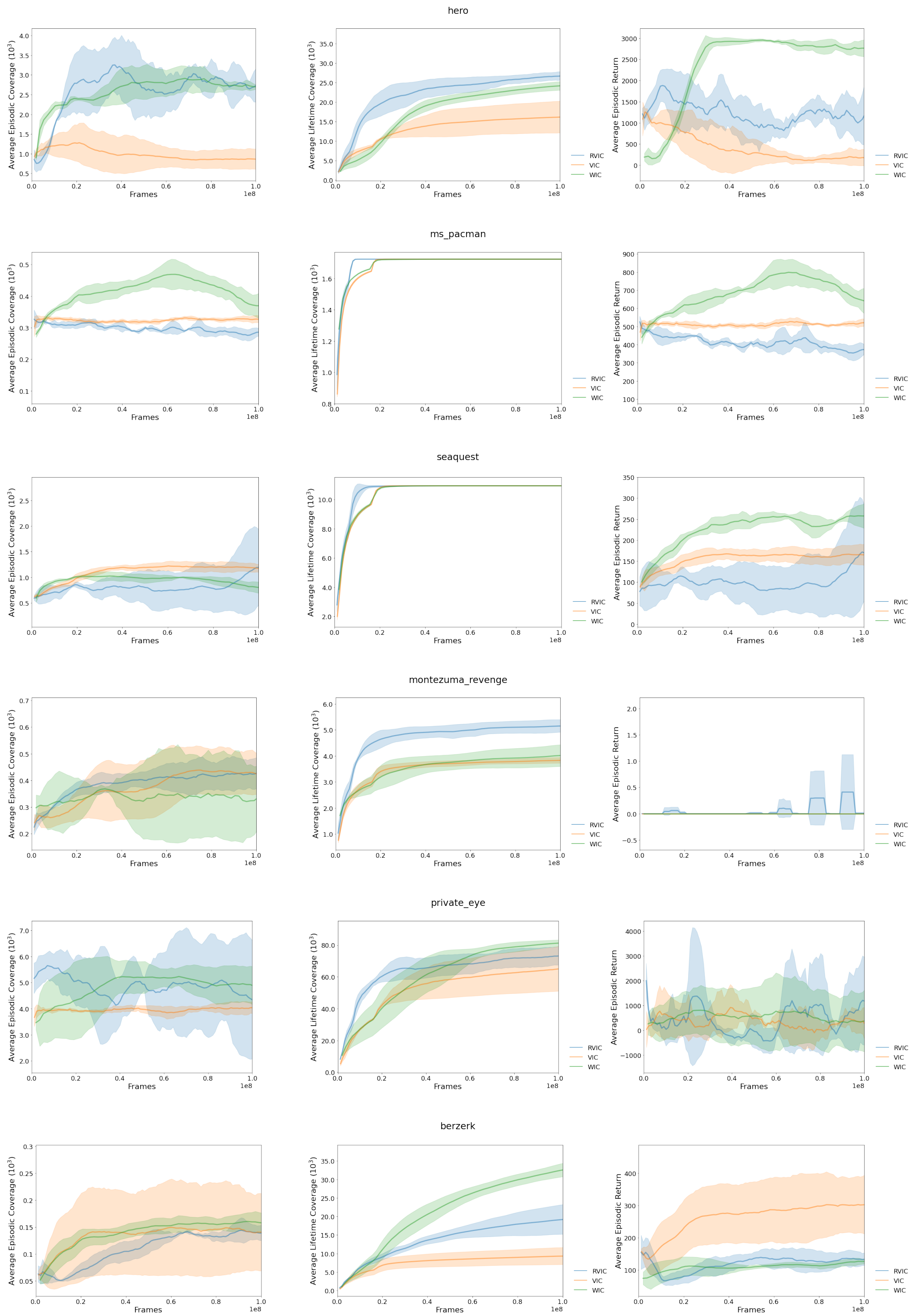


Figure 3: Comparing how well WIC does compared to mutual information based methods in Atari domains

The potential function (WIC) or discriminator (RVIC and VIC) are trained just as before, but the agent's policy is now instantiated as a Q-function trained using Q(λ) with $\lambda = 0.9$ and discount factor $\gamma = 0.98$. A replay buffer of size 5×10^5 is used to store the data, and each minibatch samples 64 trajectories of length $T = 40$ from the buffer to train from. The Q-function and the potential function share a common torso in this setting, and the architecture of this torso is equivalent to the one suggested in IMPALA [12].

As can be seen from Figure 3, the skill-conditioned policy learned through WIC leads to episodic returns better than VIC or RVIC on three of the six games we test on: Hero, Montezuma's Revenge, and Ms. Pacman, and roughly equivalent returns on the other three. These improved returns are indicative of more directed policies, even in games where the episodic coverage is similar to the mutual information based approaches (Hero and Seaquest). In the game Berzerk, we additionally see that even though the episodic returns do not outperform VIC, in terms of lifetime coverage WIC widens the gap over time.

6 Conclusion

This paper considers the question of unsupervised learning of skills in an environment, and hypothesizes that maximizing the Wasserstein distance from the start state distribution of a skill could lead to skill-conditioned policies that cover more distance in the underlying MDP than mutual information based approaches like VIC and RVIC. This approach is crystallized as Wasserstein distance maximizing intrinsinc control (WIC), and the above hypothesis is validated on a tabular grid world as well as a continuous four rooms domain. Finally, we have validated that the approach scales up to visual inputs and complex environments by evaluating it on the Atari domain.

References

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

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

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

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

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

[6] Kate Baumli, David Warde-Farley, Steven Hansen, and Volodymyr Mnih. Relative variational intrinsic control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6732–6740, 2021.

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

[8] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[9] Víctor 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 *International Conference on Machine Learning*, pages 1317–1327. PMLR, 2020.

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

[11] Ishan Durugkar, Mauricio Tec, Scott Niekum, and Peter Stone. Adversarial intrinsic motivation for reinforcement learning. *arXiv preprint arXiv:2105.13345*, 2021. URL `https://arxiv.org/abs/2105.13345`.

[12] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pages 1407–1416. PMLR, 2018.

[13] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

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

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

[16] Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

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

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

[19] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[21] Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

[22] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*, 2010.

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

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

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

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

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

[28] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

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

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

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

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

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

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

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

[36] David Williams. *Probability with martingales*. Cambridge university press, 1991.

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

[38] M. Zhang, Y. Wang, X. Ma, L. Xia, J. Yang, Z. Li, and X. Li. Wasserstein distance guided adversarial imitation learning with reward shape exploration. In *2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS)*, pages 1165–1170, Nov 2020. doi: 10.1109/DDCLS49620.2020.9275169.

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

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