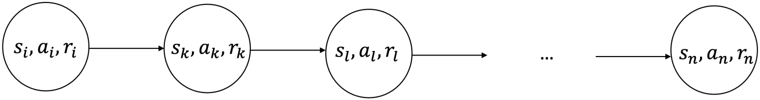
Imitation Learning with Sinkhorn Distances

Georgios Papagiannis[1] and Yunpeng Li[2]
[1]Imperial College London and [2]University of Surrey

ECML PKDD 2022
Grenoble, France
September, 2022

Reinforcement learning (RL)

Standard Markov Decision Process



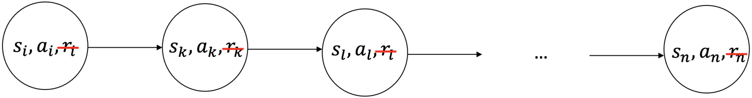
Reinforcement learning (RL) without rewards

Often reward is unavailable or hard to define

$$s_i, a_i, \cancel{r_i} \rightarrow s_k, a_k, \cancel{r_k} \rightarrow s_l, a_l, \cancel{r_l} \rightarrow \quad \dots \quad \rightarrow s_n, a_n, \cancel{r_n}$$

Reinforcement learning (RL) without rewards

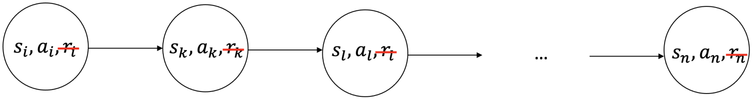
Often reward is unavailable or hard to define



- Instead, **learn** from demonstrations

Reinforcement learning (RL) without rewards

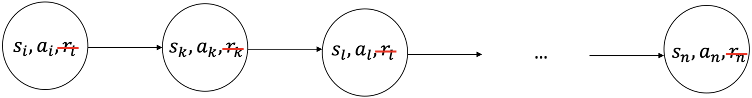
Often reward is unavailable or hard to define



- Instead, **learn** from demonstrations
- Inverse RL: Explicitly infer reward, optimize with RL (**ill posed, computationally expensive)**

Reinforcement learning (RL) without rewards

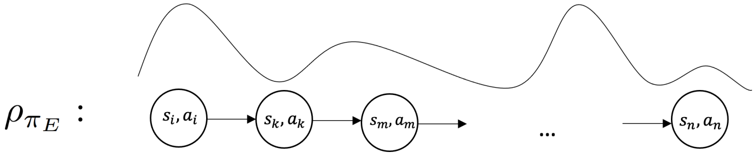
Often reward is unavailable or hard to define



- Instead, **learn** from demonstrations
- Inverse RL: Explicitly infer reward, optimize with RL (**ill posed, computationally expensive)**
- Imitation learning: Learn from demonstration directly, without explicit reward inference

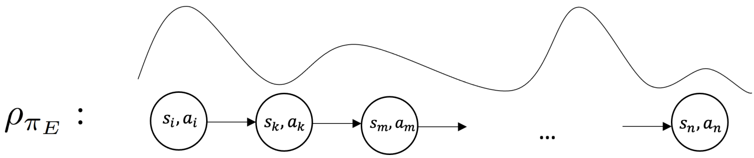
Imitation learning

Demonstrator policy π_E with occupancy measure ρ_{π_E}:

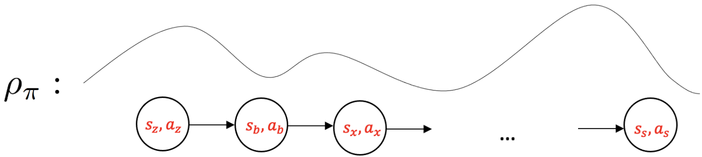


Imitation learning

Demonstrator policy π_E with occupancy measure ρ_{π_E}:

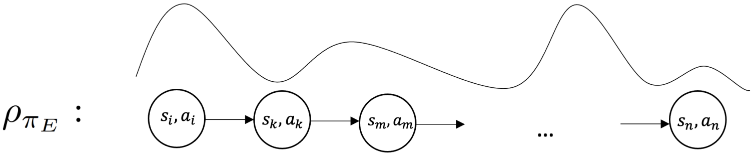


Learner policy π with occupancy measure ρ_π:

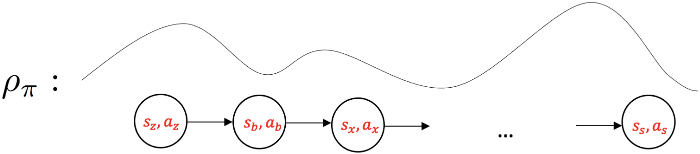


Imitation learning

Demonstrator policy π_E with occupancy measure ρ_{π_E}:



Learner policy π with occupancy measure ρ_π:



- Measure similarity with metric $\mathcal{D}(\rho_\pi, \rho_{\pi_E})$

Imitation learning

Objective: Find π such that
$\mathcal{D}(\rho_\pi, \rho_{\pi_E})$ is minimized.

Imitation learning

Objective: Find π such that
$\mathcal{D}(\rho_\pi, \rho_{\pi_E})$ is minimized.

The distribution of state-action pairs of the **demonstrator** and the **learner** policies are the **same**, hence the learner has **imitated** the expert

Different similarity metrics $\mathcal{D}(\rho_\pi, \rho_{\pi_E})$

- Supervised learning: Behaviour Cloning (BC)

Different similarity metrics $\mathcal{D}(\rho_\pi, \rho_{\pi_E})$

- Supervised learning: Behaviour Cloning (BC)
- Kullback-Leibler Divergence: Adversarial Inverse RL (AIRL)[1]

[1] Fu et al., "Learning Robust Rewards with Adversarial Inverse Reinforcement Learning", ICLR 2018

Different similarity metrics $\mathcal{D}(\rho_\pi, \rho_{\pi_E})$

- Supervised learning: Behaviour Cloning (BC)
- Kullback-Leibler Divergence: Adversarial Inverse RL (AIRL)[1]
- Jensen-Shannon divergence: Generative Adversarial Imitation Learning (GAIL)[2]

[1] Fu et al., "Learning Robust Rewards with Adversarial Inverse Reinforcement Learning", ICLR 2018

[2] Ho and Ermon ,"Generative adversarial imitation learning", NIPS 2016

Different similarity metrics $\mathcal{D}(\rho_{\pi}, \rho_{\pi_E})$

- Supervised learning: Behaviour Cloning (BC)
- Kullback-Leibler Divergence: Adversarial Inverse RL (AIRL)[1]
- Jensen-Shannon divergence: Generative Adversarial Imitation Learning (GAIL)[2]
- ... and any $f-$divergence[3]

[1] Fu et al., "Learning Robust Rewards with Adversarial Inverse Reinforcement Learning", ICLR 2018

[2] Ho and Ermon ,"Generative adversarial imitation learning", NIPS 2016

[3] Ghasemipour et al., "A Divergence Minimization Perspective on Imitation Learning", CORL 2019

Different similarity metrics $\mathcal{D}(\rho_{\pi}, \rho_{\pi_E})$

- Supervised learning: Behaviour Cloning (BC)
- Kullback-Leibler Divergence: Adversarial Inverse RL (AIRL)[1]
- Jensen-Shannon divergence: Generative Adversarial Imitation Learning (GAIL)[2]
- ... and any $f-$divergence[3]
- Dual Wasserstein: Wassertein Adversarial Imitation Learning[4]

[1] Fu et al., "Learning Robust Rewards with Adversarial Inverse Reinforcement Learning", ICLR 2018

[2] Ho and Ermon ,"Generative adversarial imitation learning", NIPS 2016

[3] Ghasemipour et al., "A Divergence Minimization Perspective on Imitation Learning", CORL 2019

[4] Xiao et al., "Wasserstein Adversarial Imitation Learning", arXiv 2019

Different similarity metrics $\mathcal{D}(\rho_\pi, \rho_{\pi_E})$

- Supervised learning: Behaviour Cloning (BC)
- Kullback-Leibler Divergence: Adversarial Inverse RL (AIRL)[1]
- Jensen-Shannon divergence: Generative Adversarial Imitation Learning (GAIL)[2]
- ... and any $f-$divergence[3]
- Dual Wasserstein: Wassertein Adversarial Imitation Learning[4]
- Bounded Wasserstein: Primal Wasserstein Imitation Learning[5]

---

[1] Fu et al., "Learning Robust Rewards with Adversarial Inverse Reinforcement Learning", ICLR 2018

[2] Ho and Ermon ,"Generative adversarial imitation learning", NIPS 2016

[3] Ghasemipour et al., "A Divergence Minimization Perspective on Imitation Learning", CORL 2019

[4] Xiao et al., "Wasserstein Adversarial Imitation Learning", arXiv 2019

[5] Dadashi et al., "Primal Wasserstein Imitation Learning", ICLR 2021

Different similarity metrics $\mathcal{D}(\rho_\pi, \rho_{\pi_E})$

Limitations:

Different similarity metrics $\mathcal{D}(\rho_\pi, \rho_{\pi_E})$

Limitations:

- Do not account for the distributions' metric space

Different similarity metrics $\mathcal{D}(\rho_\pi, \rho_{\pi_E})$

Limitations:

- Do not account for the distributions' metric space
- Not robust to disjoint measures

Different similarity metrics $\mathcal{D}(\rho_\pi, \rho_{\pi_E})$

Limitations:

- Do not account for the distributions' metric space
- Not robust to disjoint measures
- Often solved with generative adversarial training, inheriting its disadvantages such as training instability

Different similarity metrics $\mathcal{D}(\rho_\pi, \rho_{\pi_E})$

Limitations:

- Do not account for the distributions' metric space
- Not robust to disjoint measures
- Often solved with generative adversarial training, inheriting its disadvantages such as training instability
- Intractable

Different similarity metrics $\mathcal{D}(\rho_\pi, \rho_{\pi_E})$

Limitations:

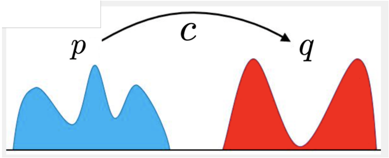
- Do not account for the distributions' metric space
- Not robust to disjoint measures
- Often solved with generative adversarial training, inheriting its disadvantages such as training instability
- Intractable
- Locally Optimal

Wasserstein Distance

Given two probability measures p, q, the k-Wasserstein[6] distance calculates the minimal transportation cost of moving measure p to measure q:

$$\mathcal{W}(\rho_\pi, \rho_{\pi_E})_c = \left(\inf_{\zeta \in \Omega(p,q)} \int c(x,y)^k d\zeta(x,y) \right)^{\frac{1}{k}}$$

where ζ corresponds to the optimal transport plan and $\Omega(p,q)$ the set of all joint distributions whose marginals correspond to p and q.



[6] Villani. "Optimal Transport: old and new", volume 338, Springer Science & Business Media, 2008

Sinkhorn Distance[7]

$$\mathcal{W}_s^{\beta}(\rho_\pi, \rho_{\pi_E})_c = \inf_{\zeta_\beta \in \Omega_\beta(p,q)} \mathbb{E}_{x,y\sim\zeta_\beta}\left[c(x,y)\right]$$

where $\Omega_\beta(p,q)$ denotes the set of all joint distributions in $\Omega(p,q)$ with entropy of at least $\mathcal{H}(p) + \mathcal{H}(q) - \beta$.

[7] Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances", NIPS 2013

Sinkhorn Distance[7]

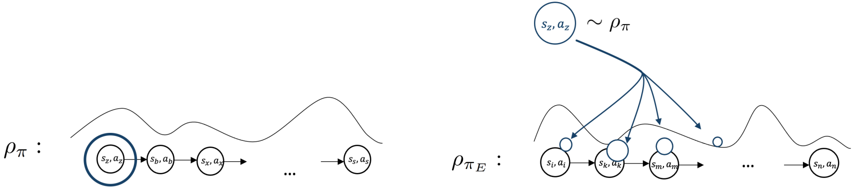
$$\mathcal{W}_s^{\beta}(\rho_{\pi}, \rho_{\pi_E})_c = \inf_{\zeta_\beta \in \Omega_\beta(p,q)} \mathbb{E}_{x,y\sim\zeta_\beta}\left[c(x,y)\right]$$

where $\Omega_\beta(p,q)$ denotes the set of all joint distributions in $\Omega(p,q)$ with entropy of at least $\mathcal{H}(p) + \mathcal{H}(q) - \beta$.

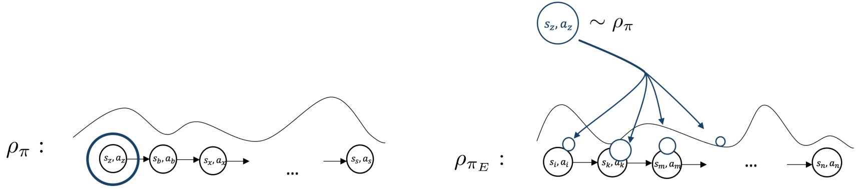
- Accounts for the distributions' metric space
- Is robust to disjoint measures
- Improves training stability
- Tractable
- Globally Optimal

[7] Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances", NIPS 2013

Sinkhorn Distance in Imitation Learning



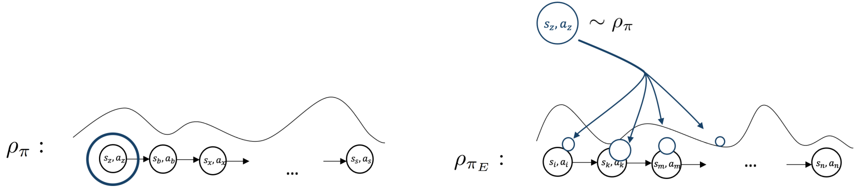
Sinkhorn Distance in Imitation Learning



Sample Transport Cost:

$$v_c\left(s_z, a_z \sim \rho_\pi \right) = \sum_{s_i, a_i \sim \rho_{\pi_E}} c\overbrace{\left(s_z, a_z \,,\, s_i, a_i \right)}^{\text{Distance cost}} \; \zeta_\beta \underbrace{\left(s_z, a_z \,,\, s_i, a_i \right)}_{\text{Optimal Transport Plan}}$$

Sinkhorn Distance in Imitation Learning



Sample Transport Cost:

$$v_c\left(s_z, a_z \sim \rho_\pi \right) = \sum_{s_i, a_i \sim \rho_{\pi_E}} c \overbrace{\left(s_z, a_z , s_i, a_i \right)}^{\text{Distance cost}} \zeta_\beta \underbrace{\left(s_z, a_z , s_i, a_i \right)}_{\text{Optimal Transport Plan}}$$

$$\mathcal{W}_s^\beta(\rho_\pi, \rho_{\pi_E})_c = \sum_{s_z, a_z \sim \rho_\pi} v_c\left(s_z, a_z \right)$$

Sinkhorn Imitation Learning (SIL)

$$\mathcal{W}_{\boldsymbol{s}}^{\beta}(\rho_{\pi}, \rho_{\pi_E})_{c_w} = \sum_{s_z, a_z \sim \rho_{\pi}} v_{c_w}(s_z, a_z)$$

$-v_{c_w}$ per sample reward proxy in reinforcement learning

Sinkhorn Imitation Learning (SIL)

$$\mathcal{W}_{\boldsymbol{s}}^{\beta}(\rho_\pi, \rho_{\pi_E})_{c_w} = \sum_{(s_z, a_z) \sim \rho_\pi} v_{c_w}((s_z, a_z))$$

$-v_{c_w}$ per sample reward proxy in reinforcement learning

SIL's Optimization Objective:

$$\min_{\pi} \max_{w} \mathcal{W}_s^{\beta}(\rho_\pi, \rho_{\pi_E})_{c_w}$$

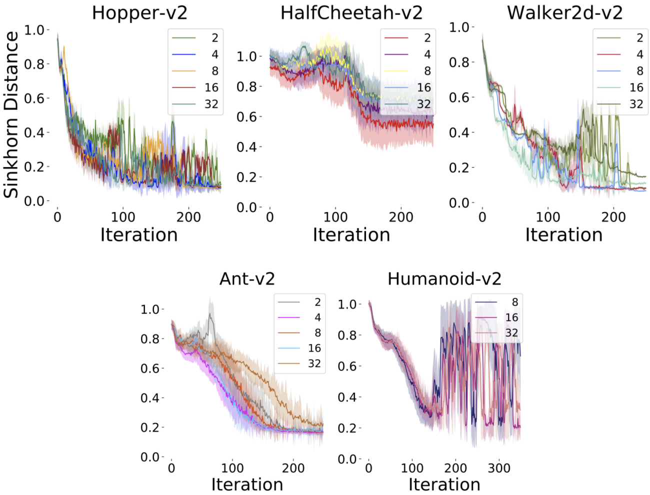
- Cost learned using a neural network (NN) parameterized by w
- Cosine distance between the output of the NN for each state-action pair

Sinkhorn Imitation Learning (SIL) Algorithm

Repeat to convergence:

Step 1: Optimize w parameterised as a NN to maximize $\mathcal{W}_s^{\beta}(\rho_{\pi}, \rho_{\pi_E})_{c_w}$
Step 2: Optimise π to minimize $\mathcal{W}_s^{\beta}(\rho_{\pi}, \rho_{\pi_E})_{c_w}$ using $-v_{c_w}$ as reward and any on-policy reinforcement learning (RL) algorithm

Results Overview



- Successful imitation learning with various numbers of demonstrations

Results Overview

Best performance on each experiment against benchmarks

Environments	Trajectories	BC	GAIL	AIRL	SIL
Hopper-v2	2	×	×	✓	×
	4	×	×	✓	×
	8	×	×	✓	×
	16	×	×	✓	×
	32	×	✓	×	×
HalfCheetah-v2	2	×	×	×	✓
	4	×	×	×	✓
	8	×	×	×	✓
	16	×	✓	×	×
	32	×	×	×	✓
Walker2d-v2	2	×	×	✓	×
	4	×	×	✓	×
	8	×	×	✓	×
	16	×	×	✓	×
	32	×	×	✓	×
	2	×	×	×	✓

	Trajectories	BC	GAIL	AIRL	SIL
Ant-v2	4	×	×	×	✓
	8	×	×	×	✓
	16	×	×	×	✓
	32	×	×	✓	×
Humanoid-v2	8	✓	×	×	×
	16	×	×	×	✓
	32	×	✓	×	×

- SIL performs SOTA against benchmarks on some environments; on par on the rest

SIL as a regularized maximum entropy Inverse RL framework

- Previous derivation of SIL is the most intuitive.
- SIL can also be derived by regularizing the objective of the maximum entropy Inverse Reinforcement Learning framework.
- Proof of the derivation available in the paper.

Summary

- Formulated Imitation learning as minimization of the Sinkhorn Distance.
- Proposed Sinkhorn Imitation Learning, SIL, a new Imitation learning method that minimizes the Sinkhorn Distance between occuppancy measures, is tractable and bypasses the drawbacks of $f-$divergence formulations.
- Derived and proved how SIL falls under the regularized maximum entropy Inverse RL frame.
- Obtained competitive or better performance on popular on-policy RL benchmarks.