NEURAL INFORMATION PROCESSING SYSTEMS

Robust Imitation via
Mirror Descent Inverse Reinforcement Learning

NeurIPS 2022

Dong-Sig Han Hyunseo Kim Hyundo Lee

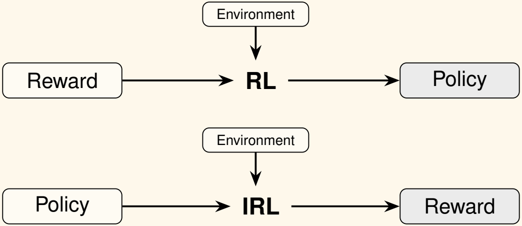
Je-Hwan Ryu Byoung-Tak Zhang

Artificial Intelligence Institute, Seoul National University

AIIS
Artificial Intelligence Institute
Seoul National University

VERI LUX
TAS MEA

Problem formulation



Reinforcement Learning (RL) & Inverse Reinforcement Learning (IRL)

Imitation Learning Problem: *Apprenticeship Learning via IRL*

Question

Can we generalize modern IRL algorithms and improve them upon the rich foundation of optimization studies?

Motivation

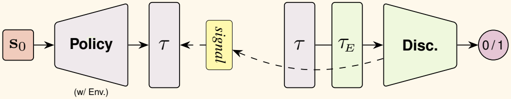
Mirror Descent (MD)[1]

For sequences $\{w_t\}_{t=1}^T$, $\{F_t\}_{t=1}^T$, and a convex function Ω:

$$\nabla\Omega(w_{t+1}) = \nabla\Omega\big(w_t\big) - \eta_t \nabla F_t\big(w_t\big)$$

$\nabla\Omega$ *links* the parametric space of $w_t \in \mathcal{W}$ to the dual space.

Adversarial Imitation Learning (AIL)[2]



- AIL tries to solve an optimization problem "directly."
- AIL does not analyze the convergence with unreliable trajectories in real-world problems.
- Through the lens of geometries, AIL does not ensure unbiased progression of its cost.

[1] Nemirovsky & Yudin (1979). Complexity of Problems and Efficiency of Optimization Methods

[2] Ho & Ermon (2016). Generative Adversarial Imitation Learning. In NeurIPS

Imitation learning in regularized MDPs

Let the cost be represented with the **Bregman divergence**[3]

With the given action space $\mathcal{A}$, it is defiend as

$$D_\Omega(\pi^s \| \hat{\pi}^s) := \Omega(\pi^s) - \Omega(\hat{\pi}^s) - \left\langle \nabla\Omega(\hat{\pi}^s),\ \pi^s - \hat{\pi}^s \right\rangle_{\mathcal{A}},$$

where π^s and $\hat{\pi}$ denote arbitrary policies for a given state s.

* Many of the AIL models can be understood with Bregman divergences[4].

Definition 1. (Regularized reward operators)

Define the regularized reward operator Ψ_Ω as

$$\psi_\pi(s,a) := \Omega'(s,a;\pi) - \left\langle \pi^s,\ \nabla\Omega(\pi^s) \right\rangle_{\mathcal{A}} + \Omega(\pi^s),$$

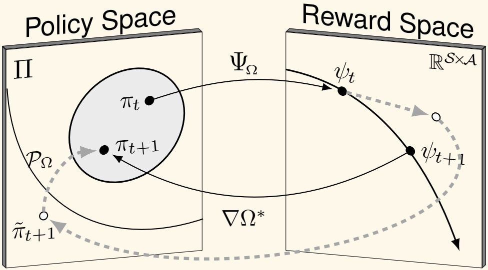
for $\Omega'(s,\cdot\,;\pi) := \nabla\Omega(\pi^s) = \left[\nabla_p\Omega(p)\right]_{p=\pi(\cdot|s)}$.

$\Rightarrow$ RL of π with reward function $\psi_{\hat{\pi}}$ is equivalent to minimizing $D_\Omega(\pi^s \| \hat{\pi}^s)$.

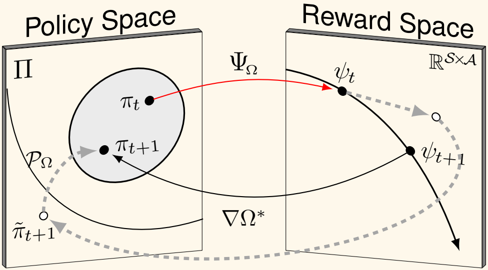
[3] Bregman (1969). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming.

[4] Jeon et al. (2021). Regularized Inverse Reinforcement Learning. In ICLR.

The MD-IRL theory: RL-IRL as a proximal method

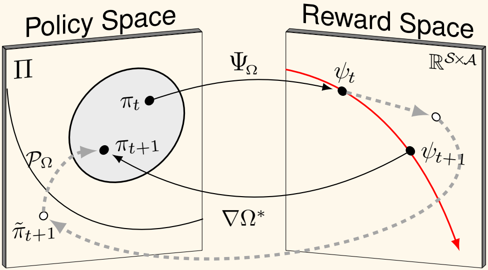


The MD-IRL theory: RL-IRL as a proximal method



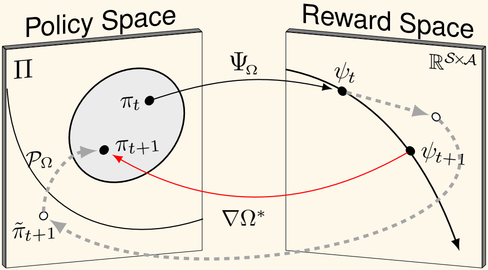
1. Policy Space ▷ Reward Space ($\psi_t \in \Psi_\Omega(\Pi)$).

The MD-IRL theory: RL-IRL as a proximal method



1. Policy Space $\triangleright$ Reward Space ($\psi_t \in \Psi_\Omega(\Pi)$).
2. Update rewards using MD update rules (MD-IRL).

The MD-IRL theory: RL-IRL as a proximal method



1. Policy Space $\triangleright$ Reward Space ($\psi_t \in \Psi_\Omega(\Pi)$).
2. Update rewards using MD update rules (MD-IRL).
3. Reward Space $\triangleright$ Policy Space ($\nabla\Omega^*$, typically by RL).

MD update rules

The proximal form of the MD update is alternatively written as[5]

$$\underset{w\in\mathcal{W}}{\text{minimize}}\,\big\langle \nabla F_t(w_t),\, w - w_t \big\rangle_{\mathcal{W}} + \alpha_t D_\Omega\big(w \,\big\|\, w_t\big),$$

where $\alpha_t := {}^1/_{\eta_t}$ denotes an inverse of the current step size η_t.

We hypothesize on existence of a random process $\{\bar{\pi}_{E,t}\}_{t=1}^{\infty}$ where each estimation $\bar{\pi}_{E,t}$ resides in a closed, convex neighborhood of π_E, generated by an arbitrary estimation algorithm.

Then, the cost is $D_\Omega\big(\pi_t^s \,\big\|\, \bar{\pi}_{E,t}^s\big)$, thus update are derived by solving the problem:

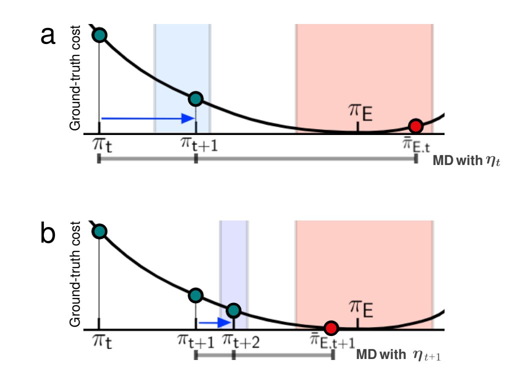
$$\begin{aligned}
&\underset{\pi^s\in\Pi^s}{\text{minimize}}\,\big\langle \underbrace{\nabla D_\Omega\big(\pi_t^s \,\big\|\, \bar{\pi}_{E,t}^s\big)}_{\nabla\Omega(\pi_t^s)-\nabla\Omega(\bar{\pi}_{E,t}^s)},\, \pi^s - \pi_t^s \big\rangle_{\mathcal{A}} + \alpha_t\, D_\Omega\big(\pi^s \,\big\|\, \pi_t^s\big) \\
&\iff \quad \underset{\pi^s\in\Pi^s}{\text{minimize}}\, D_\Omega\big(\pi^s \big\| \bar{\pi}_{E,t}^s\big) - D_\Omega\big(\pi^s \big\| \pi_t^s\big) + \alpha_t D_\Omega\big(\pi^s \big\| \pi_t^s\big) \\
&\iff \quad \underset{\pi^s\in\Pi^s}{\text{minimize}}\, \eta_t \underbrace{D_\Omega\big(\pi^s \big\| \bar{\pi}_{E,t}^s\big)}_{\text{estimated expert}} + (1-\eta_t)\, \underbrace{D_\Omega\big(\pi^s \big\| \pi_t^s\big)}_{\text{learning agent}} \qquad \forall s\in\mathcal{S},
\end{aligned}$$

where the gradient of D_Ω is taken with respect to its first argument π_t^s.

---

[5] Beck et al. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization.

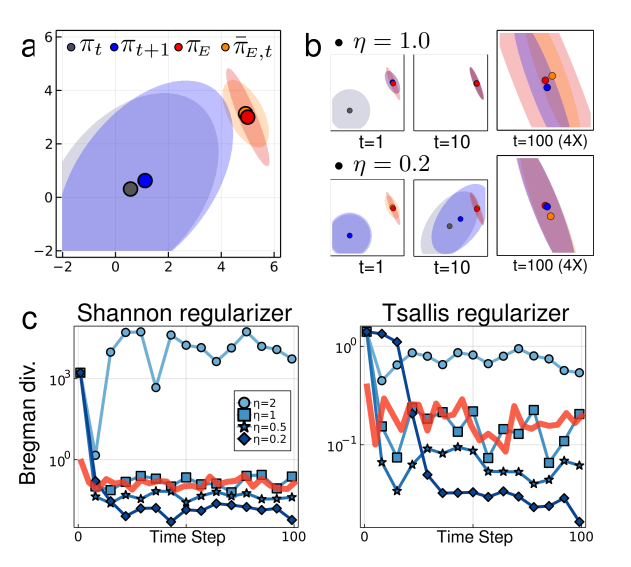
Online mirror descent on imitation learning



Illustrations of an MD-IRL process.

$$\underset{\pi^s \in \Pi^s}{\text{minimize}} \; \eta_t \underbrace{D_\Omega(\pi^s \| \bar{\pi}^s_{E,t})}_{\text{estimated expert}} + (1-\eta_t) \underbrace{D_\Omega(\pi^s \| \pi^s_t)}_{\text{learning agent}}$$

Online mirror descent on imitation learning agents



Examples of MD on Gaussian policy distributions.

Convergence analyses

Define a temporal cost function at the time step t as

$$f(\pi_t, \tau_t) \coloneqq \sum_{i=0}^{\infty} \gamma^i D_\Omega\big(\pi_t(\,\cdot \mid s_i^{(t)}) \big\| \bar{\pi}_{E,t}(\,\cdot \mid s_i^{(t)})\big),$$

Theorem 1 (Stepsize).

… $\lim_{T\to\infty} \mathbb{E}_{\tau_{1:T}}\big[\sum_{i=0}^{\infty} D_\Omega\big(\pi_*(\cdot|s_i)\big\|\pi_T(\cdot|s_i)\big)\big] = 0$ if and only if **a step size condition** is satisfied.

1. If $\lim_{t\to\infty} \eta_t = 0$, then $T \in \mathbb{N}$, $n < T$, and $c > 0$ exist s.t. $\mathbb{E}_{\tau_{1:T}}\big[f_T(\pi_T, \tau_T)\big] \geq \frac{c}{T-n}$.
2. If $\{\eta_t\}_{t\in\mathbb{R}^+}$ is $\eta_t = \frac{4}{t+1}$, then $\mathbb{E}_{\tau_{1:T}}\big[\sum_{i=0}^{\infty} D_\Omega\big(\pi_*(\cdot|s_i)\big\|\pi_T(\cdot|s_i)\big)\big] = \mathcal{O}(1/T)$.

Convergence analyses

Define a temporal cost function at the time step t as

$$f(\pi_t, \tau_t) \coloneqq \sum_{i=0}^{\infty} \gamma^i D_\Omega\Big(\pi_t(\,\cdot \mid s_i^{(t)}) \Big\| \, \bar{\pi}_{E,t}(\,\cdot \mid s_i^{(t)})\Big),$$

Theorem 1 (Stepsize).

... $\lim_{T\to\infty} \mathbb{E}_{\tau_{1:T}}\big[\sum_{i=0}^{\infty} D_\Omega\big(\pi_*(\cdot|s_i)\big\|\pi_T(\cdot|s_i)\big)\big] = 0$ if and only if **a step size condition** is satisfied.

1. If $\lim_{t\to\infty} \eta_t = 0$, then $T \in \mathbb{N}$, $n < T$, and $c > 0$ exist s.t. $\mathbb{E}_{\tau_{1:T}}\big[f_T(\pi_T, \tau_T)\big] \geq \frac{c}{T-n}$.
2. If $\{\eta_t\}_{t\in\mathbb{R}^+}$ is $\eta_t = \frac{4}{t+1}$, then $\mathbb{E}_{\tau_{1:T}}\big[\sum_{i=0}^{\infty} D_\Omega\big(\pi_*(\cdot|s_i)\big\|\pi_T(\cdot|s_i)\big)\big] = \mathcal{O}(1/T)$.

Theorem 2 (Optimal cases).

Assume $\pi_1 \neq \pi_E$ and $\inf_{\pi\in\Pi} \mathbb{E}[f(\pi, \tau_t)] = 0$. Then, $\mathbb{E}\big[f(\pi_t, \tau_t)\big] = 0$ if and only if $\sum_{t=1}^{\infty} \eta_t = \infty$. If $\eta_t \equiv \eta_1$, then there exist $c_1, c_2 \in (0, 1)$ such that $c_1^{T-1} \cdot A_1 \leq A_T \leq c_2^{T-1} \cdot A_1$, for $A_t = \sup_{s\in\mathcal{S}} \mathbb{E}_{\tau_{1:t}}\big[D_\Omega(\pi_E^s \| \pi_t^s)\big]$.

Convergence analyses

Define a temporal cost function at the time step t as

$$f(\pi_t, \tau_t) \coloneqq \sum_{i=0}^{\infty} \gamma^i D_\Omega\Big(\pi_t(\,\cdot \mid s_i^{(t)})\Big\|\, \bar{\pi}_{E,t}(\,\cdot \mid s_i^{(t)})\Big),$$

Theorem 1 (Stepsize).

... $\lim_{T\to\infty} \mathbb{E}_{\tau_{1:T}}\Big[\sum_{i=0}^{\infty} D_\Omega\Big(\pi_*(\cdot|s_i)\Big\|\pi_T(\cdot|s_i)\Big)\Big] = 0$ if and only if **a step size condition** is satisfied.

1. If $\lim_{t\to\infty} \eta_t = 0$, then $T \in \mathbb{N}$, $n < T$, and $c > 0$ exist s.t. $\mathbb{E}_{\tau_{1:T}}\Big[f_T(\pi_T, \tau_T)\Big] \geq \frac{c}{T-n}$.
2. If $\{\eta_t\}_{t\in\mathbb{R}+}$ is $\eta_t = \frac{4}{t+1}$, then $\mathbb{E}_{\tau_{1:T}}\Big[\sum_{i=0}^{\infty} D_\Omega\Big(\pi_*(\cdot|s_i)\Big\|\pi_T(\cdot|s_i)\Big)\Big] = \mathcal{O}(1/T)$.

Theorem 2 (Optimal cases).

Assume $\pi_1 \neq \pi_E$ and $\inf_{\pi\in\Pi} \mathbb{E}[f(\pi, \tau_t)] = 0$. Then, $\mathbb{E}\Big[f(\pi_t, \tau_t)\Big] = 0$ if and only if $\sum_{t=1}^{\infty} \eta_t = \infty$. If $\eta_t \equiv \eta_1$, then there exist $c_1, c_2 \in (0,1)$ such that $c_1^{T-1} \cdot A_1 \leq A_T \leq c_2^{T-1} \cdot A_1$, for $A_t = \sup_{s\in\mathcal{S}} \mathbb{E}_{\tau_{1:t}}\Big[D_\Omega(\pi_E^s \| \pi_t^s)\Big]$.

Proposition 1 (General cases).

Assume that $\pi_E \notin \Pi$, hence $\inf_{\pi\in\Pi} \mathbb{E}[f(\pi, \tau_t)] > 0$. If the step sizes satisfies **the proposed step size conditions**, then $\lim_{t\to\infty} \sum_{i=0}^{\infty} \gamma^i D_\Omega\Big(\pi_*(\,\cdot\,|s_i)\Big\|\pi_t(\,\cdot\,|s_i)\Big)$ converges to 0 almost surely.

Convergence analyses

Define a temporal cost function at the time step t as

$$f(\pi_t, \tau_t) \coloneqq \sum_{i=0}^{\infty} \gamma^i D_\Omega\big(\pi_t(\,\cdot \mid s_i^{(t)}) \,\big\|\, \bar{\pi}_{E,t}(\,\cdot \mid s_i^{(t)})\big),$$

Step size considerations

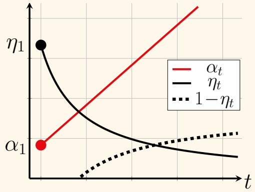
Two conditions of $\{\eta_t\}_{t=1}^{\infty}$ to guarantee convergence.

- Convergent sequence & divergent series:

  $\lim_{t\to\infty} \eta_t = 0$ and $\sum_{t=1}^{\infty} \eta_t = \infty.$

- Convergent series of squared terms:

  $\sum_{t=1}^{\infty} \eta_t = \infty$ and $\sum_{t=1}^{\infty} \eta_t^2 < \infty.$



Convergence analyses

Define a temporal cost function at the time step t as

$$f(\pi_t, \tau_t) \coloneqq \sum_{i=0}^{\infty} \gamma^i D_\Omega\big(\pi_t(\,\cdot \mid s_i^{(t)}) \,\big\|\, \bar{\pi}_{E,t}(\,\cdot \mid s_i^{(t)})\big),$$

Step size considerations

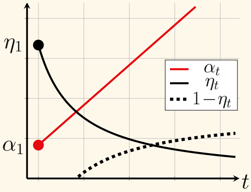
Two conditions of $\{\eta_t\}_{t=1}^{\infty}$ to guarantee convergence.

- Convergent sequence & divergent series:

  $\lim_{t\to\infty} \eta_t = 0$ and $\sum_{t=1}^{\infty} \eta_t = \infty$.

- Convergent series of squared terms:

  $\sum_{t=1}^{\infty} \eta_t = \infty$ and $\sum_{t=1}^{\infty} \eta_t^2 < \infty$.



A regret bound

In the optimal case of $\inf_{\pi\in\Pi} \mathbb{E}[f(\pi, \tau_t)] = 0$, the regret is bounded to $\mathcal{O}(1/T)$. When $\inf_{\pi\in\Pi} \mathbb{E}[f(\pi, \tau_t)] > 0$ when the step size satisfy conditions above. Thus, the regret is bounded to $\mathcal{O}(1/T)$ even for the general case.

Algorithm: MD-IRL on an adversarial framework

Dual discriminators: neural network parameters θ, ϕ, and ν are presented representing agent policy, reward, and expert policy functions.

- Matching overall state densities $D_\xi(s) = \sigma\big(d_\xi(s)\big)$.
- Imitating specific behavior $D_\nu(s, a; \theta, \xi) = \sigma\big(\log\{\pi_\nu(a|s)/\pi_\theta(a|s)\} + d_\xi(s)\big)$.

Define the objective of ϕ as direct interpretation of the update rule:

$$\mathcal{L}_{\psi_\phi} = \mathbb{E}_{s\sim\bar{\tau}_t}\big[\eta_t\, D_\Omega\big(\pi_\phi(\,\cdot\,|\,s)\big\|\pi_\nu(\,\cdot\,|\,s)\big) + (1-\eta_t)D_\Omega\big(\pi_\phi(\,\cdot\,|\,s)\big\|\pi_\theta(\,\cdot\,|\,s)\big)\big],$$

with adaptively adjusted step size coefficient η_t and a trajectory $\bar{\tau}_t$.

Define Mirror Descent Adversarial Inverse Reinforcement Learning (MD-AIRL) :

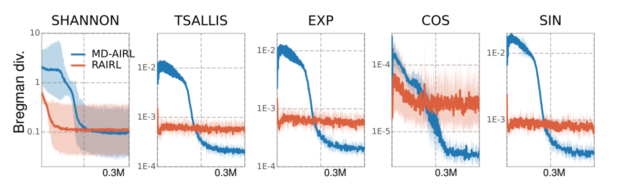
$$\psi_\phi^\lambda(s, a) = \lambda\,\psi_\phi(s, a) + d_\xi(s), \qquad \lambda \in \mathbb{R}^+$$

Train RL policy π_θ with ψ_ϕ^λ using the RAC algorithm[6]

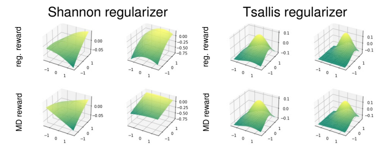
[6] Yang et al. (2019). A Regularized Approach to Sparse Optimal Policy in Reinforcement Learning. In NeurIPS.

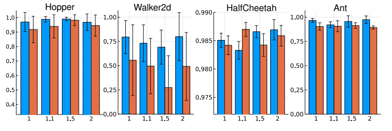
Experimental results: discrete action problems

	$\lvert\mathcal{A}\rvert = 10^2$		$\lvert\mathcal{A}\rvert = 10^3$		$\lvert\mathcal{A}\rvert = 10^4$	
Method	RAIRL	MD-AIRL	RAIRL	MD-AIRL	RAIRL	MD-AIRL
Shannon	2.55 ± 1.59	$\mathbf{2.28 \pm 1.20}$	140.3 ± 87.5	$\mathbf{125.3 \pm 61}$	-	-
Tsallis	0.21 ± 0.13	$\mathbf{0.11 \pm 0.04}$	0.55 ± 0.13	$\mathbf{0.24 \pm 0.03}$	4.95 ± 2.3	$\mathbf{4.21 \pm 0.2}$
exp	0.27 ± 0.17	$\mathbf{0.13 \pm 0.06}$	0.55 ± 0.12	$\mathbf{0.23 \pm 0.03}$	5.06 ± 2.4	$\mathbf{4.97 \pm 0.7}$
cos	0.05 ± 0.04	$\mathbf{0.02 \pm 0.01}$	0.03 ± 0.02	$\mathbf{0.01 \pm 0.01}$	0.21 ± 0.6	$\mathbf{0.05 \pm 0.1}$
sin	0.34 ± 0.25	$\mathbf{0.12 \pm 0.04}$	3.82 ± 3.46	$\mathbf{1.07 \pm 0.75}$	8.12 ± 3.8	$\mathbf{7.59 \pm 1.0}$



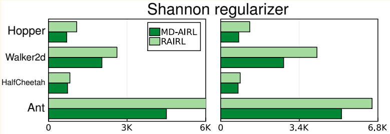
Experimental results: continuous action problems

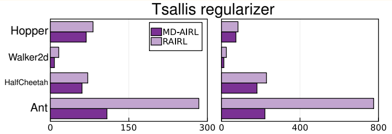




	Method	$\varepsilon = 0.01$	$\varepsilon = 0.5$
Hopper	RAIRL (Shannon)	3636.03 ± 391.09	3573.74 ± 508.14
	MD-AIRL (Shannon)	$\mathbf{3669.25 \pm 177.78}$	$\mathbf{3653.31 \pm 267.87}$
	RAIRL (Tsallis)	3671.12 ± 322.32	3576.17 ± 515.75
	MD-AIRL (Tsallis)	$\mathbf{3730.14 \pm 63.09}$	$\mathbf{3701.24 \pm 205.68}$
Walker2d	RAIRL (Shannon)	2856.56 ± 939.9	2451.00 ± 1392.6
	MD-AIRL (Shannon)	$\mathbf{3386.38 \pm 953.59}$	$\mathbf{3252.65 \pm 1395.7}$
	RAIRL (Tsallis)	2731.84 ± 1058.7	2435.10 ± 1555.2
	MD-AIRL (Tsallis)	$\mathbf{3624.00 \pm 992.63}$	$\mathbf{3093.54 \pm 963.96}$

	Method	$\varepsilon = 0.01$	$\varepsilon = 0.5$
HalfCheetah	RAIRL (Shannon)	4354.15 ± 63.83	4216.99 ± 661.17
	MD-AIRL (Shannon)	$\mathbf{4373.17 \pm 68.12}$	$\mathbf{4337.18 \pm 106.40}$
	RAIRL (Tsallis)	4364.13 ± 68.09	4216.67 ± 248.08
	MD-AIRL (Tsallis)	$\mathbf{4388.87 \pm 73.19}$	$\mathbf{4247.44 \pm 266.73}$
Ant	RAIRL (Shannon)	4493.74 ± 383.04	3777.78 ± 505.78
	MD-AIRL (Shannon)	$\mathbf{4658.29 \pm 201.37}$	$\mathbf{4284.38 \pm 329.79}$
	RAIRL (Tsallis)	4359.62 ± 168.46	3660.22 ± 508.54
	MD-AIRL (Tsallis)	$\mathbf{4705.25 \pm 130.53}$	$\mathbf{4127.37 \pm 457.25}$

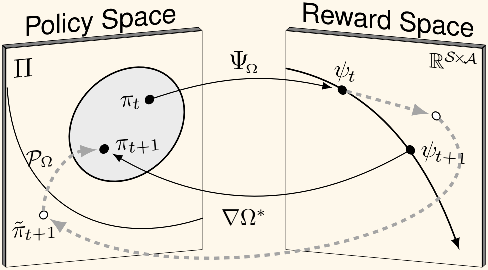




Thank you!

Robust Imitation via

Mirror Descent Inverse Reinforcement Learning



arXiv: `https://arxiv.org/abs/2210.11201`

BI lab: `https://bi.snu.ac.kr`

Dong-Sig Han: `https://dshan4585.github.io`

{dshan, hskim, hdlee, jhryu, btzhang}@bi.snu.ac.kr

BIOINTELLIGENCE