Towards Theoretical Understanding of Inverse Reinforcement Learning

Alberto Maria Metelli[1], Filippo Lazzati[1] and Marcello Restelli[1]
[1] Politecnico di Milano

40th International Conference on Machine Learning, Honolulu, HI.
July 2023

Research Question

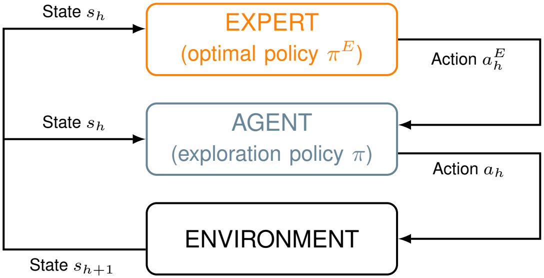
How many samples are needed to accurately solve the Inverse Reinforcement Learning (IRL) problem with high probability?

Research Question

How many samples are needed to accurately solve the Inverse Reinforcement Learning (IRL) problem with high probability?

Sample Complexity **Lower Bound** for IRL

Inverse Reinforcement Learning (IRL)

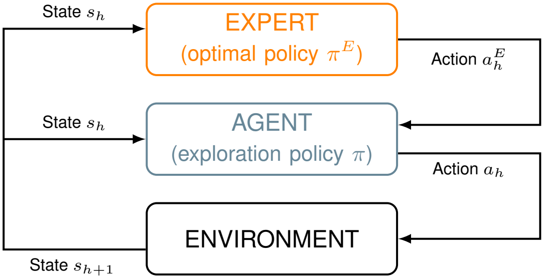


- At every stage $h \in [\![H]\!]$:
  - Observe state s_h
  - Observe **expert action** $a_h^E \sim \pi_h^E(\cdot|s_h)$
  - Play **exploratory action** $a_h \sim \pi_h(\cdot|s_h)$
  - Transition to next state $s_{h+1} \sim p_h(\cdot|s_h, a_h)$

- Traditional Goal of IRL (Arora and Doshi, 2021; Adams et al., 2022)

Find **one feasible reward function** r^* that makes the **expert's policy** π^E **optimal**, i.e.,

$$\pi^E \in \arg\max_{\pi} V^{\pi}(\cdot; r^*)$$



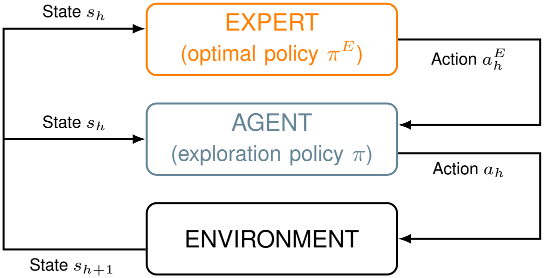
- At every stage $h \in [\![H]\!]$:
  - Observe state s_h
  - Observe **expert action** $a_h^E \sim \pi_h^E(\cdot|s_h)$
  - Play **exploratory action** $a_h \sim \pi_h(\cdot|s_h)$
  - Transition to next state $s_{h+1} \sim p_h(\cdot|s_h, a_h)$

- Traditional Goal of IRL (Arora and Doshi, 2021; Adams et al., 2022)

Find **one feasible reward function** r^* that makes the **expert's policy** π^E **optimal**, i.e.,

$$\pi^E \in \arg\max_{\pi} V^{\pi}(\cdot; r^*)$$

Inverse Reinforcement Learning (IRL)



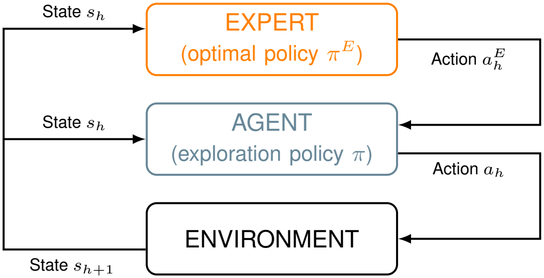
- At every stage $h \in [\![H]\!]$:
  - Observe state s_h
  - Observe **expert action** $a_h^E \sim \pi_h^E(\cdot|s_h)$
  - Play **exploratory action** $a_h \sim \pi_h(\cdot|s_h)$
  - Transition to next state $s_{h+1} \sim p_h(\cdot|s_h, a_h)$

- Traditional Goal of IRL (Arora and Doshi, 2021; Adams et al., 2022)

Find **one feasible reward function** r^* that makes the **expert's policy** π^E **optimal**, i.e.,

$$\pi^E \in \arg\max_{\pi} V^{\pi}(\cdot; r^*)$$

Inverse Reinforcement Learning (IRL)



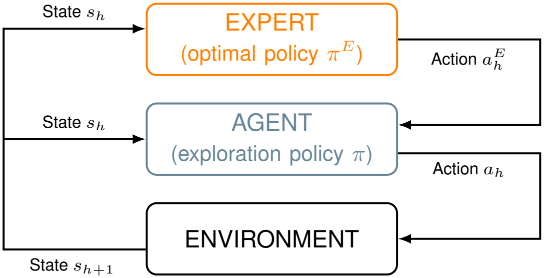
- At every stage $h \in [\![H]\!]$:
  - Observe state s_h
  - Observe **expert action** $a_h^E \sim \pi_h^E(\cdot|s_h)$
  - Play **exploratory action** $a_h \sim \pi_h(\cdot|s_h)$
  - Transition to next state $s_{h+1} \sim p_h(\cdot|s_h, a_h)$

- Traditional Goal of IRL (Arora and Doshi, 2021; Adams et al., 2022)

Find **one feasible reward function** r^* that makes the **expert's policy** π^E **optimal**, i.e.,

$$\pi^E \in \arg\max_{\pi} V^{\pi}(\cdot; r^*)$$

Inverse Reinforcement Learning (IRL)



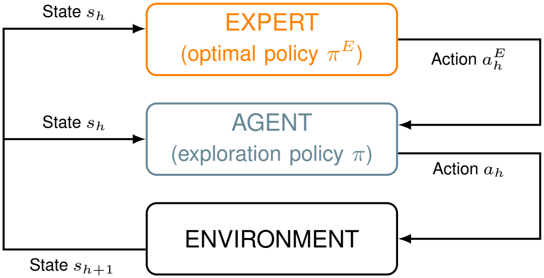
- At every stage $h \in [\![H]\!]$:
  - Observe state s_h
  - Observe **expert action** $a_h^E \sim \pi_h^E(\cdot|s_h)$
  - Play **exploratory action** $a_h \sim \pi_h(\cdot|s_h)$
  - Transition to next state $s_{h+1} \sim p_h(\cdot|s_h, a_h)$

- Traditional Goal of IRL (Arora and Doshi, 2021; Adams et al., 2022)

Find **one feasible reward function** r^* that makes the **expert's policy** π^E **optimal**, i.e.,

$$\pi^E \in \arg\max_{\pi} V^{\pi}(\cdot; r^*)$$

Inverse Reinforcement Learning (IRL)



- At every stage $h \in [\![H]\!]$:
  - Observe state s_h
  - Observe **expert action** $a_h^E \sim \pi_h^E(\cdot|s_h)$
  - Play **exploratory action** $a_h \sim \pi_h(\cdot|s_h)$
  - Transition to next state $s_{h+1} \sim p_h(\cdot|s_h, a_h)$

- Traditional Goal of IRL (Arora and Doshi, 2021; Adams et al., 2022)

Find **one feasible reward function** r^* that makes the **expert's policy** π^E **optimal**, i.e.,

$$\pi^E \in \arg\max_{\pi} V^{\pi}(\cdot; r^*)$$

Feasible Reward Set

- **Ambiguity** problem (Ng and Russell, 2000)
- Study the **full** set of feasible rewards r^* → **Feasible Reward Set** (Metelli et al., 2021)

Find **all feasible reward functions** $\mathcal{R}$ that make the **expert's policy** π^E **optimal**, i.e.,

$$\mathcal{R} := \left\{ \text{all rewards } r^* \; : \; \pi^E \in \arg\max_{\pi} V^{\pi}(\cdot; r^*) \right\}$$

- $\mathcal{R}$ defined through **linear constraints**
- If p and π^E are **known**, check if $\hat{r}$ is feasible takes $O(HS^2A)$

Feasible Reward Set

- **Ambiguity** problem (Ng and Russell, 2000)
- Study the **full** set of feasible rewards r^* → **Feasible Reward Set** (Metelli et al., 2021)

Find **all feasible reward functions** $\mathcal{R}$ that make the **expert's policy** π^E **optimal**, i.e.,

$$\mathcal{R} := \left\{ \text{all rewards } r^* \ : \ \pi^E \in \arg\max_{\pi} V^{\pi}(\cdot; r^*) \right\}$$

- $\mathcal{R}$ defined through **linear constraints**
- If p and π^E are **known**, check if $\hat{r}$ is feasible takes $O(HS^2A)$

Feasible Reward Set

- **Ambiguity** problem (Ng and Russell, 2000)
- Study the **full** set of feasible rewards r^* $\rightarrow$ **Feasible Reward Set** (Metelli et al., 2021)

Find **all feasible reward functions** $\mathcal{R}$ that make the **expert's policy** π^E **optimal**, i.e.,

$$\mathcal{R} := \left\{ \text{all rewards } r^* \; : \; \pi^E \in \arg\max_{\pi} V^{\pi}(\cdot; r^*) \right\}$$

- $\mathcal{R}$ defined through **linear constraints**
- If p and π^E are **known**, check if $\hat{r}$ is feasible takes $O(HS^2A)$

- **Ambiguity** problem (Ng and Russell, 2000)
- Study the **full** set of feasible rewards r^* $\rightarrow$ **Feasible Reward Set** (Metelli et al., 2021)

Find **all feasible reward functions** $\mathcal{R}$ that make the **expert's policy** π^E **optimal**, i.e.,

$$\mathcal{R} := \left\{ \text{all rewards } r^* \;:\; \pi^E \in \arg\max_{\pi} V^{\pi}(\cdot; r^*) \right\}$$

- $\mathcal{R}$ defined through **linear constraints**
- If p and π^E are **known**, check if $\widehat{r}$ is feasible takes $O(HS^2A)$

Feasible Reward Set

- **Ambiguity** problem (Ng and Russell, 2000)
- Study the **full** set of feasible rewards r^* $\rightarrow$ **Feasible Reward Set** (Metelli et al., 2021)

Find **all feasible reward functions** $\mathcal{R}$ that make the **expert's policy** π^E **optimal**, i.e.,

$$\mathcal{R} := \left\{ \text{all rewards } r^* \ : \ \pi^E \in \arg\max_{\pi} V^{\pi}(\cdot; r^*) \right\}$$

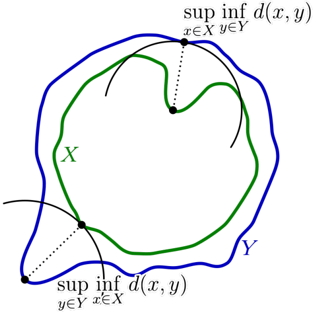
- $\mathcal{R}$ defined through **linear constraints**
- If p and π^E are **known**, check if $\widehat{r}$ is feasible takes $O(HS^2A)$

Estimating the Feasible Reward Set

- Transition model p and expert's policy π^E **unknown**
- Estimate $\widehat{p}$ and $\widehat{\pi}^E$ with **samples** inducing $\widehat{\mathcal{R}}$
- **Hausdorff distance** between $\mathcal{R}$ (true) and $\widehat{\mathcal{R}}$ (estimated) feasible reward sets

$$\mathcal{H}_d\left(\mathcal{R}, \widehat{\mathcal{R}}\right) = \max\left\{\sup_{r\in\mathcal{R}}\inf_{\widehat{r}\in\widehat{\mathcal{R}}} d(r,\widehat{r}), \sup_{\widehat{r}\in\widehat{\mathcal{R}}}\inf_{r\in\mathcal{R}} d(r,\widehat{r}),\right\}$$

where $\quad d(r,\widehat{r}) = \max_{s,a,h}\left|r_h(s,a) - \widehat{r}_h(s,a)\right|$

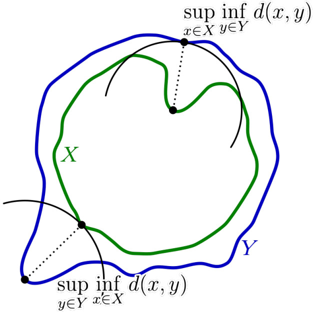


Estimating the Feasible Reward Set

- Transition model p and expert's policy π^E **unknown**
- Estimate $\widehat{p}$ and $\widehat{\pi}^E$ with **samples** inducing $\widehat{\mathcal{R}}$
- **Hausdorff distance** between $\mathcal{R}$ (true) and $\widehat{\mathcal{R}}$ (estimated) feasible reward sets

$$\mathcal{H}_d\left(\mathcal{R}, \widehat{\mathcal{R}}\right) = \max\left\{\sup_{r \in \mathcal{R}} \inf_{\widehat{r} \in \widehat{\mathcal{R}}} d(r, \widehat{r}), \sup_{\widehat{r} \in \widehat{\mathcal{R}}} \inf_{r \in \mathcal{R}} d(r, \widehat{r}), \right\}$$

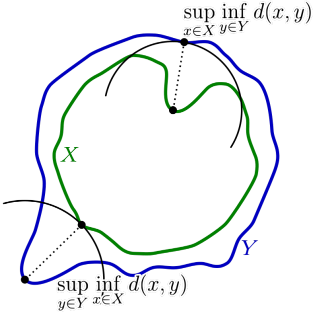
where $$d(r, \widehat{r}) = \max_{s,a,h} \left| r_h(s,a) - \widehat{r}_h(s,a) \right|$$



- Transition model p and expert's policy π^E **unknown**
- Estimate $\widehat{p}$ and $\widehat{\pi}^E$ with **samples** inducing $\widehat{\mathcal{R}}$
- **Hausdorff distance** between $\mathcal{R}$ (true) and $\widehat{\mathcal{R}}$ (estimated) feasible reward sets

$$\mathcal{H}_d\left(\mathcal{R},\widehat{\mathcal{R}}\right) = \max\left\{\sup_{r\in\mathcal{R}}\inf_{\widehat{r}\in\widehat{\mathcal{R}}} d(r,\widehat{r}), \sup_{\widehat{r}\in\widehat{\mathcal{R}}}\inf_{r\in\mathcal{R}} d(r,\widehat{r}),\right\}$$

where $$d(r,\widehat{r}) = \max_{s,a,h}|r_h(s,a) - \widehat{r}_h(s,a)|$$

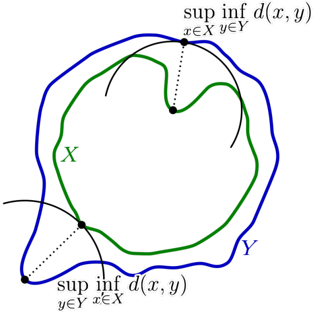


Estimating the Feasible Reward Set

- Transition model p and expert's policy π^E **unknown**
- Estimate $\widehat{p}$ and $\widehat{\pi}^E$ with **samples** inducing $\widehat{\mathcal{R}}$
- **Hausdorff distance** between $\mathcal{R}$ (true) and $\widehat{\mathcal{R}}$ (estimated) feasible reward sets

$$\mathcal{H}_d\left(\mathcal{R},\widehat{\mathcal{R}}\right) = \max\left\{\sup_{r\in\mathcal{R}}\inf_{\widehat{r}\in\widehat{\mathcal{R}}} d(r,\widehat{r}), \sup_{\widehat{r}\in\widehat{\mathcal{R}}}\inf_{r\in\mathcal{R}} d(r,\widehat{r}),\right\}$$

where $$d(r,\widehat{r}) = \max_{s,a,h}\left|r_h(s,a) - \widehat{r}_h(s,a)\right|$$



- An algorithm $\mathfrak{A}$ that outputs $\widehat{\mathcal{R}}_\tau$ after τ calls to the environment is (ϵ, δ)-PAC if

$$\mathbb{P}\left(\mathcal{H}_d\left(\mathcal{R}, \widehat{\mathcal{R}}_\tau\right) > \epsilon\right) \leqslant \delta$$

$\tau =$ sample complexity

- Sample complexity lower bound

$$\tau \geqslant \text{poly}\left(S, A, H, \frac{1}{\epsilon}, \log\left(\frac{1}{\delta}\right)\right)$$

- An algorithm $\mathfrak{A}$ that outputs $\widehat{\mathcal{R}}_\tau$ after τ calls to the environment is (ϵ, δ)-PAC if

$$\mathbb{P}\left(\mathcal{H}_d\left(\mathcal{R}, \widehat{\mathcal{R}}_\tau\right) > \epsilon\right) \leqslant \delta$$

$$\tau = \text{sample complexity}$$

- **Sample complexity lower bound**

$$\tau \geqslant \text{poly}\left(S, A, H, \frac{1}{\epsilon}, \log\left(\frac{1}{\delta}\right)\right)$$

Main Result: Lower Bound

Theorem

For any (ϵ, δ)-PAC algorithm $\mathfrak{A}$, with ϵ and δ sufficiently small, there exists an IRL problem, with S, A and H sufficiently large, such that the **expected sample complexity** *is lower bounded by:*

- *if the transition model p is time-inhomogeneous (i.e., $p_h \neq p_{h+1}$):*

$$\mathbb{E}[\tau] \geqslant \Omega\left(\frac{H^3 SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)\right);$$

- *if the transition model p is time-homogeneous (i.e., $p_h = p_{h+1}$):*

$$\mathbb{E}[\tau] \geqslant \Omega\left(\frac{H^2 SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)\right).$$

Main Result: Lower Bound

Theorem

For any (ϵ, δ)-PAC algorithm $\mathfrak{A}$, with ϵ and δ sufficiently small, there exists an IRL problem, with S, A and H sufficiently large, such that the **expected sample complexity** *is lower bounded by:*

- *if the transition model p is time-inhomogeneous (i.e., $p_h \neq p_{h+1}$):*

$$\mathbb{E}[\tau] \geqslant \Omega\left(\frac{H^3 SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)\right);$$

- *if the transition model p is time-homogeneous (i.e., $p_h = p_{h+1}$):*

$$\mathbb{E}[\tau] \geqslant \Omega\left(\frac{H^2 SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)\right).$$

Main Result: Lower Bound

Theorem

For any (ϵ, δ)-PAC algorithm $\mathfrak{A}$, with ϵ and δ sufficiently small, there exists an IRL problem, with S, A and H sufficiently large, such that the **expected sample complexity** *is lower bounded by:*

- *if the transition model p is time-inhomogeneous (i.e., $p_h \neq p_{h+1}$):*

$$\mathbb{E}[\tau] \geqslant \Omega\left(\frac{H^3 SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)\right);$$

- *if the transition model p is time-homogeneous (i.e., $p_h = p_{h+1}$):*

$$\mathbb{E}[\tau] \geqslant \Omega\left(\frac{H^2 SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)\right).$$

Main Result: Lower Bound

Comparison when p time-homogeneous

- Inverse Reinforcement Learning (IRL, **ours**)

$$\Omega\left(\frac{H^2SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)\right)$$

- Reinforcement Learning (RL, Domingues et al., 2021) (**tight**)

$$\Omega\left(\frac{H^2SA}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$$

IRL harder than RL

- Reward-Free Exploration (RFE, Jin et al., 2020; Dann and Brunskill, 2015) (**tight**)

$$\Omega\left(\frac{H^2SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)\right)$$

IRL as hard as RFE

Main Result: Lower Bound

Comparison when p time-homogeneous

- Inverse Reinforcement Learning (IRL, **ours**)

$$\Omega\left(\frac{H^2SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)\right)$$

- Reinforcement Learning (RL, Domingues et al., 2021) (**tight**)

$$\Omega\left(\frac{H^2SA}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$$

IRL harder than RL

- Reward-Free Exploration (RFE, Jin et al., 2020; Dann and Brunskill, 2015) (**tight**)

$$\Omega\left(\frac{H^2SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)\right)$$

IRL as hard as RFE

Main Result: Lower Bound

Comparison when p time-homogeneous

- Inverse Reinforcement Learning (IRL, **ours**)

$$\Omega\left(\frac{H^2SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)\right)$$

- Reinforcement Learning (RL, Domingues et al., 2021) (**tight**)

$$\Omega\left(\frac{H^2SA}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$$

IRL harder than RL

- Reward-Free Exploration (RFE, Jin et al., 2020; Dann and Brunskill, 2015) (**tight**)

$$\Omega\left(\frac{H^2SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)\right)$$

IRL as hard as RFE

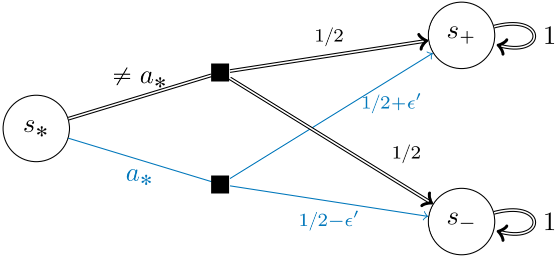
Proof Sketch

Two **regimes** of δ

Two **regimes** of $\delta \rightarrow$ **Small-δ regime**

- Expert's policy $\pi^E(s) = a_0$
- **Hard** to identify which action behaves like a_*
- Construct $\Theta(A)$ hard instances
- **Technical tool**: Bretagnolle-Huber inequality (Lattimore and Szepesvári, 2020)

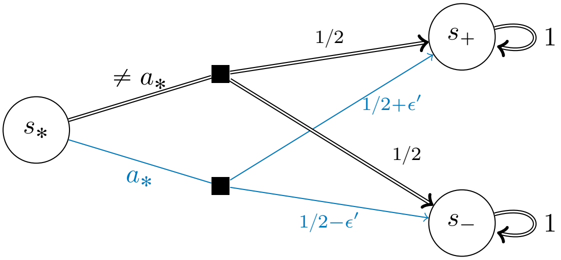
$$\Omega\left(\frac{AH^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$$



Two **regimes** of $\delta \rightarrow$ **Small-δ regime**

- Expert's policy $\pi^E(s) = a_0$
- **Hard** to identify which action behaves like a_*
- Construct $\Theta(A)$ hard instances
- **Technical tool**: Bretagnolle-Huber inequality (Lattimore and Szepesvári, 2020)

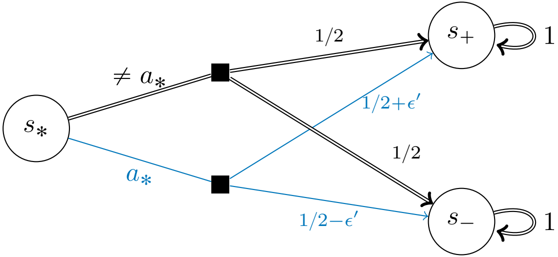
$$\Omega\left(\frac{AH^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$$



Two **regimes** of $\delta \to$ **Small-δ regime**

- Expert's policy $\pi^E(s) = a_0$
- **Hard** to identify which action behaves like a_*
- Construct $\Theta(A)$ hard instances
- **Technical tool**: Bretagnolle-Huber inequality (Lattimore and Szepesvári, 2020)

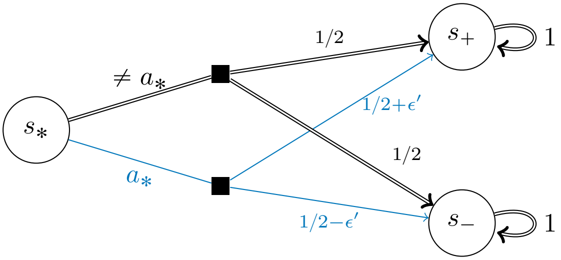
$$\Omega\left(\frac{AH^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$$



Two regimes of $\delta \to$ **Small-δ regime**

- Expert's policy $\pi^E(s) = a_0$
- **Hard** to identify which action behaves like a_*
- Construct $\Theta(A)$ hard instances
- **Technical tool**: Bretagnolle-Huber inequality (Lattimore and Szepesvári, 2020)

$$\Omega\left(\frac{AH^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$$



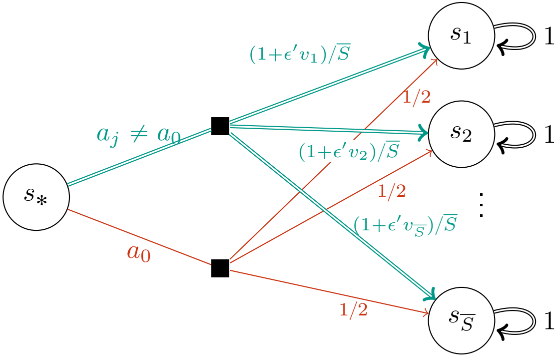
Proof Sketch

Two **regimes** of δ

Two **regimes** of δ → **Large-δ regime**

- Expert's policy $\pi^E(s) = a_0$
- **Hard** to distinguish all actions a_j
- Construct $\Theta(2^S)$ hard instances via a **packing** argument based on **Hamming coding**
- **Technical tool**: Fano's inequality (Gerchinovitz et al., 2020)

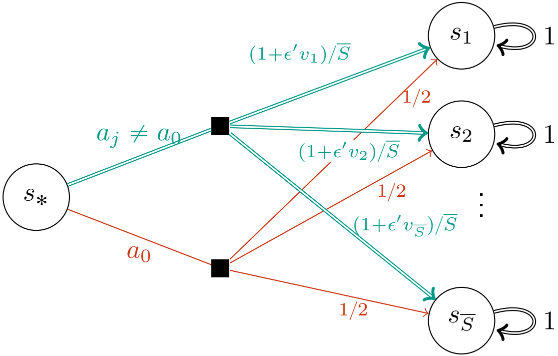
$$\Omega\left(\frac{SAH^2}{\epsilon^2}\right)$$



Two **regimes** of $\delta \to$ **Large-δ regime**

- Expert's policy $\pi^E(s) = a_0$
- **Hard** to distinguish all actions a_j
- Construct $\Theta(2^S)$ hard instances via a **packing** argument based on **Hamming coding**
- **Technical tool**: Fano's inequality (Gerchinovitz et al., 2020)

$$\Omega\left(\frac{SAH^2}{\epsilon^2}\right)$$

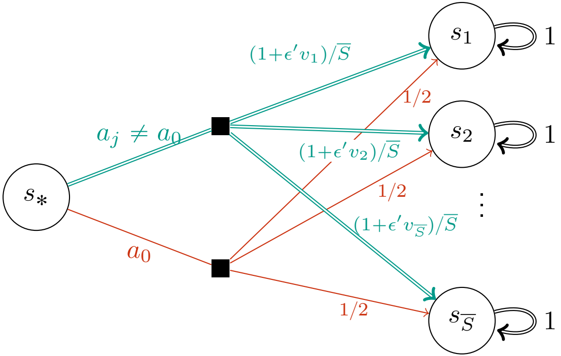


Proof Sketch

Two **regimes** of $\delta \to$ **Large-δ regime**

- Expert's policy $\pi^E(s) = a_0$
- **Hard** to distinguish all actions a_j
- Construct $\Theta(2^S)$ hard instances via a **packing** argument based on **Hamming coding**
- **Technical tool**: Fano's inequality (Gerchinovitz et al., 2020)

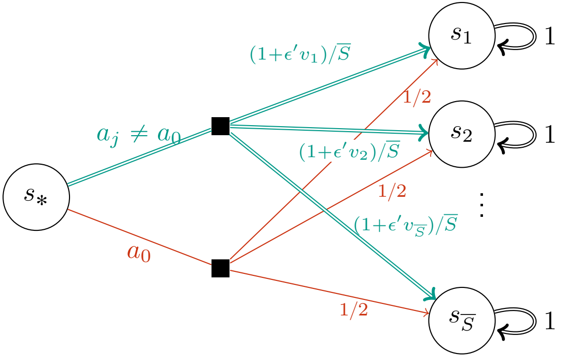
$$\Omega\left(\frac{SAH^2}{\epsilon^2}\right)$$



Two **regimes** of $\delta \rightarrow$ **Large-δ regime**

- Expert's policy $\pi^E(s) = a_0$
- **Hard** to distinguish all actions a_j
- Construct $\Theta(2^S)$ hard instances via a **packing** argument based on **Hamming coding**
- **Technical tool**: Fano's inequality (Gerchinovitz et al., 2020)

$$\Omega\left(\frac{SAH^2}{\epsilon^2}\right)$$



What it Missing in this Talk?

- **Lipschitz** properties of the feasible reward set $\mathcal{R}$
  - **Impossibility** results for learning strict subsets of $\mathcal{R}$ (e.g., state-only rewards)
- **Uniform sampling** algorithm nearly matches the lower bound
- Relation with different **dissimilarity index** between rewards

What it Missing in this Talk?

- **Lipschitz** properties of the feasible reward set $\mathcal{R}$
  - **Impossibility** results for learning strict subsets of $\mathcal{R}$ (e.g., state-only rewards)
- **Uniform sampling** algorithm nearly matches the lower bound
- Relation with different **dissimilarity index** between rewards

What it Missing in this Talk?

- **Lipschitz** properties of the feasible reward set $\mathcal{R}$
  - **Impossibility** results for learning strict subsets of $\mathcal{R}$ (e.g., state-only rewards)
- **Uniform sampling** algorithm nearly matches the lower bound
- Relation with different **dissimilarity index** between rewards

- **Lipschitz** properties of the feasible reward set $\mathcal{R}$
  - **Impossibility** results for learning strict subsets of $\mathcal{R}$ (e.g., state-only rewards)
- **Uniform sampling** algorithm nearly matches the lower bound
- Relation with different **dissimilarity index** between rewards

Thank You for Your Attention!

Contacts: albertomaria.metelli@polimi.it
Link: https://icml.cc/virtual/2023/poster/24193

Adams, S. C., Cody, T., and Beling, P. A. (2022). A survey of inverse reinforcement learning. *Artif. Intell. Rev.*, 55(6):4307–4346.

Arora, S. and Doshi, P. (2021). A survey of inverse reinforcement learning: Challenges, methods and progress. *Artif. Intell.*, 297:103500.

Dann, C. and Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 2818–2826.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2021). Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory (ALT)*, volume 132 of *Proceedings of Machine Learning Research*, pages 578–598. PMLR.

Gerchinovitz, S., Ménard, P., and Stoltz, G. (2020). Fano's inequality for random variables. *Statistical Science*, 35(2):178–201.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. (2020). Reward-free exploration for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 4870–4879. PMLR.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Metelli, A. M., Ramponi, G., Concetti, A., and Restelli, M. (2021). Provably efficient learning of transferable rewards. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 7665–7676. PMLR.

Ng, A. Y. and Russell, S. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 663–670. Morgan Kaufmann.