Balancing Sample Efficiency and Suboptimality in Inverse Reinforcement Learning

A.Damiani, Giorgio Manganini, A.Metelli, M.Restelli

Gran Sasso Science Institute (GSSI), L'Aquila, Italy
giorgio.manganini@gssi.it

July 2022
Thirty-ninth International Conference on Machine Learning

Inverse Reinforcement Learning (IRL)

- IRL [1] is the process of **recovering**, from (demonstrations of) an expert's policy, the **expert's reward** function

  π_E expert's policy

  r_E, γ_E expert's reward and discount factor

- The learned reward is intended to be successively used in **forward Reinforcement Learning** [2]

  M finite-sample budget for the forward RL phase

  $\widehat{Q}^\star_M$ approximation of optimal $Q^\star_{r,\gamma}$, under a pair (γ, r)

[1] [Ng and Russell, 2000]
[2] [RL, Sutton and Barto, 2018]

G S
S I

Balancing Sample Efficiency and Suboptimality

IRL

A reward r is **compatible** [a] with with the expert's policy π_E if

$$\pi \in \mathcal{G}\left[Q^{\star}_{r,\gamma}\right]$$

[a] [Ng and Russell, 2000]

Sample Complexity

- *How much data must we collect in order to achieve "learning"?* [a]
- **Number of samples** required to attain a near-optimal estimate of the **optimal value-function**

$$\sim \frac{1}{1-\gamma}$$ [b]

[a] [Kakade, 2003]

[b] e.g.,[Munos and Szepesvári, 2008, Farahmand et al., 2010, Lazaric et al., 2012, Azar et al., 2013]

Novel IRL Formulation for Efficient Forward Learning

$$\min_{r\in\mathcal{R},\gamma\in[0,1)} \max_{\pi\in\mathcal{G}[\widehat{Q}^\star_M]} \left\| Q^{\pi_E}_{r_E,\gamma_E} - Q^{\pi}_{r_E,\gamma_E} \right\|$$

$$\text{s.t. } \left\| \widehat{Q}^\star_M - Q^\star_{r,\gamma} \right\| \leq \epsilon^\star(M,\gamma)$$

Reward r compatibility with expert's π_E

- Worst-case distance between expert's π_E and the learned policy π under optimized r in the successive forward RL task

Sample complexity of forward RL phase

- Tuned by directly optimizing γ

Forward RL phase with finite samples M

- Confidence region of the future estimated optimal Q-function $\widehat{Q}^\star_M$ under the optimized reward and discount (r,γ)

Novel IRL Formulation for Efficient Forward Learning

$$\min_{r \in \mathcal{R}, \gamma \in [0,1)} \max_{\pi \in \mathcal{G}\left[\widehat{Q}^\star_M\right]} \left\| Q^{\pi_E}_{r_E,\gamma_E} - Q^{\pi}_{r_E,\gamma_E} \right\|$$

$$\text{s.t. } \left\| \widehat{Q}^\star_M - Q^\star_{r,\gamma} \right\| \leq \epsilon^\star(M,\gamma)$$

Reward r compatibility with expert's π_E

- Worst-case distance between expert's π_E and the learned policy π under optimized r in the successive forward RL task

Sample complexity of forward RL phase

- Tuned by directly optimizing γ

Forward RL phase with finite samples M

- Confidence region of the future estimated optimal Q-function $\widehat{Q}^\star_M$ under the optimized reward and discount (r, γ)

G S S I

Novel IRL Formulation for Efficient Forward Learning

$$\min_{r \in \mathcal{R},\ \gamma \in [0,1)} \max_{\pi \in \mathcal{G}[\widehat{Q}^\star_M]} \left\| Q^{\pi_E}_{r_E, \gamma_E} - Q^{\pi}_{r_E, \gamma_E} \right\|$$

$$\text{s.t.} \quad \left\| \widehat{Q}^\star_M - Q^\star_{r,\gamma} \right\| \leq \epsilon^\star(M, \gamma)$$

Reward r compatibility with expert's π_E

- Worst-case distance between expert's π_E and the learned policy π under optimized r in the successive forward RL task

Sample complexity of forward RL phase

- Tuned by directly optimizing γ

Forward RL phase with finite samples M

- Confidence region of the future estimated optimal Q-function $\widehat{Q}^\star_M$ under the optimized reward and discount (r, γ)

GS SI

Novel IRL Formulation for Efficient Forward Learning

$$\min_{r\in\mathcal{R},\gamma\in[0,1)} \max_{\pi\in\mathcal{G}[\widehat{Q}^\star_M]} \left\| Q^{\pi_E}_{r_E,\gamma_E} - Q^{\pi}_{r_E,\gamma_E} \right\|$$

$$\text{s.t.} \quad \left\| \widehat{Q}^\star_M - Q^\star_{r,\gamma} \right\| \leq \epsilon^\star(M,\gamma)$$

Reward r compatibility with expert's π_E

- Worst-case distance between expert's π_E and the learned policy π under optimized r in the successive forward RL task

Sample complexity of forward RL phase

- Tuned by directly optimizing γ

Forward RL phase with finite samples M

- Confidence region of the future estimated optimal Q-function $\widehat{Q}^\star_M$ under the optimized reward and discount (r,γ)

GSSI

Novel IRL Formulation for Efficient Forward Learning

$$\min_{r \in \mathcal{R},\, \gamma \in [0,1)} \max_{\pi \in \mathcal{G}\left[\widehat{Q}^\star_M\right]} \left\| Q^{\pi_E}_{r_E,\gamma_E} - Q^{\pi}_{r_E,\gamma_E} \right\|$$

$$\text{s.t.} \quad \left\| \widehat{Q}^\star_M - Q^\star_{r,\gamma} \right\| \leq \epsilon^\star(M,\gamma)$$

Reward r compatibility with expert's π_E

- Worst-case distance between expert's π_E and the learned policy π under optimized r in the successive forward RL task

Sample complexity of forward RL phase

- Tuned by directly optimizing γ

Forward RL phase with finite samples M

- Confidence region of the future estimated optimal Q-function $\widehat{Q}^\star_M$ under the optimized reward and discount (r,γ)

Objective function

- ✗ Exper's reward r_E and discount γ_E are **unknow**
- ✓ **Surrogate objective function**
  - from value-function distance to **policy divergence** (Theorem 4.1)
- ✓ Computable from an **offline dataset** available at IRL time

$$\left\| Q^{\pi_E}_{\boldsymbol{r_E},\boldsymbol{\gamma_E}} - Q^{\pi}_{\boldsymbol{r_E},\boldsymbol{\gamma_E}} \right\|$$

$\downarrow$

Theorem 4.1

$\downarrow$

$$\int_{\mathcal{S}} W_2(\pi_E(\cdot|s), \pi(\cdot|s))\, \mathrm{d}s$$

Dealing with forward Q-function $Q^\star_{r,\gamma}$

- ✗ Forward optimal Q-function $Q^\star_{r,\gamma}$ with the optimized pair (r,γ) is **unknown**
- ✗ Might be estimated with an inner loop of **forward RL**
- ✓ We replace it with $Q^{\pi_E}_{r,\gamma}$, since when (r,γ) are **compatible with the expert**, $Q^\star_{r,\gamma} = Q^{\pi_E}_{r,\gamma}$ holds

$$\left\| \widehat{Q}^\star_M - \boldsymbol{Q^\star_{r,\gamma}} \right\| \le \epsilon^\star(M,\gamma)$$

$$\downarrow$$

$$\left\| \widehat{Q}^{\pi_E}_M - Q^{\pi_E}_{r,\gamma} \right\| \le \epsilon_1(M,\gamma)$$

G S
S I

Relaxing the greedy constraint

- ✗ Computation of greedy policy is **complicated** within maximization
- ✓ We perform two **relaxations**
  - transition from a greedy policy to all policy with at least a **performance improvement**
  - we enforce the constraint over a **finite subset of states** $\mathcal{D}_{\text{IRL}} \subseteq \mathcal{S}$

$$\pi \in \mathcal{G}\left[\widehat{Q}_M^{\pi_E}\right]$$

$$\downarrow$$

$$\widehat{Q}_M^{\pi_E}(s, \pi(s)) \geq \widehat{Q}_M^{\pi_E}(s, \pi_E(s)) \quad \forall s \in \mathcal{S}$$

$$\downarrow$$

$$\sum_{s \in \mathcal{D}_{\text{IRL}}} \widehat{Q}_M^{\pi_E}(s, \pi(s)) - \widehat{Q}_M^{\pi_E}(s, \pi_E(s)) \geq 0$$

G S
S I

Enforcing the confidence region

$$\left\| \widehat{Q}_M^{\pi_E} - Q_{r,\gamma}^{\pi_E} \right\| \leq \epsilon_1(M, \gamma)$$

$$\downarrow$$

Proposition 4.3

$$\downarrow$$

$$\sum_{s \in \mathcal{D}_{\text{IRL}}} \widehat{Q}_N^{\pi_E}(s, \pi(s)) - \widehat{Q}_N^{\pi_E}(s, \pi_E(s)) + 2\epsilon_1(M, \gamma) + 2\epsilon_2(N, \gamma) \geq 0$$

- ✖ The confidence region on the **forward** $\widehat{Q}_M^{\pi_E}$ depends on the **expert's Q-function** $Q_{r,\gamma}^{\pi_E}$
- ✔ Compute a **looser constraint** by introducing the expert's Q-function approximation known at IRL time $Q_N^{\pi_E}$

GS
SI

The solvable IRL formulation

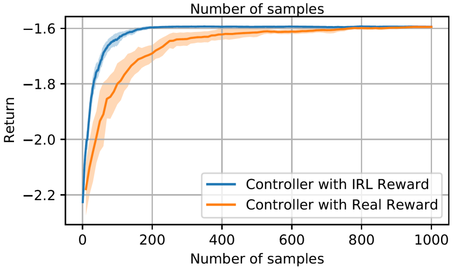
$$\min_{\substack{\boldsymbol{\theta}\in\mathbb{R}^{d_\theta}\\ \gamma\in[0,1)}} \max_{\boldsymbol{\eta}\in\mathbb{R}^{d_\eta}} \sum_{s\in\mathcal{D}_{\text{IRL}}} W_2\big(\pi^E(s), \pi_{\boldsymbol{\eta}}(s)\big)$$

$$\sum_{s\in\mathcal{D}_{\text{IRL}}} \widehat{Q}_N^{\pi_E}(s, \pi_{\boldsymbol{\eta}}(s)) - \widehat{Q}_N^{\pi_E}(s, \pi_E(s)) + 2\epsilon_M + 2\epsilon_N \geq 0$$

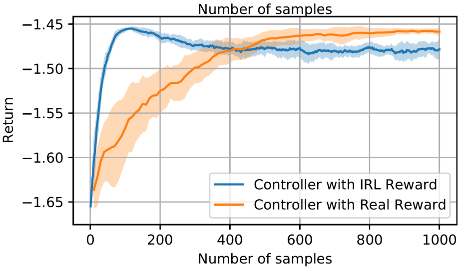
- We **parametrize**
  - $r_{\boldsymbol{\theta}}(s,a) = \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbb{R}^{d_\theta}$
  - $\pi_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathbb{R}^{d_\eta}$
- $\widehat{Q}_N^{\pi_E}$ is estimated by **policy evaluation** (e.g., LSTDQ [3])
- **Min-max optimization** is solved following the potential function approach, and minimizing it via gradient descent [4]

[3] [Lagoudakis and Parr, 2003]
[4] [Razaviyayn et al., 2020]

LQ [5]: forward learning results



(a) Expert's environment



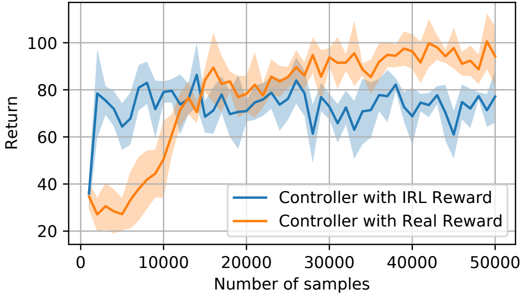
(b) Modified environment

(a) IRL and expert's rewards share the **same optimality**, but IRL optimal pair $(r_{\boldsymbol{\theta}}, \gamma)$ is more **sample efficient** (i.e., $\gamma < \gamma_E$)

(b) IRL reward peforms a (tunable) **trade-off** between the **bias** and the **sample efficiency** of the optimized pair $(r_{\boldsymbol{\theta}}, \gamma)$

[5] [Dorato et al., 1994]

Mountain Car [6]: forward learning results



- Expert's reward leads to **optimal** policy, but requires large γ
- IRL reward leads to a **sub-optimal** policy but admits a smaller γ, preferred for small values of M

GS SI

[6] [Moore, 1990]

Novel IRL formulation in a nutshell

- **Trade-off** between
  - error introduced on the learned policy when potentially choosing a **sub-optimal reward**
  - **sample efficiency** in the subsequent forward RL phase
- Completely **model-free**
- **No interaction** with the environment
- **No planning or forward RL** problem to be solved

GS
SI

Balancing Sample Efficiency and Suboptimality in Inverse Reinforcement Learning

A.Damiani, Giorgio Manganini, A.Metelli, M.Restelli

Gran Sasso Science Institute (GSSI), L'Aquila, Italy
giorgio.manganini@gssi.it

July 2022
Thirty-ninth International Conference on Machine Learning

References I

M. G. Azar, R. Munos, and H. J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013. doi: 10.1007/s10994-013-5368-1.

P. Dorato, V. Cerone, and C. Abdallah. *Linear-quadratic control: an introduction*. Simon & Schuster, Inc., 1994.

A. M. Farahmand, R. Munos, and C. Szepesvári. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pages 568–576, 2010.

S. M. Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, UCL (University College London), 2003.

M. G. Lagoudakis and R. Parr. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4:1107–1149, 2003.

A. Lazaric, M. Ghavamzadeh, and R. Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13:3041–3074, 2012.

A. W. Moore. Efficient memory-based learning for robot control. Technical report, University of Cambridge, 1990.

R. Munos and C. Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.

A. Y. Ng and S. J. Russell. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 663–670. Morgan Kaufmann Publishers Inc., 2000.

M. Razaviyayn, T. Huang, S. Lu, M. Nouiehed, M. Sanjabi, and M. Hong. Non-convex min-max optimization: Applications, challenges, and recent theoretical advances. *arXiv:2006.08141*, Aug 2020. arXiv: 2006.08141.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.