

Theory and Methods for Reinforcement Learning

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 6: Imitation Learning

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-618 (Spring 2023)



Google AI



FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION



License Information for Theory and Methods for Reinforcement Learning (EE-618)

- ▷ This work is released under a [Creative Commons License](#) with the following terms:
- ▷ **Attribution**
 - ▶ The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.
- ▷ **Non-Commercial**
 - ▶ The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor's permission.
- ▷ **Share Alike**
 - ▶ The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor's work.
- ▷ [Full Text of the License](#)

Overview

Introduction

Behavioral Cloning

Online Imitation Learning

Inverse Reinforcement Learning

Generative Adversarial Imitation Learning

Non-adversarial imitation learning

Convex-analytic methods

Learning from Demonstration (LfD)

- Motivation:**
- In RL, the reward function is known and we maximize the cumulative reward.
 - The reward functions are often manually designed to define the task.
 - Can we instead learn a policy by capitalizing an expert's behavior?

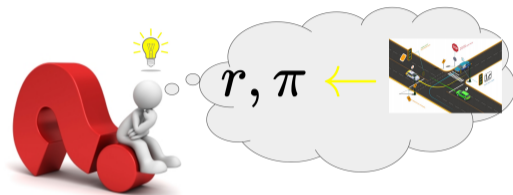


(a)



(b)

Learning from demonstration (LfD) (cont'd)



Real world problems:

- The reward function is unknown or is difficult to be designed.
- It is easier/more natural to use “demonstrations” by experts.

Imitation learning (IL) vs inverse reinforcement learning (IRL)

- Setting:

- ▶ Given an expert's demonstrations $\{(s_i, \pi_E(s_i))\}$ (offline trajectories or online queries)
- ▶ Reward signal is unobserved
- ▶ Transition model may be known or unknown

- Goals and approaches:

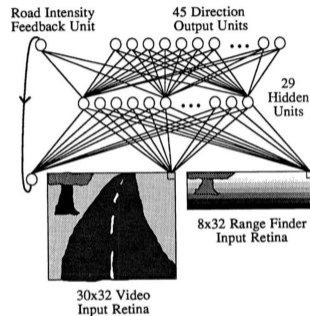
- ▶ Recover the expert's policy π_E directly: **imitation learning (IL)**
- ▶ Recover the expert's latent reward function $r_{\text{true}}(s, a)$: **inverse reinforcement learning (IRL)**

A historic application

- o Inverse Reinforcement Learning has been formally introduced by [25].



(c)



(d)

Figure: One of the first imitation learning systems using neural networks.

- o ALVINN: Autonomous Land Vehicle In a Neural Network, 1989 [28].

<https://www.youtube.com/watch?v=2KMAAmkz9go&t=205s>.

One of the latest applications

- Large language models: ChatGPT



<https://www.forbes.com/sites/bernardmarr/2022/12/28/what-does-chatgpt-really-mean-for-businesses/?sh=27bc344f7d1e>

- The last training step is based on Reinforcement Learning from Human Feedback (RLHF) (see [26]).
- A recent work [37] shows a close connection between IRL and RLHF.

More applications

- Simulated highway driving [2]
- Helicopter acrobatics [1]
- Urban navigation [38]
- Human goal inference [22]
- Object manipulation [34, 12]



(a)



(b)

Figure: Helicopter model and instance of its acrobatics [10].

Big Picture: Taxonomy of learning from demonstration methods

Method	Reward learning	Access to environment	Interactive demonstrations	Pre-collected demonstrations
Behavioural Cloning	NO	NO	NO	YES
Online IL	NO	YES	YES	MAYBE
Inverse RL	YES	YES	NO	YES
Adversarial IL	MAYBE	YES	NO	YES
Non-adversarial IL	MAYBE	YES	NO	YES

Remarks:

- BC avoids interaction with the environment, but can suffer from cascading errors.
- Online IL helps with the cascading errors but requires (expensive) expert queries.
- IRL explains the expert's behavior but has poor sample complexity and scalability.
- Adversarial IL avoids solving the RL problem repeatedly but are unstable due to adversarial training.
- Non-adversarial IL enjoys stable performance but is hampered by limited theoretical understanding.

Overview

Introduction

Behavioral Cloning

Online Imitation Learning

Inverse Reinforcement Learning

Generative Adversarial Imitation Learning

Non-adversarial imitation learning

Convex-analytic methods

Offline imitation learning: Behavioral cloning

- We assume there is an expert that has the optimal policy π_E .
- Input: offline data from expert's demonstration $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^n$, where $a_i \sim \pi_E(s_i)$.

Offline imitation learning: Behavioral cloning

- We assume there is an expert that has the optimal policy π_E .
- Input: offline data from expert's demonstration $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^n$, where $a_i \sim \pi_E(s_i)$.
- **Idea:** Directly learn the expert's policy via supervised learning.

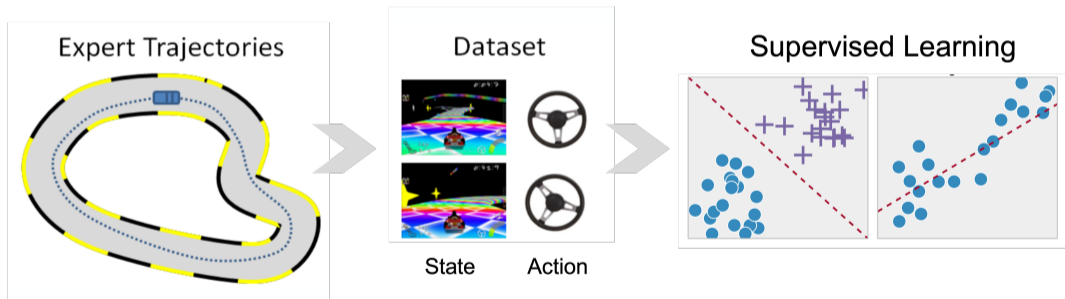


Figure: Source: <https://smartlabai.medium.com/a-brief-overview-of-imitation-learning-8a8a75c44a9c>

Behavioral cloning

Maximum Likelihood Estimation (MLE)

$$\max_{\pi \in \Pi} \sum_{(s,a) \in \mathcal{D}} \log \pi(a|s). \quad (1)$$

Behavioral cloning

Maximum Likelihood Estimation (MLE)

$$\max_{\pi \in \Pi} \sum_{(s,a) \in \mathcal{D}} \log \pi(a|s). \quad (1)$$

Risk Minimization

$$\min_{\theta} \mathbb{E}_{s \sim \lambda^{\pi_E}(\cdot|s)} \left[\ell(\pi_{\theta}(\cdot|s), \pi_E(\cdot|s)) \right], \quad (2)$$

where λ^{π_E} is the state visitation distribution under policy π_E and ℓ is a loss function.

Theoretical Guarantee

Theorem (Behavior Cloning) [4]

Let Π be a discrete and realizable policy class, i.e., $\pi_E \in \Pi$. With probability at least $1 - \delta$, behavioral cloning returns a policy that obeys:

$$V^{\pi_E} - V^{\hat{\pi}_{\text{MLE}}} \leq \mathcal{O} \left(\frac{1}{(1 - \gamma)^2} \sqrt{\frac{\log(|\Pi|/\delta)}{|\mathcal{D}|}} \right),$$

where $|\Pi|$ is the size of the policy class, and $|\mathcal{D}|$ is the length of the provided dataset.

Theoretical Guarantee

Theorem (Behavior Cloning) [4]

Let Π be a discrete and realizable policy class, i.e., $\pi_E \in \Pi$. With probability at least $1 - \delta$, behavioral cloning returns a policy that obeys:

$$V^{\pi_E} - V^{\hat{\pi}_{\text{MLE}}} \leq \mathcal{O} \left(\frac{1}{(1-\gamma)^2} \sqrt{\frac{\log(|\Pi|/\delta)}{|\mathcal{D}|}} \right),$$

where $|\Pi|$ is the size of the policy class, and $|\mathcal{D}|$ is the length of the provided dataset.

Remarks:

- BC only ensures the learned policy $\hat{\pi}_{\text{MLE}}$ is close to π_E under the support of distribution λ^{π_E} .
- The term $\sqrt{\frac{\log(|\Pi|/\delta)}{|\mathcal{D}|}}$ reflects the error $\hat{\pi}_{\text{MLE}}$ and π_E under the distribution λ^{π_E} .
- The term $\frac{1}{(1-\gamma)^2}$ reflects the cascading errors when performing under policy $\hat{\pi}_{\text{MLE}}$.
- The quadratic dependency on the effect horizon $H = \frac{1}{1-\gamma}$ is not avoidable in worst case [32].
- Can be improved if knowing the transition model [4].

Proof Sketch

- Recall the advantage defined as $A^{\hat{\pi}}(s, a) = Q^{\hat{\pi}}(s, a) - V^{\hat{\pi}}(s)$ and notice that $\mathbb{E}_{a \sim \hat{\pi}(\cdot|s)} A^{\hat{\pi}}(s, a) = 0, \quad \forall s.$
- We will use also that $A^{\hat{\pi}}(s, a) \leq \frac{1}{1-\gamma}$ if $\max_{s,a} |r(s, a)| \leq 1.$

Proof.

- ▶ Based on performance difference lemma [20], we have

$$\begin{aligned} V^{\pi_E} - V^{\hat{\pi}} &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \lambda^{\pi_E}, a \sim \pi_E(\cdot|s)} A^{\hat{\pi}}(s, a) \\ &= \frac{1}{1-\gamma} \left[\mathbb{E}_{s \sim \lambda^{\pi_E}, a \sim \pi_E(\cdot|s)} A^{\hat{\pi}}(s, a) - \mathbb{E}_{s \sim \lambda^{\pi_E}, a \sim \hat{\pi}(\cdot|s)} A^{\hat{\pi}}(s, a) \right] \\ &\leq \frac{1}{(1-\gamma)^2} \mathbb{E}_{s \sim \lambda^{\pi_E}} \|\hat{\pi}(\cdot|s) - \pi_E(\cdot|s)\|_1. \end{aligned}$$

□

Proof Sketch

- Recall the advantage defined as $A^{\hat{\pi}}(s, a) = Q^{\hat{\pi}}(s, a) - V^{\hat{\pi}}(s)$ and notice that $\mathbb{E}_{a \sim \hat{\pi}(\cdot|s)} A^{\hat{\pi}}(s, a) = 0, \quad \forall s.$
- We will use also that $A^{\hat{\pi}}(s, a) \leq \frac{1}{1-\gamma}$ if $\max_{s,a} |r(s, a)| \leq 1.$

Proof.

- ▶ Based on performance difference lemma [20], we have

$$\begin{aligned} V^{\pi_E} - V^{\hat{\pi}} &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \lambda^{\pi_E}, a \sim \pi_E(\cdot|s)} A^{\hat{\pi}}(s, a) \\ &= \frac{1}{1-\gamma} \left[\mathbb{E}_{s \sim \lambda^{\pi_E}, a \sim \pi_E(\cdot|s)} A^{\hat{\pi}}(s, a) - \mathbb{E}_{s \sim \lambda^{\pi_E}, a \sim \hat{\pi}(\cdot|s)} A^{\hat{\pi}}(s, a) \right] \\ &\leq \frac{1}{(1-\gamma)^2} \mathbb{E}_{s \sim \lambda^{\pi_E}} \|\hat{\pi}(\cdot|s) - \pi_E(\cdot|s)\|_1. \end{aligned}$$

- ▶ MLE guarantee [5] is given by

$$\mathbb{E}_{s \sim \lambda^{\pi_E}} \|\hat{\pi} - \pi_E\|_{TV}^2 \leq \frac{\log(|\Pi|/\delta)}{|\mathcal{D}|}.$$

□

Proof Sketch

- Recall the advantage defined as $A^{\hat{\pi}}(s, a) = Q^{\hat{\pi}}(s, a) - V^{\hat{\pi}}(s)$ and notice that $\mathbb{E}_{a \sim \hat{\pi}(\cdot|s)} A^{\hat{\pi}}(s, a) = 0, \quad \forall s.$
- We will use also that $A^{\hat{\pi}}(s, a) \leq \frac{1}{1-\gamma}$ if $\max_{s,a} |r(s, a)| \leq 1.$

Proof.

- Based on performance difference lemma [20], we have

$$\begin{aligned} V^{\pi_E} - V^{\hat{\pi}} &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \lambda^{\pi_E}, a \sim \pi_E(\cdot|s)} A^{\hat{\pi}}(s, a) \\ &= \frac{1}{1-\gamma} \left[\mathbb{E}_{s \sim \lambda^{\pi_E}, a \sim \pi_E(\cdot|s)} A^{\hat{\pi}}(s, a) - \mathbb{E}_{s \sim \lambda^{\pi_E}, a \sim \hat{\pi}(\cdot|s)} A^{\hat{\pi}}(s, a) \right] \\ &\leq \frac{1}{(1-\gamma)^2} \mathbb{E}_{s \sim \lambda^{\pi_E}} \|\hat{\pi}(\cdot|s) - \pi_E(\cdot|s)\|_1. \end{aligned}$$

- MLE guarantee [5] is given by

$$\mathbb{E}_{s \sim \lambda^{\pi_E}} \|\hat{\pi} - \pi_E\|_{TV}^2 \leq \frac{\log(|\Pi|/\delta)}{|\mathcal{D}|}.$$

- Then the result follows from Jensen's inequality and that $\|\cdot\|_{TV} = \frac{1}{2} \|\cdot\|_1.$

□

Behavioral cloning: Advantages and disadvantages

- Advantages
 - Simple.
 - Effective. For example in ALVINN [28].

Behavioral cloning: Advantages and disadvantages

- Advantages
 - Simple.
 - Effective. For example in ALVINN [28].
- Disadvantages
 - No long-term planning.
 - Cascading errors.
 - Possible mismatch between training and testing distributions.

Quote from Pomerleau [32]

When driving for itself, the network (ALVINN) may occasionally stray from the center of road and so must be prepared to recover by steering the vehicle back to the center of the road.

A key difference with supervised learning

- The dataset \mathcal{D} is collected according to π_E , therefore behavioural cloning outputs the policy with parameters

$$\arg \min_{\theta} \mathbb{E}_{s \sim \lambda \pi_E} \left[\ell(\pi_{\theta}(\cdot|s), \pi_E(\cdot|s)) \right].$$

- However when we act in the environment with π_{θ} the states are sampled accordingly to $\lambda^{\pi_{\theta}}$.
- It follows that we would like to minimize

$$\min_{\theta} \mathbb{E}_{s \sim \lambda^{\pi_{\theta}}} \left[\ell(\pi_{\theta}(\cdot|s), \pi_E(\cdot|s)) \right].$$

- Scenario different from supervised learning where the classification decisions do not affect the data distribution.

Overview

Introduction

Behavioral Cloning

Online Imitation Learning

Inverse Reinforcement Learning

Generative Adversarial Imitation Learning

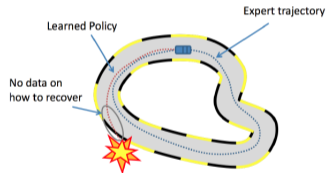
Non-adversarial imitation learning

Convex-analytic methods

Another variation along the theme: Behavioral cloning and interactive IL

- Behavioral cloning (BC) is a supervised learning approach to learning from demonstrations
 - ▶ Given an expert's demonstrations $\{(s_i, \pi_E(s_i))\}$ (offline trajectories or online queries)
 - ▶ Fix a loss: $\mathcal{L} : \mathcal{A} \rightarrow \mathbb{R}$
 - ▶ Output $\pi_* \in \operatorname{argmin}_{\pi} \sum_i^N \mathcal{L}(a_i, \pi(s_i))$ with a_i, s_i in the dataset provided by the expert.

- BC can result in cascading errors
 - ▶ Any error at a state can accumulate over an episode.
 - ▶ It can have catastrophic consequences...



- **Solution:** *Interactive IL* allows to query the expert policy from a particular state

Figure: <https://smartlabai.medium.com/a-brief-overview-of-imitation-learning-8a8a75c44a9c>

Interactive imitation learning

- Aims to mitigate the cascading errors through interacting with the expert.
- We assume that we can query the expert π_E at any time and any state sampled from λ^{π_θ} .

Interactive imitation learning

- Aims to mitigate the cascading errors through interacting with the expert.
- We assume that we can query the expert π_E at any time and any state sampled from λ^{π_θ} .
- **Idea:** Learn the expert's policy via **online learning**.

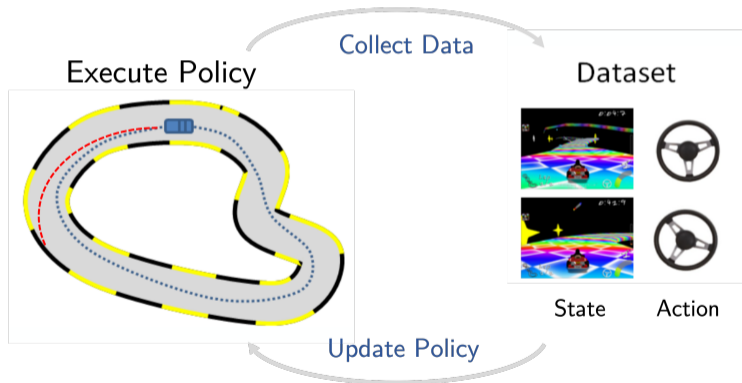


Figure: <https://smartlabai.medium.com/a-brief-overview-of-imitation-learning-8a8a75c44a9c>

Interactive imitation learning

- **Dataset Aggregation** (DAgger) [31]: iteratively build up a policy via supervised learning on aggregated data from the expert.
- **Policy Aggregation** (e.g., SMILe [32]): iteratively build up a policy by mixing newly trained policies.

Interactive imitation learning

- **Dataset Aggregation** (DAgger) [31]: iteratively build up a policy via supervised learning on aggregated data from the expert.
- **Policy Aggregation** (e.g., SMILe [32]): iteratively build up a policy by mixing newly trained policies.

Interactive imitation learning

Initialize π_0

for each iteration $t = 1, \dots, T$ **do**

 Generate trajectories τ following π_t

 Collect new data $\mathcal{D}_t = \{(s, \pi_E(s)) \mid s \in \tau\}$ based on expert's feedback

Data Aggregation: run behavioral cloning with $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_1 \cup \dots \cup \mathcal{D}_t$ and obtain π_t

Policy Aggregation: run behavioral cloning with \mathcal{D}_t and obtain $\hat{\pi}_t$, set $\pi_t = \beta \hat{\pi}_t + (1 - \beta) \pi_{t-1}$

end for

- Remark:**
- In the dataset \mathcal{D}_t the states are sampled accordingly to λ^{π_t} .
 - However, the actions are sampled from π_E . We need to assume that the expert is interactive.

Reduction to no-regret online learning

- Classical online optimization framework [39, 15, 9].
- Repeated game between the learner/player and the environment/adversary for any round $t = 1, \dots, T$.

Online learning protocol

- The learner picks a decision $\mathbf{x}_t \in X$;
- The adversary picks a loss $\ell_t(\cdot) : X \rightarrow \mathbb{R}$
- The learner suffers from the loss $\ell_t(\mathbf{x}_t)$ and observes some information about ℓ_t
- The goal is to minimize the player's regret against the best decision in hindsight:

$$\mathcal{R}_T := \sum_{t=1}^T \ell_t(\mathbf{x}_t) - \min_{\mathbf{x} \in X} \sum_{t=1}^T \ell_t(\mathbf{x}).$$

- **Follow-the-Regularized Leader** Algorithm (FTRL) [3]:

$$\mathbf{x}_t = \arg \min_{\mathbf{x} \in X} \sum_{i=1}^T \ell_i(\mathbf{x}) + \lambda R(\mathbf{x}), t = 1, \dots, T$$

- $R(\cdot)$ is a strongly convex and smooth function.

Offline vs online imitation learning

- o DAgger is equivalent to Follow-the-Leader, which ensures no regret $o(T)$ for strongly convex loss [35].
- o BC can be viewed as a special case with only one iteration.
- o DAgger has improved performance bound from $\mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right)$ to $\mathcal{O}\left(\frac{\max_{s,a} |A^{\pi_E}(s,a)|}{1-\gamma}\right)$ [4].

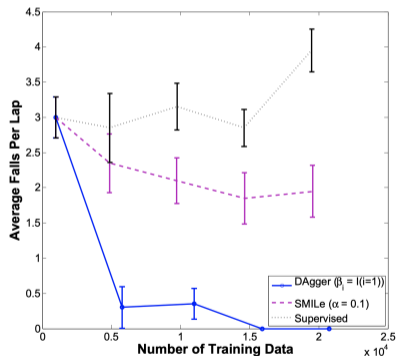
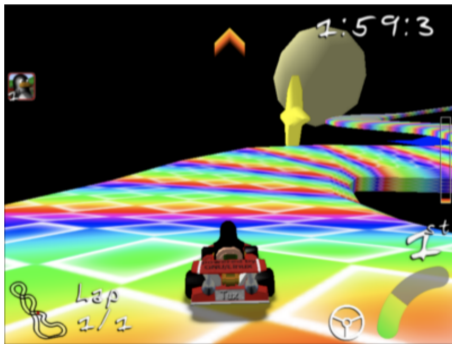


Figure: 3D racing car [31]

Overview

Introduction

Behavioral Cloning

Online Imitation Learning

Inverse Reinforcement Learning

Generative Adversarial Imitation Learning

Non-adversarial imitation learning

Convex-analytic methods

The imitation learning perspective

- Standard imitation learning

- ▶ copy the *actions* performed by the expert
- ▶ no reasoning about outcomes of actions



Figure: Robot imitation

- Human imitation learning

- ▶ copy the *intent* of the expert
- ▶ might take very different actions!



Figure: Human imitation

Inverse reinforcement learning (IRL) [25, 33]

IRL Objective

Find reward function $r(\cdot, \cdot) : \mathcal{S} \times \mathcal{A} \rightarrow [-1, 1]$ that explains the expert's behavior:

$$\pi_E \in \arg \max_{\pi \in \Pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 \sim \mu, \pi \right].$$

Inverse reinforcement learning (IRL) [25, 33]

IRL Objective

Find reward function $r(\cdot, \cdot) : \mathcal{S} \times \mathcal{A} \rightarrow [-1, 1]$ that explains the expert's behavior:

$$\pi_E \in \arg \max_{\pi \in \Pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 \sim \mu, \pi \right].$$

Namely, it holds that

$$\mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 \sim \mu, \pi_E \right] \geq \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 \sim \mu, \pi \right], \forall \pi \in \Pi.$$

- Remarks:**
- Assume the expert is optimizing some reward function r_{true} .
 - The true reward function is unknown; π_E is the optimal policy of the MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r_{\text{true}}, \gamma)$.
 - Unlike BC, IRL uses the MDP structure for the learning from expert demonstration.
 - IRL recovers a reward function and avoids the distribution shift issue in BC [2, 38].
 - Note that this is a convex feasibility problem: It has different solution challenges.

The RL and IRL dichotomy

	IRL	RL
Input	Expert Demonstrations	Reward Function
Output	Optimal policy Reward function	Optimal Policy

- RL recovers a nearly optimal behavior from reward functions
- IRL recovers a nearly optimal behavior from demonstrations by an expert

Challenges with inverse reinforcement learning

- Reward function ambiguity; A trivial solution is $r = 0$.

Challenges with inverse reinforcement learning

- Reward function ambiguity; A trivial solution is $r = 0$.

Theorem (Reward shaping)

An expert policy π_E optimal in the MDP \mathcal{M} with reward r is optimal also in the MDP \mathcal{M} with reward function \hat{r} given by

$$\hat{r}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\Phi(s')] - \Phi(s),$$

where $\Phi : S \rightarrow \mathbb{R}$ is called potential function.

Challenges with inverse reinforcement learning

- Reward function ambiguity; A trivial solution is $r = 0$.

Theorem (Reward shaping)

An expert policy π_E optimal in the MDP \mathcal{M} with reward r is optimal also in the MDP \mathcal{M} with reward function \hat{r} given by

$$\hat{r}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\Phi(s')] - \Phi(s),$$

where $\Phi : S \rightarrow \mathbb{R}$ is called potential function.

- ▶ **Solution:** Add regularization, restrict reward assumptions

Challenges with inverse reinforcement learning

- Reward function ambiguity; A trivial solution is $r = 0$.

Theorem (Reward shaping)

An expert policy π_E optimal in the MDP \mathcal{M} with reward r is optimal also in the MDP \mathcal{M} with reward function \hat{r} given by

$$\hat{r}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\Phi(s')] - \Phi(s),$$

where $\Phi : \mathcal{S} \rightarrow \mathbb{R}$ is called potential function.

- ▶ **Solution:** Add regularization, restrict reward assumptions
- IRL is computationally expensive if we want to enumerate all policies to form the constraints.

Challenges with inverse reinforcement learning

- Reward function ambiguity; A trivial solution is $r = 0$.

Theorem (Reward shaping)

An expert policy π_E optimal in the MDP \mathcal{M} with reward r is optimal also in the MDP \mathcal{M} with reward function \hat{r} given by

$$\hat{r}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\Phi(s')] - \Phi(s),$$

where $\Phi : \mathcal{S} \rightarrow \mathbb{R}$ is called potential function.

- ▶ **Solution:** Add regularization, restrict reward assumptions
- IRL is computationally expensive if we want to enumerate all policies to form the constraints.
 - ▶ **Solution:** Consider a tractable apprenticeship learning formalism

Challenges with inverse reinforcement learning

- Reward function ambiguity; A trivial solution is $r = 0$.

Theorem (Reward shaping)

An expert policy π_E optimal in the MDP \mathcal{M} with reward r is optimal also in the MDP \mathcal{M} with reward function \hat{r} given by

$$\hat{r}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\Phi(s')] - \Phi(s),$$

where $\Phi : \mathcal{S} \rightarrow \mathbb{R}$ is called potential function.

- ▶ **Solution:** Add regularization, restrict reward assumptions
- IRL is computationally expensive if we want to enumerate all policies to form the constraints.
 - ▶ **Solution:** Consider a tractable apprenticeship learning formalism
- In practice, we do not observe π_E but only trajectories from π_E .

Challenges with inverse reinforcement learning

- Reward function ambiguity; A trivial solution is $r = 0$.

Theorem (Reward shaping)

An expert policy π_E optimal in the MDP \mathcal{M} with reward r is optimal also in the MDP \mathcal{M} with reward function \hat{r} given by

$$\hat{r}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\Phi(s')] - \Phi(s),$$

where $\Phi : \mathcal{S} \rightarrow \mathbb{R}$ is called potential function.

- ▶ **Solution:** Add regularization, restrict reward assumptions
- IRL is computationally expensive if we want to enumerate all policies to form the constraints.
 - ▶ **Solution:** Consider a tractable apprenticeship learning formalism
- In practice, we do not observe π_E but only trajectories from π_E .
 - ▶ **Solution:** Use sample averages of total returns under π_E

Challenges with inverse reinforcement learning

- Reward function ambiguity; A trivial solution is $r = 0$.

Theorem (Reward shaping)

An expert policy π_E optimal in the MDP \mathcal{M} with reward r is optimal also in the MDP \mathcal{M} with reward function \hat{r} given by

$$\hat{r}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\Phi(s')] - \Phi(s),$$

where $\Phi : \mathcal{S} \rightarrow \mathbb{R}$ is called potential function.

- ▶ **Solution:** Add regularization, restrict reward assumptions
- IRL is computationally expensive if we want to enumerate all policies to form the constraints.
 - ▶ **Solution:** Consider a tractable apprenticeship learning formalism
- In practice, we do not observe π_E but only trajectories from π_E .
 - ▶ **Solution:** Use sample averages of total returns under π_E
- May be infeasible if the expert's policy is not optimal.

Challenges with inverse reinforcement learning

- Reward function ambiguity; A trivial solution is $r = 0$.

Theorem (Reward shaping)

An expert policy π_E optimal in the MDP \mathcal{M} with reward r is optimal also in the MDP \mathcal{M} with reward function \hat{r} given by

$$\hat{r}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\Phi(s')] - \Phi(s),$$

where $\Phi : S \rightarrow \mathbb{R}$ is called potential function.

- ▶ **Solution:** Add regularization, restrict reward assumptions
- IRL is computationally expensive if we want to enumerate all policies to form the constraints.
 - ▶ **Solution:** Consider a tractable apprenticeship learning formalism
- In practice, we do not observe π_E but only trajectories from π_E .
 - ▶ **Solution:** Use sample averages of total returns under π_E
- May be infeasible if the expert's policy is not optimal.
 - ▶ **Solution:** Relax the constraints; add slack variables

Identifiability in inverse reinforcement learning

- The reward function ambiguity problem can be solved leveraging two experts. The following holds:

Theorem (Theorem 2 in [30])

Consider two Markov decision problems on the same set of states and actions, but with different transition matrices P^1, P^2 and discount factors γ_1, γ_2 . Suppose that we observe two experts acting each in one of these environments, optimally with respect to the same reward function, in the sense that their policies maximize the entropy regularized reward in their respective environments. Then, the reward function can be recovered up to the addition of a constant if and only if

$$\text{rank} \begin{pmatrix} I - \gamma_1 P_{a_1}^1 & -(I - \gamma_2 P_{a_1}^2) \\ \vdots & \vdots \\ I - \gamma_1 P_{a_{|\mathcal{A}|}}^1 & -(I - \gamma_2 P_{a_{|\mathcal{A}|}}^2) \end{pmatrix} = 2|\mathcal{S}| - 1. \quad (3)$$

- Remark:**
- This result has been stated very recently in [8] under a limited form.
 - This stronger statement is a new result.
 - Identifying the reward is important when one needs to predict how the expert would behave under different dynamics but same reward.

Feature-based reward

Theorem

Assumption Let $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \ni (s, a) \mapsto \phi(s, a) \in \mathbb{R}^d$ be a feature mapping. Assume linear true reward function, i.e.,

$$r_{true} \in \{r \mid r(s, a) = w^\top \phi(s, a), \text{ where } w \in \mathbb{R}^d \text{ and } \|w\|_2 \leq 1\}$$

o The expected total reward when $r(s, a) = w^\top \phi(s, a)$ can then be expressed as:

$$J_r(\pi) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid \pi \right] = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t w^\top \phi(s_t, a_t) \mid \pi \right] = w^\top \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) \mid \pi \right] = w^\top \rho_\phi(\pi),$$

where $\rho_\phi(\pi) \in \mathbb{R}^d$ is the **feature expectation vector** of policy π .

Goal

Find $w \in \mathbb{R}^d$ such that

$$\underbrace{w^\top \rho_\phi(\pi_E)}_{=J_w(\pi_E)} \geq \underbrace{w^\top \rho_\phi(\pi)}_{=J_w(\pi)}, \quad \forall \pi \in \Pi.$$

Feature-based reward (cont'd)

Goal

Find $w \in \mathbb{R}^d$ such that

$$\underbrace{w^\top \rho_\phi(\pi_E)}_{=J_w(\pi_E)} \geq \underbrace{w^\top \rho_\phi(\pi)}_{=J_w(\pi)}, \quad \forall \pi \in \Pi.$$

- Remark:**
- Note that $\rho_\phi(\pi)$ can be readily estimated from sampled trajectories.
 - By Hoeffding's Lemma we need $\mathcal{O}\left(\frac{d \log(\frac{1}{\delta})}{(1-\gamma)^2 \varepsilon^2}\right)$ expert trajectories to have an ε -small ℓ_∞ -error with probability at least $1 - \delta$.

Max margin IRL [Ratliff et al., 2006][29]

Standard max margin

$$\begin{aligned} \min_w \quad & \|w\|_2^2 \\ \text{s.t.} \quad & w^\top \rho_\phi(\pi_E) \geq w^\top \rho_\phi(\pi) + 1, \quad \text{for all } \pi \end{aligned}$$

Structured prediction max margin

$$\begin{aligned} \min_w \quad & \|w\|_2^2 \\ \text{s.t.} \quad & w^\top \rho_\phi(\pi_E) \geq w^\top \rho_\phi(\pi) + m(\pi_E, \pi), \quad \text{for all } \pi \end{aligned}$$

- Remarks:**
- Margin should be larger for policies that are very different from π_E
 - Example: $m(\pi_E, \pi)$ = number of states in which π_E was observed and in which π and π_E disagree.

Max margin IRL [Ratliff et al., 2006][29] (cont')

Structured prediction max margin with slack variables

$$\begin{aligned} \min_{w, \xi} \quad & \|w\|_2^2 + C\xi \\ \text{s.t.} \quad & w^\top \rho_\phi(\pi_E) \geq w^\top \rho_\phi(\pi) + m(\pi_E, \pi) - \xi, \quad \text{for all } \pi \end{aligned}$$

- Remarks:**
- Resolved: access to π_E , reward ambiguity, expert suboptimality
 - One challenge remains: very large number of constraints
 - Assuming access to an RL subroutine, it can be solved, e.g., by constraint generation.

Feature expectation matching

- Note that if $\|\rho_\phi(\pi_E) - \rho_\phi(\pi)\|_2 \leq \epsilon$, then

$$w^\top \rho_\phi(\pi_E) - w^\top \rho_\phi(\pi) \leq \|w\|_2 \|\rho_\phi(\pi) - \rho_\phi(\pi_E)\|_2 \leq \epsilon.$$

- Therefore, it suffices to solve [2]

$$\min_{\pi} \|\rho_\phi(\pi) - \rho_\phi(\pi_E)\|_2^2 \quad (4)$$

Apprenticeship learning formalism

Assume that $r_{\text{true}} \in \mathcal{R}$. Solve

$$\min_{\pi} \max_{r \in \mathcal{R}} J_r(\pi_E) - J_r(\pi) = \min_{\pi} \max_{r \in \mathcal{R}} \langle \lambda^{\pi_E} - \lambda^{\pi}, r \rangle \quad (5)$$

Remark:

- When $\mathcal{R} = \{\sum_{i=1}^d w_i \phi_i \mid \|w\|_2 \leq 1\}$ the minimax problem (5) is reduced to (4)
- The quantity $\delta_{\mathcal{R}} := \max_{r \in \mathcal{R}} \langle \lambda^{\pi_E} - \lambda^{\pi}, r \rangle$ is called \mathcal{R} -distance and is an integral probability metric between λ^{π} and λ^{π_E}
- Different choices of \mathcal{R} lead to different \mathcal{R} -distances.

Maximum entropy inverse reinforcement learning [Ziebart et al, 2008 [38]]

- Consider the constrained optimization for feature expectation matching:

Max-Ent IRL

Let λ^π be the state-action occupancy measure of policy π . Consider the following problem:

$$\begin{aligned} \max_{\pi \in \Pi} \quad & \mathbb{E}_{(s,a) \sim \lambda^\pi} [-\log \pi(a|s)] \\ \text{s.t.} \quad & \mathbb{E}_{(s,a) \sim \lambda^\pi} [\phi(s, a)] = \mathbb{E}_{s \sim \lambda^{\pi_E}, a \sim \pi_E(\cdot|s)} [\phi(s, a)]. \end{aligned}$$

measure.

- This is equivalent to the min-max optimization reformulation:

$$\min_w \max_{\pi \in \Pi} w^\top \left(\mathbb{E}_{(s,a) \sim \lambda^\pi} [\phi(s, a)] - \mathbb{E}_{(s,a) \sim \lambda^{\pi_E}} [\phi(s, a)] \right) + \mathbb{E}_{(s,a) \sim \lambda^\pi} [-\log \pi(a|s)].$$

- Game-theoretic perspective: zero-sum game between the reward and the policy.

Solving the saddle point problem

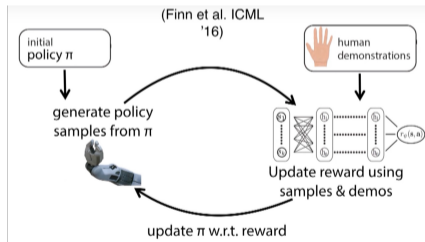
- One can employ a bi-level optimization algorithm
- Let $f(w) = \max_{\pi \in \Pi} w^\top \left(\mathbb{E}_{s,a \sim \lambda \pi} [\phi(s, a)] - \mathbb{E}_{s,a \sim \lambda \pi_E} [\phi(s, a)] \right) + \mathbb{E}_{s,a \sim \lambda \pi} [-\log \pi(a|s)]$.
- Evaluating $f(w)$ requires solving an RL problem with reward $w^\top \phi(s, a) - \log \pi(a|s)$.
- Let π_* be the optimal policy for this reward.
- By Danskin's theorem [11], we can compute $\nabla_w f(w) = \left(\mathbb{E}_{s,a \sim \lambda \pi_*} [\phi(s, a)] - \mathbb{E}_{s,a \sim \lambda \pi_E} [\phi(s, a)] \right)$.
- And update the reward weights w by gradient descent.

Maximum entropy inverse reinforcement learning

Max-Ent IRL Algorithm

Alternatively update

- update w by GD (with fixed π);
- update π by any RL algorithm for the corresponding entropy-regularized MDP (with fixed w)



Overview

Introduction

Behavioral Cloning

Online Imitation Learning

Inverse Reinforcement Learning

Generative Adversarial Imitation Learning

Non-adversarial imitation learning

Convex-analytic methods

Generative adversarial imitation learning (GAIL): A primal dual perspective

- In Maximum Causal Entropy IRL [38], we need to solve an RL problem for every reward update.
- This is a major computation bottleneck.
- We can develop a more efficient method if we use alternating updates.

Derivation: ○ We will follow the same steps from [16]

GAIL objective

Let $h : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \rightarrow \mathbb{R}$ be a convex function that serves as reward regularizer. GAIL solves the following minimax problem:

$$\min_r \max_{\pi \in \Pi} h(r) + \mathbb{E}_{s,a \sim \lambda \pi} [r(s,a)] - \mathbb{E}_{s,a \sim \lambda \pi_E} [r(s,a)] + \mathbb{E}_{s,a \sim \lambda \pi} [-\log \pi(a|s)]$$

- Use Fenchel conjugation, we can obtain

$$\max_{\pi \in \Pi} -h^*(\lambda^{\pi_E} - \lambda^{\pi}) + \mathbb{E}_{s,a \sim \lambda \pi} [-\log \pi(a|s)].$$

An important choice for the regularizer h .

- Choosing h as

$$h(r) = \begin{cases} \mathbb{E}_{s,a \sim \lambda^{\pi_E}} [g(r(s,a))], & \text{if } r(s,a) < 0; \\ \infty, & \text{otherwise.} \end{cases}$$

with $g(x) = -x - \log(1 - e^x)$.

- The Fenchel conjugate of h is given by:

$$h^*(\lambda^{\pi_E} - \lambda^\pi) = \max_{D \in [0,1]} \mathbb{E}_{s,a \sim \lambda^\pi} [\log D(s,a)] + \mathbb{E}_{s,a \sim \lambda^{\pi_E}} [\log(1 - D(s,a))]$$

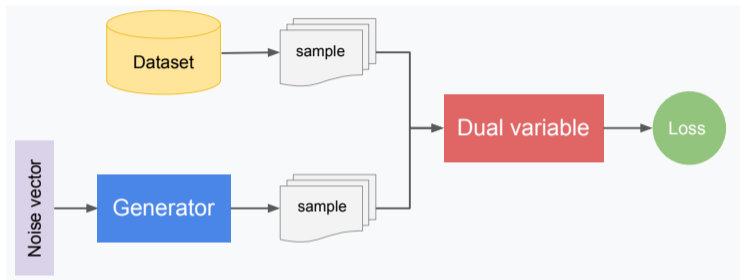
that is widely known as the (vanilla) GAN loss.

- Therefore, we can learn a policy from demonstrations solving the following saddle point problem:

$$\min_{\pi \in \Pi} \max_{D \in [0,1]} \mathbb{E}_{s,a \sim \lambda^\pi} [\log D(s,a)] + \mathbb{E}_{s,a \sim \lambda^{\pi_E}} [\log(1 - D(s,a))] - \mathbb{E}_{s,a \sim \lambda^\pi} [-\log \pi(a|s)].$$

Generative Adversarial Network (GANs)

- o GAN [14] is framed as a min-max game between a generator and a discriminator.



- o **GAN:** (\Rightarrow minimizing the Jensen-Shannon divergence)

$$\min_{G_\phi} \max_{D_\theta} \mathbb{E}_{x \sim p_{\text{data}}} [\log D_\theta(x)] + \mathbb{E}_z [\log(1 - D_\theta(G_\phi(z)))]$$

- o **Wasserstein GAN:** (\Rightarrow minimizing the Wasserstein divergence)

$$\min_{G_\phi} \max_{f_\theta: 1\text{-Lipschitz}} \mathbb{E}_{x \sim p_{\text{data}}} [f_\theta(x)] - \mathbb{E}_z [f_\theta(G_\phi(z))]$$

Generative Adversarial Networks (GANs)



2014
GAN



2018
GAN

Generative Adversarial Imitation Learning (GAIL)

- GAIL [17] aims to solve the min-max game for learning the policy given an expert policy π_E .

$$\min_{\theta} \max_{\phi} \mathbb{E}_{s,a \sim \lambda \pi_{\theta}} [\log(C_{\phi}(s,a))] + \mathbb{E}_{s,a \sim \lambda \pi_E} [\log(1 - C_{\phi}(s,a))] - \lambda H(\pi_{\theta}).$$

- Remarks:**
- We assume a differentiable parametrized policy π_{θ} .
 - The discriminator tries to separate the data generated from learned policy from expert data.
 - Equivalent to minimize the Jensen-Sannon divergence between the state-action distributions of the expert policy and the learned policy.
 - Unlike Max-Entropy IRL, does not require expensive RL subroutines to learn the reward.

Numerical Performance [17]

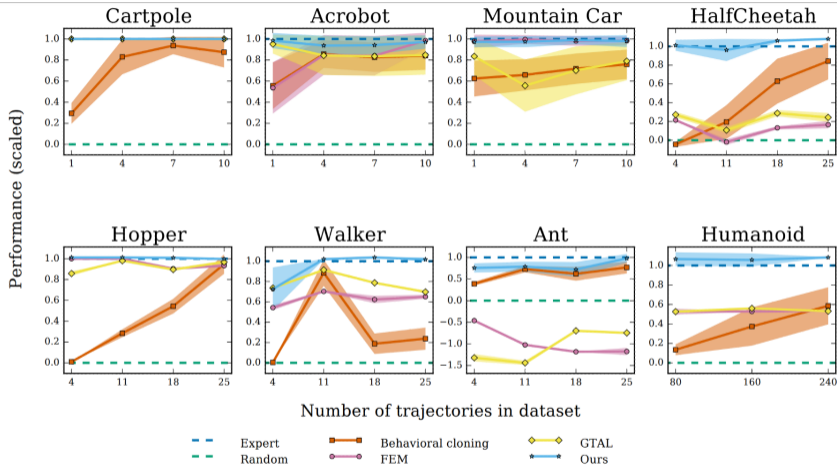


Figure: Performance of learned policies among GAIL, Behavior Cloning (BC), Feature Expectation Matching (FEM), and Game-theoretic Apprenticeship Learning (GATL)

Overview

Introduction

Behavioral Cloning

Online Imitation Learning

Inverse Reinforcement Learning

Generative Adversarial Imitation Learning

Non-adversarial imitation learning

Convex-analytic methods

The IQ-Learn optimization problem [13]

- The core idea is to use the expert data to learn a state action value function.
- We can see IQ-Learn as a double smoothing approach.
- We add a strongly convex function occupancy measure dependent function $H(\cdot|\lambda_0)$
- Analogously, we add a strongly concave function dependent on the reward variable r .

$$\min_{\lambda \in \tilde{\mathcal{X}}} \max_r \langle \lambda_{\pi_E} - \lambda, r \rangle + \frac{1}{\chi} \psi(r) + \frac{1}{\eta} H(\lambda, \lambda_0),$$

where H is the relative conditional entropy defined as $H(\lambda, \lambda_0) := \sum_{x,a} \lambda(x,a) \log \frac{\lambda(x,a) \sum_a \lambda_{\pi_0}(x,a)}{\lambda_{\pi_0}(x,a) \sum_a \lambda(x,a)}$.

- $\psi(r)$ is restricted to a particular form, i.e. $\psi(r) = \langle \lambda_{\pi_E}, r - \phi(r) \rangle$, with $\phi : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \rightarrow \mathbb{R}$ being a convex and non-increasing function.

IQ-Learn equivalent unconstrained problem

IQ-Learn Program over Q -functions

Replacing the optimal policy $\pi_Q(a|s) \propto \exp(Q(s, a))$, we obtain an unconstrained problem.

$$\tilde{Q} \approx \arg \max_Q (1 - \gamma) \langle \mu, V_Q^{\pi_Q} \rangle - \langle \lambda_{\pi_E}, \phi(Q - \gamma P V_Q^{\pi_Q}) \rangle$$

Remarks:

- The approach is very similar to REPS.
- However, the derivation of the unconstrained problem is not straightforward and requires assumptions on ψ .
- The formulation is concave w.r.t. Q .
- The empirical performance of this algorithm is very convincing.
- Lack of convergence guarantees.

Overview

Introduction

Behavioral Cloning

Online Imitation Learning

Inverse Reinforcement Learning

Generative Adversarial Imitation Learning

Non-adversarial imitation learning

Convex-analytic methods

Linear programming approach for imitation learning

- Let \mathcal{R} be a class of reward functions.
- The following LP outputs the occupancy measure under the worst case reward in \mathcal{R} .

LP for imitation learning

$$\max_{\lambda} \min_{r \in \mathcal{R}} \langle \lambda - \lambda^{\pi_E}, r \rangle \quad (6)$$

$$\text{s.t. } E^{\top} \lambda = \gamma P^{\top} \lambda + (1 - \gamma) \mu \quad (7)$$

Remarks:

- There are $|\mathcal{S}| + |\mathcal{S}||\mathcal{A}|$ decision variables.
- There are $|\mathcal{S}|$ constraints.
- To avoid the large number of constraints, [21] propose to study the Lagrangian.
- To reduce the number of decision variables, [21] uses linear function approximation.

The Lagrangian

- Let \mathcal{R} be a class of reward functions such that $r_{\text{true}} \in \mathcal{R}$
- The following LP outputs the occupancy measure under the worst case reward in \mathcal{R} .

Saddle point formulation for imitation learning

$$\max_{\lambda} \min_{r \in \mathcal{R}} \min_V \langle \lambda - \lambda^{\pi_E}, r \rangle + \langle V, -E^\top \lambda + \gamma P^\top \lambda + (1 - \gamma)\mu \rangle \quad (8)$$

Remarks:

- Notice that the number of decision variables is $|\mathcal{S}| + 2|\mathcal{S}||\mathcal{A}|$.
- Hence, we can parameterize the occupancy measure as $\lambda_\theta = \Phi\theta$, $V_w = \Psi w$ and $r = C\beta$.
- This parametrization helps reduce the number of decision variables significantly.
- The value parametrization has precedence in earlier RL literature.
- The occupancy measure parameterization is done out of necessity.

The reduced Lagrangian

- Introducing the linear function approximation we obtain the reduced Lagrangian.
- The number of decision variables is now $\dim(\theta) + \dim(w) + \dim(\beta)$.

Saddle Point for imitation learning

$$\max_{\theta \in \Delta} \min_{\beta \in \Delta} \min_{\|w\|_{\infty} \leq C} \langle \Phi\theta - \lambda^{\pi_E}, C\beta \rangle + \langle \Psi w, -E^{\top} \Phi\theta + \gamma P^{\top} \Phi\theta + (1 - \gamma)\mu \rangle \quad (9)$$

Remarks:

- We can solve the problem applying stochastic mirror prox [19].
- With this approach we get an ϵ optimal policy with $\mathcal{O}(\epsilon^{-2})$ samples.
- The sample complexity is independent of $|\mathcal{S}|$ and $|\mathcal{A}|$ due to the parametrization.
- A drawback is that one needs a strong assumption on the feature choice (see [21, 6]).

The Linear MDP Assumption

Linear MDP [18]

There exist mappings $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^m$ and $g : \mathcal{S} \rightarrow \mathbb{R}^m$ and a vector $w \in \mathcal{W} := \{w \in \mathbb{R}^m : \|w\|_2 \leq 1\}$ such that

$$r(s, a) = \langle \phi(s, a), w \rangle$$

$$P(s'|s, a) = \langle \phi(s, a), g(s') \rangle$$

that is, in matrix form

$$r = \Phi w$$

$$P = \Phi M$$

Remarks:

- The Linear MDP is a standard setting in RL theory literature.
- It justifies an alternative LP formulation.

The constraint splitting trick

- P²IL [36] is derived from the primal problem for imitation learning.
- We plug in the (Linear MDP) structure in (Primal IL) (6) and we split the as follows ¹

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}} \quad & \min_{w \in \mathcal{W}} \langle \lambda - \lambda^{\pi_E}, \Phi w \rangle \\ \text{s.t.} \quad & E^\top \lambda = (1 - \gamma)\mu + \gamma M^\top \Phi^\top \lambda \end{aligned}$$

⇓

$$\begin{aligned} \max_{\rho \in \Delta^m, \lambda \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}} \quad & \min_{w \in \mathcal{W}} \langle \rho - \Phi^\top \lambda^{\pi_E}, w \rangle \\ \text{s.t.} \quad & E^\top \lambda - \gamma M^\top \rho = (1 - \gamma)\mu \\ & \Phi^\top \lambda = \rho \end{aligned}$$

- Now we can apply on the Lagrangian, inexact proximal point updates for λ and ρ .

¹A similar trick appeared outside the imitation learning in [24], [23] and [7]

The algorithm: P²IL

Proximal Point Imitation Learning: P²IL

Initialize π_0 as uniform distribution over \mathcal{A}

for $k = 1, \dots, K$ **do**

// Policy evaluation

$$(w_k, \theta_k) \approx \arg \min_{w \in \mathcal{W}, \theta \in \Theta} \mathcal{G}_k(w, \theta)$$

// Policy improvement

$$\pi_k(a|s) \propto \pi_{k-1}(a|s) e^{-\alpha Q_{\theta_k}(s,a)}$$

end for

◦ $\mathcal{G}_k(w, \theta)$, called logistic Bellman error [7], is the following convex and smooth function:

$$\mathcal{G}_k(w, \theta) \triangleq \frac{1}{\eta} \log \sum_{i=1}^m (\Phi^\top \lambda_{k-1})(i) e^{\eta \delta_{w, \theta}^k(i)} + (1 - \gamma) \langle \mu, V_\theta^k \rangle - \langle \lambda_{\pi_E}, \Phi^\top w \rangle,$$

$$\delta_{w, \theta}^k \triangleq w + \gamma M V_\theta^k - \theta \quad \text{and} \quad V_\theta^k \triangleq \frac{1}{\alpha} \log \left(\sum_a \pi_{\lambda_{k-1}}(a|s) e^{\alpha Q_\theta(s,a)} \right) \quad \text{where} \quad Q_\theta = \Phi \theta$$

Sample Complexity Guarantees for P²IL

- We consider errors in the maximization of $\mathcal{G}_k(w, \theta)$, i.e. $\epsilon_k = \mathcal{G}_k(w_k^*, \theta_k^*) - \mathcal{G}_k(w_k, \theta_k)$.
- First, we show how errors propagate.
- Second, we control that the errors are small using a Biased Stochastic Gradient Ascent subroutine.

Error propagation

Let $\hat{\pi}_K$ be the average iterate. Then, with probability at least $1 - \delta$, it holds that

$$d_C(\lambda_{\hat{\pi}_K}, \lambda_{\pi_E}) \leq \frac{1}{K} \left(\log(m|\mathcal{A}|) + C \sum_k \sqrt{\epsilon_k} + \sum_k \epsilon_k \right).$$

Error control

Let (w_k, θ_k) be the output of the **Biased Stochastic Gradient Ascent** subroutine for T iterations. Then, $\epsilon_k = \max_{w, \theta} \mathcal{G}_k(w, \theta) - \mathcal{G}_k(w_k, \theta_k) \leq \mathcal{O}\left(\frac{\max\{\eta, 1\}m}{\beta \sqrt{T}}\right)$, with probability $1 - \delta$.

A downside: exploration assumptions

Remarks:

- Choosing $K = \Omega(\epsilon^{-1})$ and $T = \Omega(\epsilon^{-4})$ we obtain $\mathcal{O}(\epsilon^{-5})$ sample complexity.
- We use samples to approximate the gradients $\nabla_{\theta} \mathcal{G}_k$ and $\nabla_w \mathcal{G}_k$.
- In REPS, [27] required the following assumption.

Exploration assumption

We can sample state action pairs from an occupancy measure $\lambda_{\pi_0}(s, a) > 0 \quad \forall s, a \in \mathcal{S} \times \mathcal{A}$.

- In our extension to Linear MDP, we require the following assumption.

Positive Definite Covariance Matrix

We can sample state action pairs from an occupancy measure λ_{π_0} such that.

$$\sigma_{\min} \left(\mathbb{E}_{s, a \sim \lambda_{\pi_0}} \phi(s, a) \phi(s, a)^{\top} \right) \geq \beta > 0.$$

Summary of imitation learning

Method	Reward learning	Access to environment	Interactive demonstrations	Pre-collected demonstrations
Behavioural Cloning	NO	NO	NO	YES
Online IL	NO	YES	YES	MAYBE
Inverse RL	YES	YES	NO	YES
Adversarial IL	MAYBE	YES	NO	YES
Non-adversarial IL	MAYBE	YES	NO	YES

Remarks:

- BC avoids interaction with the environment, but can suffer from cascading errors.
- Online IL helps with the cascading errors but requires (expensive) expert queries.
- IRL explains the expert's behavior but has poor sample complexity and scalability.
- Adversarial IL avoids solving the RL problem repeatedly but are unstable due to adversarial training.
- Non-adversarial IL enjoys stable performance but is hampered by limited theoretical understanding.

References I

- [1] Pieter Abbeel, Adam Coates, Morgan Quigley, and Andrew Ng.
An application of reinforcement learning to aerobatic helicopter flight.
Advances in neural information processing systems, 19, 2006.
9

- [2] Pieter Abbeel and Andrew Y Ng.
Apprenticeship learning via inverse reinforcement learning.
In Proceedings of the twenty-first international conference on Machine learning, page 1, 2004.
9, 34, 35, 51

- [3] Jacob Abernethy, Elad Hazan, and Alexander Rakhlin.
Competing in the dark: An efficient algorithm for bandit linear optimization.
In In Proceedings of the 21st Annual Conference on Learning Theory (COLT, 2008.
30

- [4] Alekh Agarwal, Nan Jiang, Sham M. Kakade, and Wen Sun.
Reinforcement learning: Theory and algorithms, 2020.
16, 17, 31

- [5] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun.
Flambe: Structural complexity and representation learning of low rank mdps.
In Neural Information Processing Systems (NeurIPS), 2020.
18, 19, 20

References II

- [6] J. Bas-Serrano and G. Neu.
Faster saddle-point optimization for solving large-scale Markov decision processes.
In *Conference on Learning for Dynamics and Control (L4DC)*, 2020.
68
- [7] Joan Bas-Serrano, Sebastian Curi, Andreas Krause, and Gergely Neu.
Logistic Q-learning.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.
70, 71
- [8] Haoyang Cao, Samuel N. Cohen, and Lukasz Szpruch.
Identifiability in inverse reinforcement learning, 2021.
46
- [9] Nicolo Cesa-Bianchi, Gabor Lugosi, and Learning Prediction.
Games, 2006.
30
- [10] Adam Coates, Pieter Abbeel, and Andrew Y Ng.
Learning for control from multiple demonstrations.
In *Proceedings of the 25th international conference on Machine learning*, pages 144–151, 2008.
9

References III

[11] J. Danskin.

The theory of max-min, with applications.

SIAM Journal on Applied Mathematics, 14(4):641–664, 1966.

53

[12] Chelsea Finn, Sergey Levine, and Pieter Abbeel.

Guided cost learning: Deep inverse optimal control via policy optimization.

In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 49–58, New York, New York, USA, 20–22 Jun 2016. PMLR.

9

[13] Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon.

IQ-learn: Inverse soft-Q learning for imitation.

In *Advances in Neural Information Processing Systems*, 2021.

63

[14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio.

Generative Adversarial Networks.

ArXiv e-prints, June 2014.

58

[15] Elad Hazan.

Introduction to online convex optimization.

Foundations and Trends® in Optimization, 2(3-4):157–325, 2016.

30

References IV

- [16] Jonathan Ho and Stefano Ermon.
Generative adversarial imitation learning.
In *Advances in Neural Information Processing Systems*, pages 4565–4573, 2016.
56
- [17] Jonathan Ho and Stefano Ermon.
Generative adversarial imitation learning.
In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
60, 61
- [18] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan.
Provably efficient reinforcement learning with linear function approximation.
In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.
69
- [19] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel.
Solving variational inequalities with stochastic mirror-prox algorithm.
Stochastic Systems, 1(1):17–58, 2011.
68
- [20] Sham Kakade and John Langford.
Approximately optimal approximate reinforcement learning.
In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.
18, 19, 20

References V

- [21] Angeliki Kamoutsis, Goran Banjac, and John Lygeros.
Efficient performance bounds for primal-dual reinforcement learning from demonstrations.
In *International Conference on Machine Learning (ICML)*.
66, 68
- [22] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert.
Activity forecasting.
In *European conference on computer vision*, pages 201–214. Springer, 2012.
9
- [23] Donghwan Lee and Niao He.
Stochastic primal-dual q-learning algorithm for discounted mdps.
In *2019 american control conference (acc)*, pages 4897–4902. IEEE, 2019.
70
- [24] Prashant G Mehta and Sean P Meyn.
Convex q-learning, part 1: Deterministic optimal control.
arXiv preprint arXiv:2008.03559, 2020.
70
- [25] A. Y. Ng and S. J. Russell.
Algorithms for inverse reinforcement learning.
In *International Conference on Machine Learning (ICML)*, 2000.
7, 34, 35

References VI

- [26] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.
Training language models to follow instructions with human feedback.
arXiv preprint arXiv:2203.02155, 2022.
8
- [27] Aldo Pacchiano, Jonathan Lee, Peter Bartlett, and Ofir Nachum.
Near optimal policy optimization via REPS.
In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
73
- [28] Dean A. Pomerleau.
Alvinn: An autonomous land vehicle in a neural network.
In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1989.
7, 21, 22
- [29] Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich.
Maximum margin planning.
In *Proceedings of the 23rd international conference on Machine learning*, pages 729–736. ACM, 2006.
49, 50
- [30] Paul Rolland, Luca Viano, Norman Schürhoff, Boris Nikolov, and Volkan Cevher.
Identifiability and generalizability from multiple experts in inverse reinforcement learning.
arXiv preprint arXiv:2209.10974, 2022.
46

References VII

- [31] S. Ross, G. Gordon, and D. Bagnell.
A reduction of imitation learning and structured prediction to no-regret online learning.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.
28, 29, 31
- [32] Stéphane Ross and Drew Bagnell.
Efficient reductions for imitation learning.
In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 661–668. JMLR Workshop and Conference Proceedings, 2010.
16, 17, 21, 22, 28, 29
- [33] Stuart Russell.
Learning agents for uncertain environments (extended abstract).
In *Annual Conference on Computational Learning Theory (COLT)*, 1998.
34, 35
- [34] Stefan Schaal.
Is imitation learning the route to humanoid robots?, 1999.
9
- [35] Shai Shalev-Shwartz.
Online learning and online convex optimization.
Foundations and Trends® in Machine Learning, 4(2):107–194, 2012.
31

References VIII

- [36] Luca Viano, Angeliki Kamoutsis, Gergely Neu, Igor Krawczuk, and Volkan Cevher.
Proximal point imitation learning.
arXiv preprint arXiv:2209.10968, 2022.
70
- [37] Banghua Zhu, Jiantao Jiao, and Michael I Jordan.
Principled reinforcement learning with human feedback from pairwise or k -wise comparisons.
arXiv preprint arXiv:2301.11270, 2023.
8
- [38] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al.
Maximum entropy inverse reinforcement learning.
volume 8, pages 1433–1438. Chicago, IL, USA, 2008.
9, 34, 35, 52, 56
- [39] Martin Zinkevich.
Online convex programming and generalized infinitesimal gradient ascent.
In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.
30