Proximal point imitation learning.

lions@epfl

Luca Viano Angeliki Kamoutsi Gergely Neu Igor Krawzuck Volkan Cevher

ETH zürich upf. Universitat Pompeu Fabra Barcelona

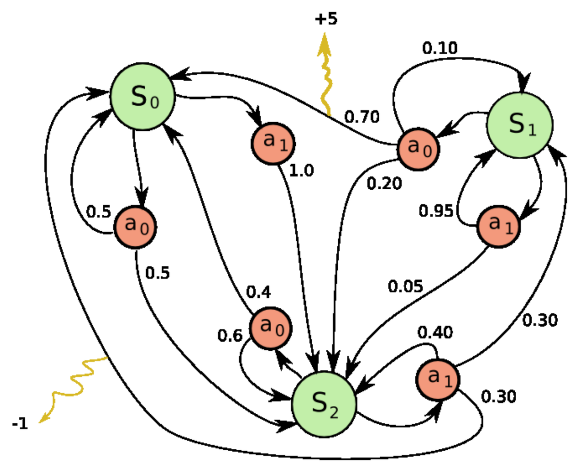
① What type of learning

Environment: MDP, unknown dynamics, unknown cost

INPUT: A finite set of expert demonstrations

GOAL: Learn a policy that performs at least as *good* as the expert

② Markov decision processes (MDPs)



- **Markov decision model** $\mathcal{M}_c \triangleq (\mathcal{X}, \mathcal{A}, P, \gamma, \nu_0, c)$
- $P(x'|x,a) = \text{Prob}\,(x_{t+1} = x' | x_t = x, a_t = a)$,
- Π_0 set of stationary Markov policies π, $\pi(a|x) = \text{Prob}\,(a_t = a | x_t = x)$,
- $\nu_0 \in \Delta_{\mathcal{S}}$ initial state distribution, $c \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ cost function, $\gamma \in (0,1)$ discount factor.
- $a_t \sim \pi(\cdot|x_t)$; $x_{t+1} \sim P(\cdot|x_t, a_t)$; $c(x_t, a_t)$
- Occupancy measure μ_π induced by a policy

$$\mu_\pi(x,a) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t P(x_t = x, a_t = a | x_0 \sim \nu_0, \pi)$$

- **Minimize a cost criterion**

$$\min_{\pi \in \Pi} \rho_c(\pi), \quad \text{where} \quad \rho_c(\pi) \triangleq (1-\gamma)\mathbb{E}^{\pi}_{\nu_0}\left[\sum_{t=0}^{\infty} \gamma^t c(x_t, a_t)\right].$$

- **LP formulation**

$$\min_{\mu \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \max_{c \in \mathcal{C}} \langle \mu - \mu_{\pi_E}, c \rangle \qquad \text{(Primal IL)}$$
$$\text{s.t.} \quad E^\intercal \mu = (1-\gamma)\nu_0 + \gamma P^\intercal \mu, \quad \mu \geq 0.$$

- **Linear MDP assumption** There exists mappings $\boldsymbol{\phi} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^m$ and $g : \mathcal{X} \to \mathbb{R}^m$ and a vector $w \in \mathcal{W} := \{w \in \mathbb{R}^m : \|w\|_2 \leq 1\}$ such that

$$c(s,a) = \langle \boldsymbol{\phi}(s,a), w \rangle \quad P(s'|s,a) = \langle \boldsymbol{\phi}(s,a), g(s') \rangle$$

that is, in matrix form

$$c = \Phi w \quad P = \Phi M$$

③ The constraint splitting trick

- We plug in the (Linear MDP) structure in (Primal IL) as follows. A similar trick appeared outside the imitation learning in in (Mehta and Meyn, 2020), (Lee and He, 2019) and (Bas-Serrano et al., 2021).

$$\min_{\mu \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \max_{w \in \mathcal{W}} \langle \mu - \mu_{\pi_E}, \Phi w \rangle \quad \text{s.t.} \quad E^\intercal \mu = (1-\gamma)\nu_0 + \gamma M^\intercal \Phi^\intercal \mu \quad \Rightarrow \quad \min_{\lambda \in \Delta^m, \mu \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}} \max_{w \in \mathcal{W}} \langle \lambda - \Phi^T \mu^{\pi_E}, w \rangle$$
$$\text{s.t.} \quad E^T \mu - \gamma M^T \lambda = (1-\gamma)\nu_0$$
$$\Phi^T \mu = \lambda \qquad \text{(Primal}' \text{ IL)}$$

- Our algorithm consists in applying inexact proximal point updates for μ and λ on the Lagrangian of Primal′ IL.

④ The algorithm

Proximal Point Imitation Learning: P^2IL

Initialize π_0 as uniform distribution over $\mathcal{A}$

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

`Policy evaluation` :

$$(w_k, \theta_k) \approx \text{argmax}_{w \in \mathcal{W}, \theta \in \Theta} \mathcal{G}_k(w, \theta)$$

`Policy improvement` :

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

Where $\mathcal{G}_k(w,\theta)$ is called (negative) logistic Bellman error (Bas-Serrano et al., 2021) and it is the following concave and smooth function. [a]

$$-\frac{1}{\eta}\log\sum_{i=1}^{m}(\Phi^T \mu_{k-1})(i) e^{-\eta \delta^k_{w,\theta}(i)} + (1-\gamma)\langle \nu_0, V^k_{\boldsymbol{\theta}} \rangle - \langle \lambda_{\pi_E}, \Phi^T w \rangle,$$

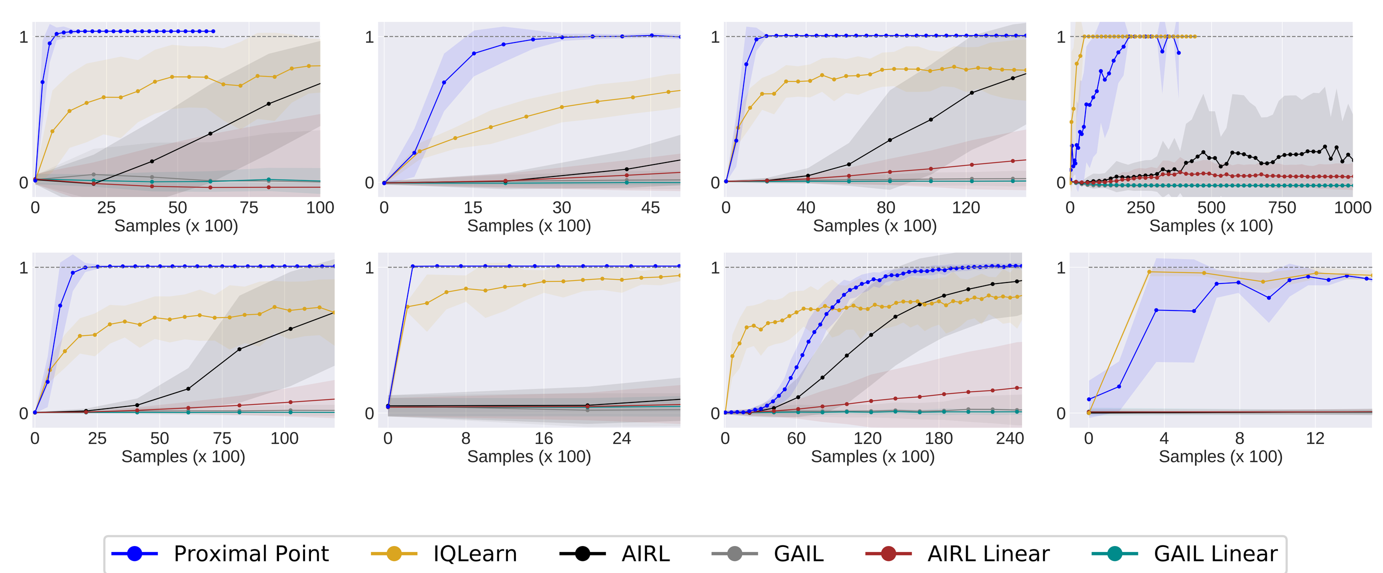
[a] We use $\delta^k_{w,\theta} \triangleq w + \gamma M V^k_\theta - \theta$ and $V^k_\theta \triangleq -\frac{1}{\alpha}\log\left(\sum_a \pi_{\mu_{k-1}}(a|s) e^{-\alpha Q_\theta(s,a)}\right)$ and $Q_\theta = \Phi\theta$.

Theorem 1 (Resources Guarantees) *Let us define the $\mathcal{C}$-distance between π and π_{E}, $d_{\mathcal{C}}(\pi, \pi_{\text{E}}) \triangleq \max_{\mathbf{c} \in \mathcal{C}} (\rho_{\mathbf{c}}(\pi) - \rho_{\mathbf{c}}(\pi_{\text{E}}))$. Using $\Omega(KT) = \Omega(\epsilon^{-5})$ sample transitions, $\Omega(\varepsilon^{-2})$ expert trajectories and approximately solving $\Omega(\epsilon^{-1})$ concave maximization problems, we can ensure $d_{\mathcal{C}}(\widehat{\pi}, \pi_{\text{E}}) \leq \mathcal{O}(\epsilon + \varepsilon)$, with high probability.*

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

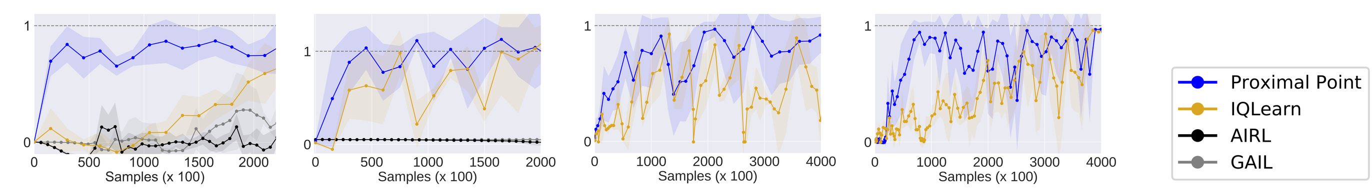
⑤ Discrete Actions Experiments

- From the left to the right: WideTree, RiverSwim, SingleChain, DoubleChain, Cartpole, Two State, Gridworld and Acrobot.



⑥ Continuous Control Experiments

- From the left to the right: HalfCheetah, Ant, Hopper, Walker .



- Nonlinear function approximation and continuous actions are not covered by our theory.
- However the empirical performance is convincing vs the state-of-the-art IQLearn (Garg et al., 2021).

Full paper

HASLERSTIFTUNG SDSC swisscom ZEISS FNSNF Fonds National Suisse Schweizerischer Nationalfonds Fondo Nazionale Svizzero Swiss National Science Foundation E4S Enterprise for Society erc EPFL