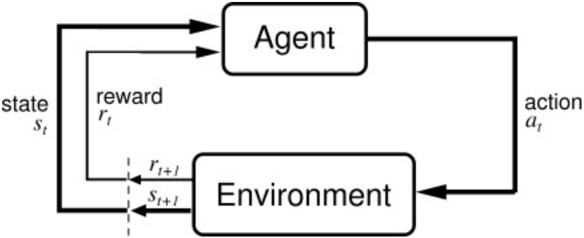
End-to-End Differentiable Adversarial Imitation Learning

Shie Mannor
EE @ Technion

Joint work with Nir Baram, Oron Anschel, Itai Caspi
May 9th, ICRI

Reinforcement Learning
(Agent interacts with an environment)



Applications:

1. Games (Go, Chess, Atari)
2. Robotics (self-driving cars)
3. Power grids, healthcare, ...

Reinforcement Learning
(Different approaches)

Value-based methods

$$Q^\pi(s,a) = \mathbb{E}\Big[\sum_{t=0}^{N} \gamma^t r_t | \pi, s, a\Big]$$

$$\pi(s) \leftarrow \operatorname*{argmax}_a Q(s,a)$$

Policy-based Gradient

$$J(\theta) = \mathbb{E}\Big[\sum_{t=0}^{N} \gamma^t r_t | \pi_\theta\Big]$$

Gradient Descent: $\frac{\partial J}{\partial \theta}$

Imitation Learning

$$D = \Big\{(s_0, a_0), (s_1, a_1), \ldots\Big\}$$

ERM: $\mathcal{F} : s \rightarrow a$

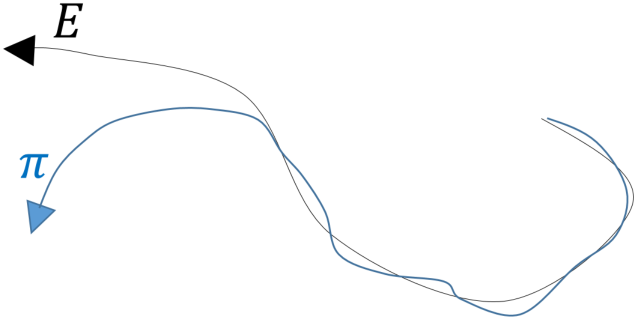
Why is imitation learning useful?

- Imitating is easy!
- Alleviate the temporal-credit assignment problem
- Learn complex behaviour quickly (dancing!)
- Combine simulator and data

But, need to use stochastic policies.

Imitation Learning as a Supervised Problem
(Behavioral Cloning)

$$p_E(s) \neq p_\pi(s)$$



Impossible to avoid a stochastic policy.

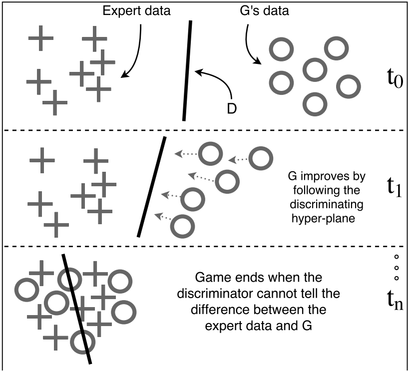
Imitation Learning as a Reinforcement Problem

(Composition of two difficult problems)

$$D = \left\{ (s_0, a_0), (s_1, a_1), \ldots \right\} \xrightarrow[\text{(Difficult and ill-posed problem)}]{\text{Inverse RL}} r(s, a) \xrightarrow{\text{RL}} \pi$$

Imitation Learning using a GAN

(What are GANs?)



Imitation Learning using GANs

(**G**enerative **A**dversarial **I**mitation **L**earning)

GAN

$\text{argmin}_G \ \text{argmax}_{D \in (0,1)} \quad \mathbb{E}_{x \backsim p_E}[\log D(x)] + \mathbb{E}_{z \backsim p_z}\left[\log\left(1 - D(G(z))\right)\right]$

A game between G (generator) and D (judge)

GAIL

$\text{argmin}_\pi \ \text{argmax}_{D \in (0,1)} \ \mathbb{E}_\pi[\log D(s,a)] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))] - \lambda H(\pi)$

A game between π (policy) and D (judge)

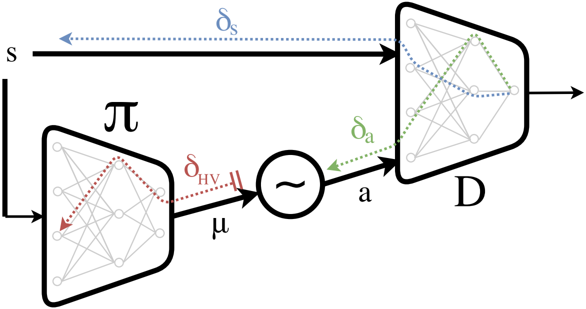
Main technical issue:

$\mathbb{E}_\pi[\log D(s,a)] = \mathbb{E}_{s \sim \rho(\pi)} \mathbb{E}_{a \sim \pi(\cdot|s)}[\log D(s,a)]$.

How to differentiate w.r.t. (parameters of) π?

Imitation Learning using GAN: Model free approach

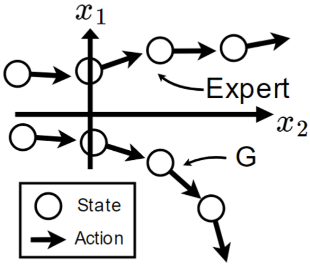
(The problem of training stochastic policies)

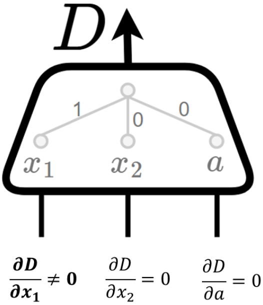


Imitation Learning using GANs: Model free approach

(The importance of $\frac{\partial D}{\partial s}$)

Perfect discrimination based on x_1!





Imitation Learning using GAN

(The role of D in imitation problems)

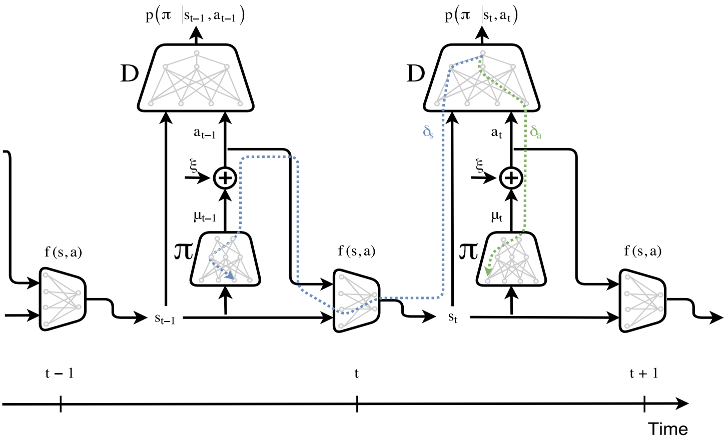
$D(s,a) = p(\pi|s,a)$ where $\pi \in \{\pi_E, \pi\}$.
$D(s,a)$ represents the likelihood ratio that the pair (s,a) is generated by π rather than by π_E.

Can show:
Policy likelihood ratio: $\varphi(s,a) = \frac{p(a|s,\pi_E)}{p(a|s,\pi)}$,
State distribution likelihood ratio: $\psi(s) = \frac{p(s|\pi_E)}{p(s|\pi)}$,
$D(s,a) = \frac{1}{1+\varphi(s,a)\cdot\psi(s)}$

Conclusion: Can easily compute $\bigtriangledown_a D$ and $\bigtriangledown_s D$.

model-based Imitation Learning using GAN

(How can we avoid throwing away $\frac{\partial D}{\partial s}, \frac{\partial D}{\partial a}$)



Main Conclusions

(The benefits of using a model)

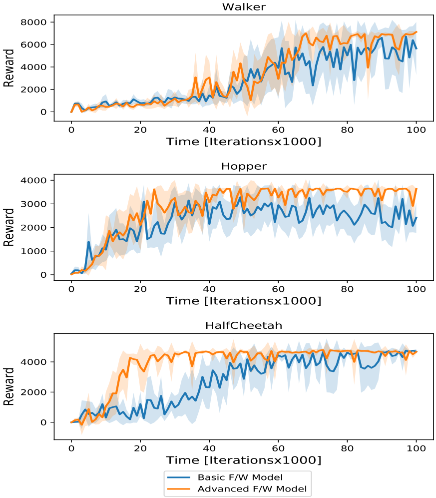
Model Free (GAIL)	Model Based (mGAIL)
un-biased high-variance gradient Discards $\frac{\partial D}{\partial s}$	biased low-variance gradient Uses $\frac{\partial D}{\partial s}$

Some of the domains

Movie

Results

(GAIL vs. mGAIL)



Results

(GAIL vs. mGAIL)

Task	Dataset size	Behavioral cloning	GAIL	MGAIL
Cartpole	1	72.02 ± 35.82	$\mathbf{200.00 \pm 0.00}$	$\mathbf{200.00 \pm 0.00}$
	4	169.18 ± 59.18	$\mathbf{200.00 \pm 0.00}$	$\mathbf{200.00 \pm 0.00}$
	7	188.60 ± 29.61	$\mathbf{200.00 \pm 0.00}$	$\mathbf{200.00 \pm 0.00}$
	10	177.19 ± 52.83	$\mathbf{200.00 \pm 0.00}$	$\mathbf{200.00 \pm 0.00}$
Mountain Car	1	-136.76 ± 34.44	$\mathbf{-101.55 \pm 10.32}$	$\mathbf{-107.4 \pm 10.89}$
	4	-133.25 ± 29.97	$\mathbf{-101.35 \pm 10.63}$	$\mathbf{-100.23 \pm 11.52}$
	7	-127.34 ± 29.15	$\mathbf{-99.90 \pm 7.97}$	$\mathbf{-104.23 \pm 14.31}$
	10	-123.14 ± 28.26	$\mathbf{-100.83 \pm 11.40}$	$\mathbf{-99.25 \pm 8.74}$
Acrobot	1	-130.60 ± 55.08	$\mathbf{-77.26 \pm 18.03}$	$\mathbf{-85.65 \pm 23.74}$
	4	-93.20 ± 35.58	$\mathbf{-83.12 \pm 23.31}$	$\mathbf{-81.91 \pm 17.41}$
	7	-96.92 ± 34.51	$\mathbf{-82.56 \pm 20.95}$	$\mathbf{-80.74 \pm 14.02}$
	10	-95.09 ± 33.33	$\mathbf{-78.91 \pm 15.76}$	$\mathbf{-77.93 \pm 14.78}$
Hopper	4	50.57 ± 0.95	3614.22 ± 7.17	$\mathbf{3669.53 \pm 6.09}$
	11	1025.84 ± 266.86	3615.00 ± 4.32	$\mathbf{3649.98 \pm 12.36}$
	18	1949.09 ± 500.61	3600.70 ± 4.24	$\mathbf{3661.78 \pm 11.52}$
	25	3383.96 ± 657.61	3560.85 ± 3.09	$\mathbf{3673.41 \pm 7.73}$
Walker	4	32.18 ± 1.25	4877.98 ± 2848.37	$\mathbf{6916.34 \pm 115.20}$
	11	5946.81 ± 1733.73	6850.27 ± 91.48	$\mathbf{7197.63 \pm 38.34}$
	18	1263.82 ± 1347.74	6964.68 ± 46.30	$\mathbf{7128.87 \pm 141.98}$
	25	1599.36 ± 1456.59	6832.01 ± 254.64	$\mathbf{7070.45 \pm 30.68}$
Half-Cheetah	4	-493.62 ± 246.58	4515.70 ± 549.49	$\mathbf{4891.56 \pm 654.43}$
	11	637.57 ± 1708.10	4280.65 ± 1119.93	$\mathbf{4844.61 \pm 138.78}$
	18	2705.01 ± 2273.00	4749.43 ± 149.04	$\mathbf{4876.34 \pm 85.74}$
	25	3718.58 ± 1856.22	4840.07 ± 95.36	$\mathbf{4989.95 \pm 351.14}$
Ant	4	1611.75 ± 359.54	3186.80 ± 903.57	$\mathbf{4645.12 \pm 179.29}$
	11	3065.59 ± 635.19	3306.67 ± 988.39	$\mathbf{4657.92 \pm 94.27}$
	18	2579.22 ± 1366.57	3033.87 ± 1460.96	$\mathbf{4664.44 \pm 183.11}$
	25	3235.73 ± 1186.38	4132.90 ± 878.67	$\mathbf{4637.52 \pm 45.66}$
Humanoid	80	1397.06 ± 1057.84	10200.73 ± 1324.47	$\mathbf{10312.34 \pm 388.54}$
	160	3655.14 ± 3714.28	10119.80 ± 1254.73	$\mathbf{10428.39 \pm 46.12}$
	240	5660.53 ± 3600.70	10361.94 ± 61.28	$\mathbf{10470.94 \pm 54.35}$

End-to-End Differentiable Adversarial Imitation Learning

- Imitation learning is the key to practical RL in many domains where expert trajectories are available
- A new architecture that allows differentiable backprop
- Main issue: the need to learn a forward model
- Transfer learning and multi-task are next