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, ...

End-to-End Differentiable Adversarial Imitation Learning
Reinforcement Learning

(Different approaches)

**Value-based methods**

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

\[
\pi(s) \leftarrow \arg\max_a Q(s, a)
\]

**Policy-based Gradient**

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

Gradient Descent:

\[
\frac{\partial J}{\partial \theta}
\]

**Imitation Learning**

\[
D = \left\{ (s_0, a_0), (s_1, a_1), \ldots \right\}
\]

ERM: \( \mathcal{F} : s \to a \)
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)

\[ D = \{(s_0, a_0), (s_1, a_1), \ldots \} \]

Inverse RL
(Difficult and ill-posed problem)

\[ r(s, a) \]

RL

\[ \pi \]
Imitation Learning using a GAN
(What are GANs?)
Imitation Learning using GANs

(Generative Adversarial Imitation Learning)

GAN

\[
\arg\min_G \arg\max_D \mathbb{E}_{x \sim p_E} [\log D(x)] + \\
\mathbb{E}_{z \sim p_z} [\log (1 - D(G(z)))]
\]

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

GAIL

\[
\arg\min_{\pi} \arg\max_D \mathbb{E}_{\pi} [\log D(s, a)] + \\
\mathbb{E}_{\pi E} [\log(1 - D(s, a))] - \lambda H(\pi)
\]

A game between \( \pi \) (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) \( \pi \)?
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 $\pi$ rather than by $\pi_E$.

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 $\nabla_a D$ and $\nabla_s D$. 
model-based Imitation Learning using GAN

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

Shie Mannor EE @ Technion
End-to-End Differentiable Adversarial Imitation Learning
Main Conclusions
(The benefits of using a model)

<table>
<thead>
<tr>
<th>Model Free (GAIL)</th>
<th>Model Based (mGAIL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>un-biased high-variance gradient</td>
<td>biased low-variance gradient</td>
</tr>
<tr>
<td>Discards $\frac{\partial D}{\partial s}$</td>
<td>Uses $\frac{\partial D}{\partial s}$</td>
</tr>
</tbody>
</table>
Some of the domains

Movie
Results

(GAIL vs. mGAIL)
## Results

(GAIL vs. mGAIL)

<table>
<thead>
<tr>
<th>Task</th>
<th>Dataset size</th>
<th>Behavioral cloning</th>
<th>GAIL</th>
<th>MGAIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cartpole</td>
<td>1</td>
<td>72.02 ± 35.82</td>
<td>200.00 ± 0.00</td>
<td>200.00 ± 0.00</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>169.18 ± 50.18</td>
<td>200.00 ± 0.00</td>
<td>200.00 ± 0.00</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>188.60 ± 29.61</td>
<td>200.00 ± 0.00</td>
<td>200.00 ± 0.00</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>177.19 ± 52.83</td>
<td>200.00 ± 0.00</td>
<td>200.00 ± 0.00</td>
</tr>
<tr>
<td>Mountain Car</td>
<td>1</td>
<td>−136.76 ± 34.44</td>
<td>−101.55 ± 10.32</td>
<td>−107.4 ± 10.89</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>−133.25 ± 29.97</td>
<td>−101.85 ± 10.63</td>
<td>−100.23 ± 11.52</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>−127.34 ± 29.15</td>
<td>−99.90 ± 7.97</td>
<td>−104.23 ± 14.31</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>−123.14 ± 28.26</td>
<td>−100.83 ± 11.40</td>
<td>−99.25 ± 8.74</td>
</tr>
<tr>
<td>Acrobat</td>
<td>1</td>
<td>−130.60 ± 55.08</td>
<td>−77.26 ± 18.03</td>
<td>−85.65 ± 23.74</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>−93.20 ± 35.58</td>
<td>−83.12 ± 23.31</td>
<td>−81.91 ± 17.41</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>−96.92 ± 34.51</td>
<td>−82.56 ± 20.95</td>
<td>−80.74 ± 14.02</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>−95.09 ± 33.33</td>
<td>−78.91 ± 15.76</td>
<td>−77.93 ± 14.78</td>
</tr>
<tr>
<td>Hopper</td>
<td>4</td>
<td>50.57 ± 0.95</td>
<td>3614.22 ± 7.17</td>
<td>3669.53 ± 6.09</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>1025.84 ± 266.86</td>
<td>3615.00 ± 4.32</td>
<td>3649.98 ± 12.36</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>1949.09 ± 500.61</td>
<td>3600.70 ± 4.24</td>
<td>3661.78 ± 11.52</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>3383.96 ± 657.61</td>
<td>3560.85 ± 3.09</td>
<td>3673.41 ± 7.73</td>
</tr>
<tr>
<td>Walker</td>
<td>4</td>
<td>32.18 ± 1.25</td>
<td>4877.98 ± 2848.37</td>
<td>6916.34 ± 115.20</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>5946.81 ± 1733.73</td>
<td>6850.27 ± 91.48</td>
<td>7197.63 ± 38.34</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>1263.82 ± 1347.74</td>
<td>6964.68 ± 46.30</td>
<td>7128.87 ± 141.98</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>1599.36 ± 1456.59</td>
<td>6832.01 ± 254.64</td>
<td>7070.45 ± 30.68</td>
</tr>
<tr>
<td>Half-Cheetah</td>
<td>4</td>
<td>−493.62 ± 246.58</td>
<td>4515.70 ± 549.49</td>
<td>4891.56 ± 654.43</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>637.57 ± 1708.10</td>
<td>4280.65 ± 1119.93</td>
<td>4844.61 ± 138.78</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>2705.01 ± 2273.00</td>
<td>4749.43 ± 149.04</td>
<td>4876.34 ± 85.74</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>3718.58 ± 1856.22</td>
<td>4840.07 ± 95.36</td>
<td>4989.95 ± 351.14</td>
</tr>
<tr>
<td>Ant</td>
<td>4</td>
<td>1611.75 ± 359.54</td>
<td>3186.80 ± 903.57</td>
<td>4645.12 ± 179.29</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>3065.59 ± 635.19</td>
<td>3306.67 ± 988.39</td>
<td>4657.92 ± 94.27</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>2579.22 ± 1366.57</td>
<td>3033.87 ± 1460.96</td>
<td>4664.44 ± 183.11</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>3235.73 ± 1186.38</td>
<td>4132.90 ± 878.67</td>
<td>4637.52 ± 45.66</td>
</tr>
<tr>
<td>Humanoid</td>
<td>80</td>
<td>1397.06 ± 1057.84</td>
<td>10200.73 ± 1324.47</td>
<td>10312.34 ± 388.54</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>3655.14 ± 3714.28</td>
<td>10119.80 ± 1254.73</td>
<td>10428.39 ± 46.12</td>
</tr>
<tr>
<td></td>
<td>240</td>
<td>5660.53 ± 3600.70</td>
<td>10361.94 ± 61.28</td>
<td>10470.94 ± 54.35</td>
</tr>
</tbody>
</table>
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