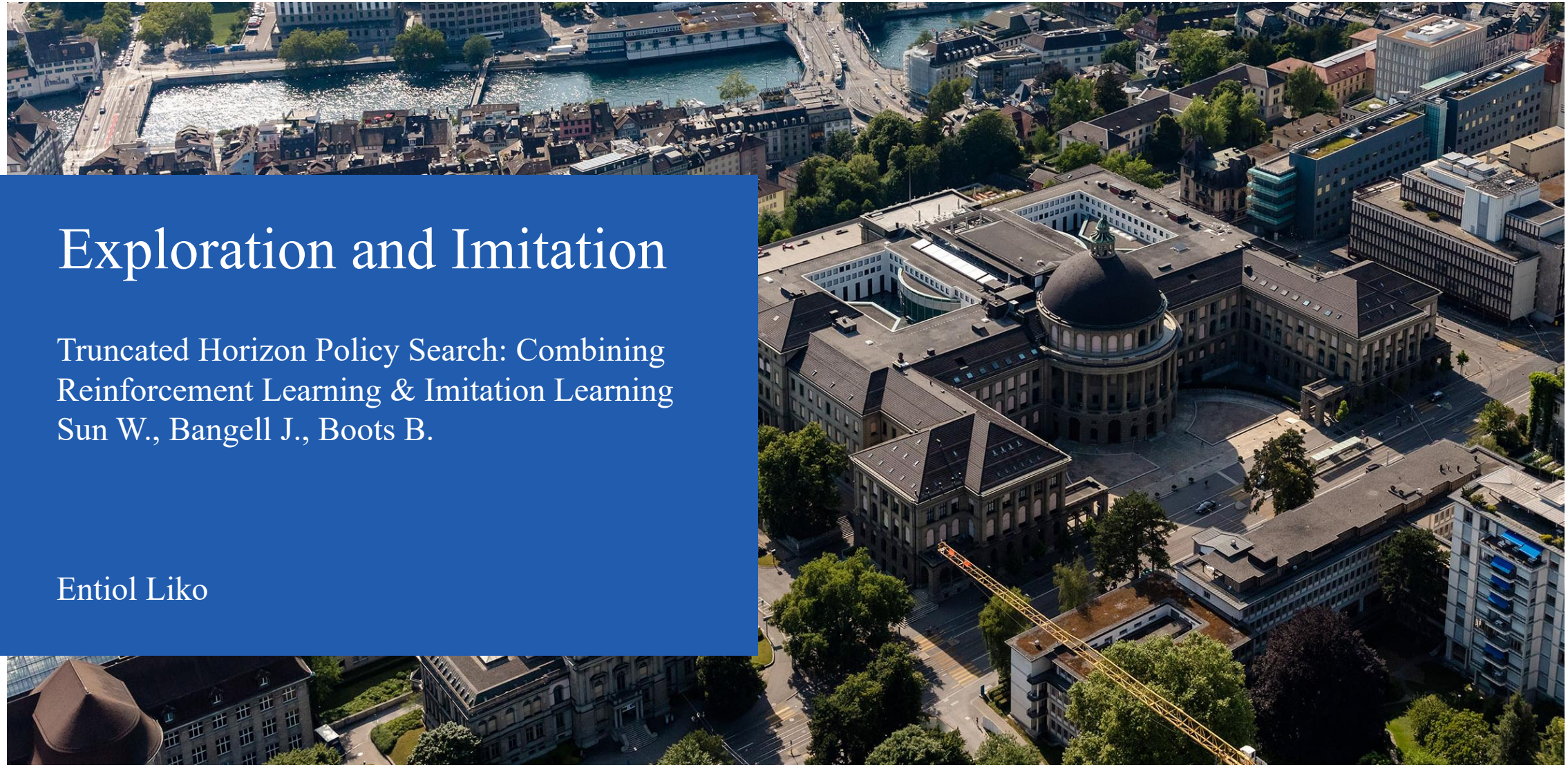
ETH zürich



Exploration and Imitation

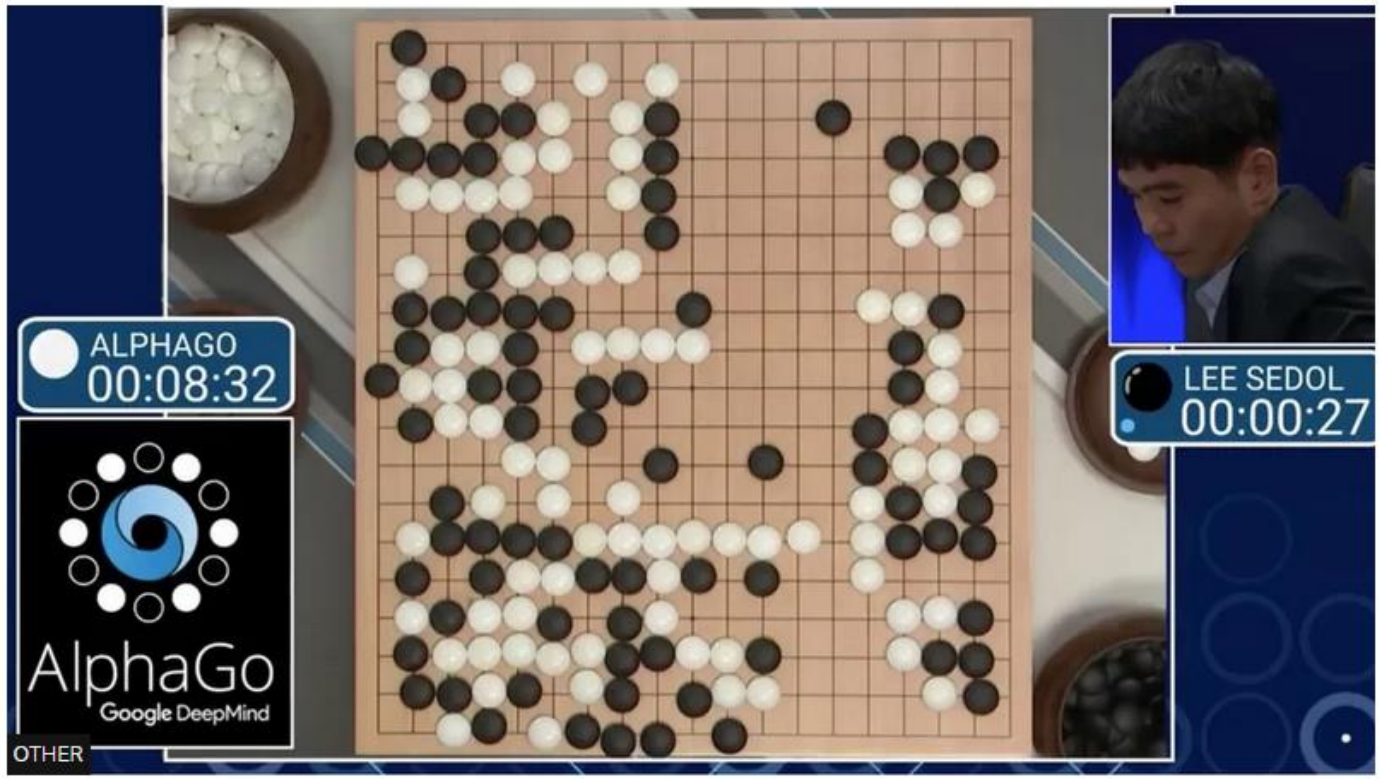
Truncated Horizon Policy Search: Combining Reinforcement Learning & Imitation Learning
Sun W., Bangell J., Boots B.

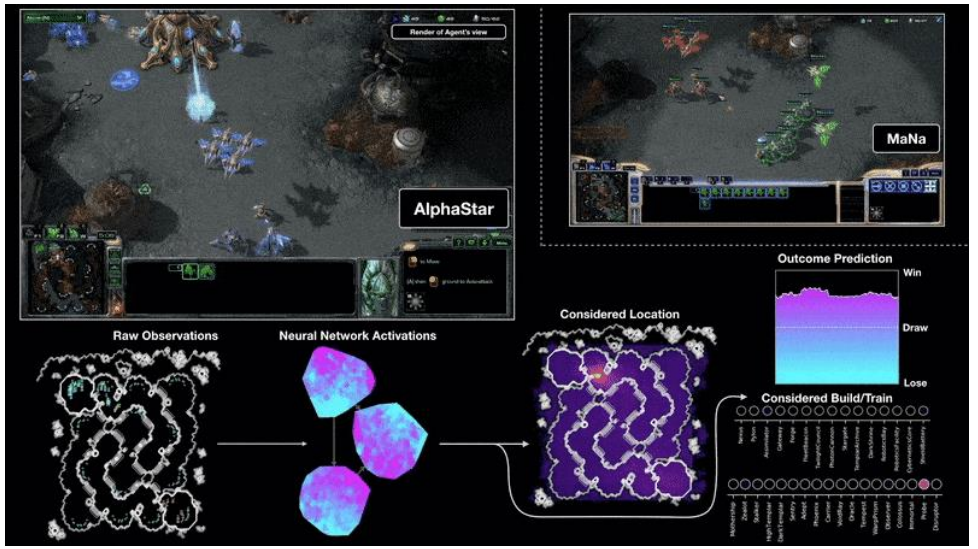
Entiol Liko

Reinforcement Learning

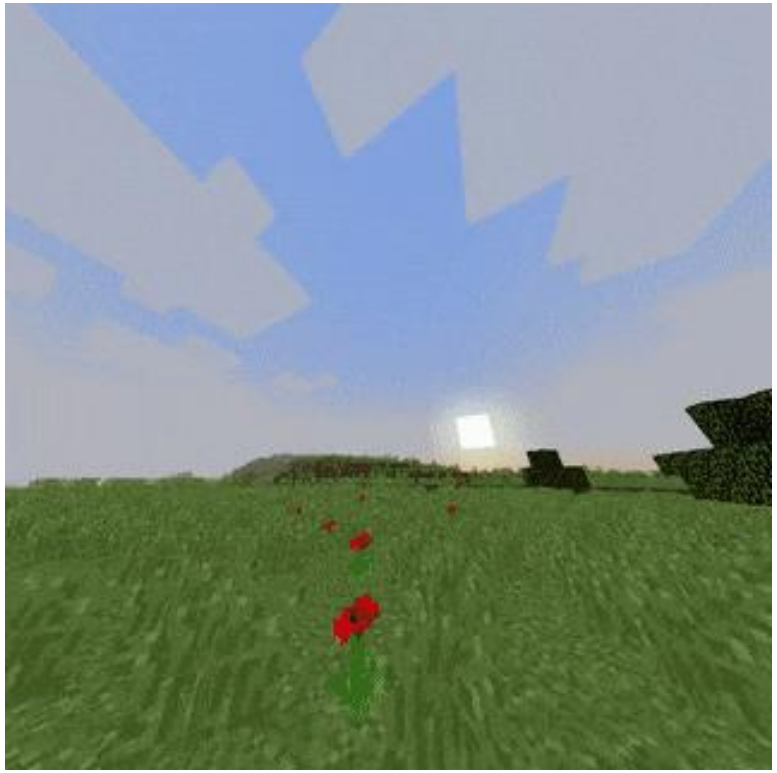
Artificial intelligence: Google's AlphaGo beats Go master Lee Se-dol

12 March 2016

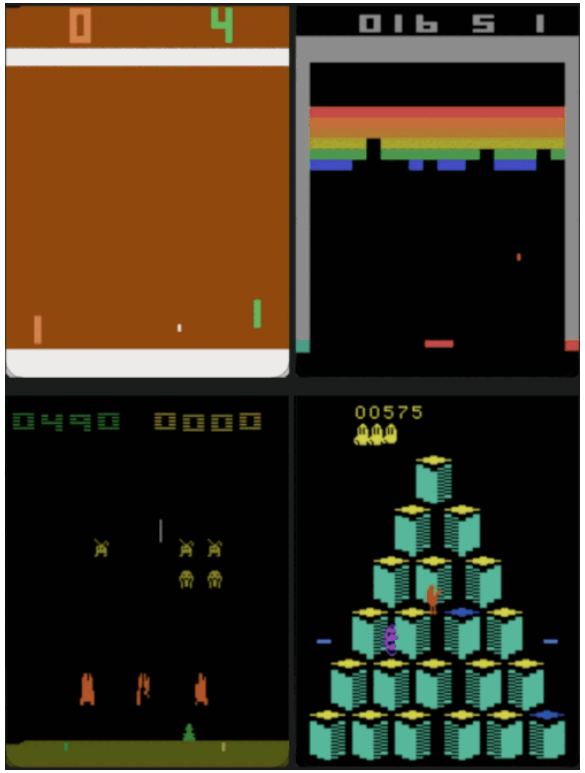



Render of Agent's view
AlphaStar
MaNa
Outcome Prediction
Win
Draw
Lose
Considered Location
Raw Observations
Neural Network Activations
Considered Build/Train

Imitation Learning



Artificial Intelligence (AI) playing Minecraft and exploring the world to find caves after being trained on only a few videos of human players, using our new approach (IQ-Learn). *Many approaches have difficulty doing anything in an open-world Minecraft map, but IQ-Learn can navigate the world with ease.*

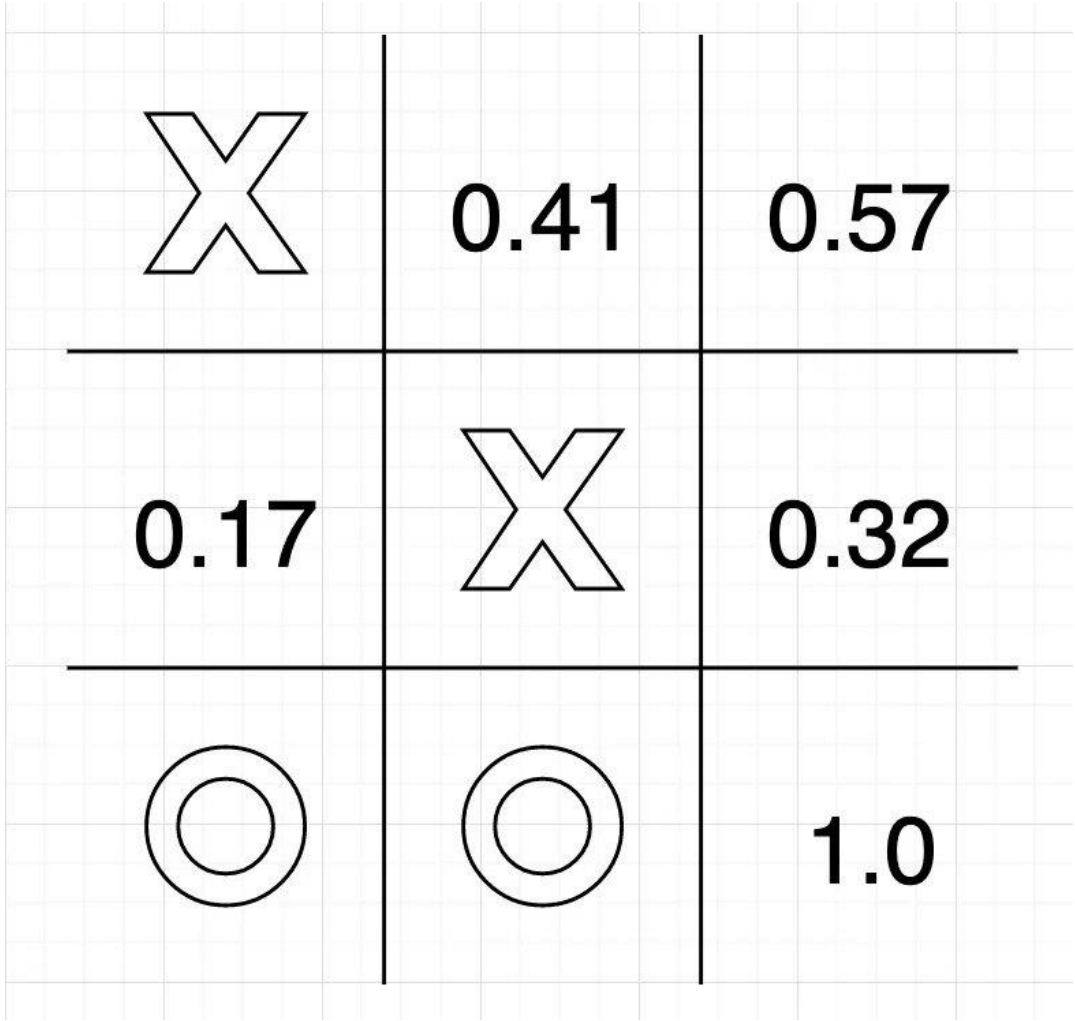


IQ-Learn on Atari. Our trained agent reaches human performance in all games (using 20 expert demos).

Imitation Learning

- Sun et al. (2017)[1] shows that, with access to an optimal expert, IL can **exponentially** lower sample complexity.
- The performance of our policy is **limited** by the performance of the expert, which is often sub-optimal.
- Data Aggregation(DAgger) (Ross et al., 2011)[2] and Aggregation with Values(AggreVaTe) (Ross & Bagnell, 2014)[3] **can only guarantee** a policy which performs **as good as** the expert policy.

The Value Function



$$V^{\pi}_{\mathcal{M}_0}(s) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) | s_0 = s, a \sim \pi\Big].$$

$$\pi^* = \arg\min_{\pi} V^{\pi}(s), \forall s \in \mathcal{S}.$$

Cost Reshaping

Given the original MDP M_0 and any potential function, we can reshape the cost as follows:

$$c'(s,a) = c(s,a) + \gamma\Phi(s') - \Phi(s), \quad s' \sim P_{sa}$$

We can use the cost-to-go oracle V^e as a potential function.

$$A^e(s,a) = c(s,a) + \gamma\mathbb{E}_{s'\sim P_{sa}}[V^e(s')] - V^e(s)$$

This means we can find the optimal policy as the following optimization problem:

$$\pi^* = \arg\min_{\pi} \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t A^e(s_t,a_t)|s_0 = s, a \sim \pi].$$

Main Idea:

Combine IL and RL through the idea of **Reward Shaping** (Ng et al., 1999)[5].

The cost-to-go **oracle** can serve as a potential function for cost shaping.

Previous work:

Chang et al (2015)[4] attempted to combine IL and RL by stochastically interleaving incremental RL and IL updates.

Ng (2003)[17] considers the setting where the potential function is close to the optimal value function

Contributions:

Attemps to unify IL and RL by varying the planning horizon from 1 to infinity, based on how close the oracle is to the optimal value function.

A lower bound analysis of AGGREVATE with an imperfect oracle, which is missing in Ross & Bangell (2014)[3]

Suggests a way to understand previous IL appraches throught reward shaping.

A model-free, actor-critic style algorithm that can be used for continuous state and action spaces.

Main Results

Theorem 3.1. *There exists an MDP and an imperfect oracle* $\hat{V}^e(s)$ *with* $|\hat{V}^e(s) - V^*_{\mathcal{M}_0,h}(s)| = \epsilon$, *such that the performance of the induced policy from the cost-to-go oracle* $\hat{\pi}^* = \arg\min_a \left[c(s,a) + \gamma \mathbb{E}_{s' \sim P_{sa}}[\hat{V}^e(s')]\right]$ *is at least* $\Omega(\gamma\epsilon/(1-\gamma))$ *away from the optimal policy* π^**:*

$$J(\hat{\pi}^*) - J(\pi^*) \geq \Omega\left(\frac{\gamma}{1-\gamma}\epsilon\right). \tag{3}$$

$$\mathbb{E}[c'(s,a)] = [c(s,a) + \gamma \mathbb{E}_{s' \sim P_{sa}} \hat{V}^e(s') - \hat{V}^e(s)],$$

With the k steps we have:

$$\mathbb{E}\Big[\sum_{i=1}^{k}\gamma^{i-1}c'(s_i,a_i)|s_1=s;a\sim\pi\Big].$$

$$\mathbb{E}\Big[\sum_{i=1}^{k}\gamma^{i-1}c(s_i,a_i)+\gamma^k\hat{V}^e(s_{k+1})-\hat{V}^e(s_1)|s_1=s;a\sim\pi\Big],\forall s\in\mathcal{S}.$$

With the main result being:

Theorem 3.2. *Assume* $\hat{\pi}^*$ *minimizes Eq.* 6 *for every state* $s\in\mathcal{S}$ *with* $k>1$ *and* $|\hat{V}^e(s)-V^*(s)|=\Theta(\epsilon),\forall s$. *We have :*

$$J(\hat{\pi}^*)-J(\pi^*)\leq O\left(\frac{\gamma^k}{1-\gamma^k}\epsilon\right) \tag{7}$$

THOR

We start by defining the k-step truncated value function and action-value function:

$$V_{\mathcal{M}}^{\pi,k}(s) = \mathbb{E}\Big[\sum_{t=1}^{k} \gamma^{t-1} c'(s_t, a_t) | s_1 = s, a \sim \pi\Big],$$

$$Q_{\mathcal{M}}^{\pi,k}(s,a) = \mathbb{E}\Big[c'(s,a) + \sum_{i=1}^{k-1} \gamma^{i} c'(s_i, a_i) | s_i \sim P_{sa}, a_i \sim \pi(s_i)\Big],$$

$$\min_{\pi \in \Pi} \mathbb{E}_{s_0 \sim \nu}\left[\mathbb{E}\Big[\sum_{i=1}^{k} \gamma^{i} c'(s_i, a_i) | a \sim \pi\Big]\right].$$

We can use gradient-based uupdate procedures (Stochastic Gradient Descent, Natural Gradient (Kakade, 2002[6]; Bagnell & Schneider, 2003[7])) in the policy's parameter space.

THOR

After taking the derivative with respect to the policy's parameters we have:

$$\mathbb{E}_{S\sim P_{T_{\pi_n}}}\left[\mathbb{E}_{\tau^k\sim\pi_n}[\sum_{i=1}^{k}\nabla_\theta \ln\pi(a_i|s_i;\theta)(\sum_{j=i}^{k+i}\gamma^{j-i}c'(s_j,a_j))]\right]$$

$$\approx \mathbb{E}_{S\sim P_{T_{\pi_n}}}\left[\mathbb{E}_{\tau^k\sim\pi_n}[\sum_{i=1}^{k}\nabla_\theta \ln\pi(a_i|s_i;\theta)Q_{\mathcal{M}}^{\pi,k}(s_i,a_i)]\right]$$

By replacing the expectation by empirical samples from π_n, replacing $Q_M^{\pi,k}$ by a critic approximated by Generalized disadvantage Estimator (GAE) $\hat{A}_M^{\pi,k}$ (Schulman el al., 2016)[8], we have:

$$\mathbb{E}_{S\sim P_{T_{\pi_n}}}\left[\mathbb{E}_{\tau_k\sim\pi_n}[\sum_{i=1}^{k}\nabla_\theta \ln\pi(a_i|s_i;\theta)Q_{\mathcal{M}}^{\pi,k}(s_i,a_i)]\right]$$

$$\approx k\sum_{\tau}\left(\sum_{t=1}^{|\tau|}\nabla_\theta \ln(\pi(a_t|s_t;\theta))\hat{A}_{\mathcal{M}}^{\pi,k}(s_t,a_t)\right)/H,$$

THOR

Algorithm 1 Truncated Horizon Policy Search (THOR)

Input: The original MDP $\mathcal{M}_0$. Truncation Step k. Oracle V^e.
Initialize policy π_{θ_0} with parameter θ_0 and truncated advantage estimator $\hat{A}^{0,k}_{\mathcal{M}}$.
for n $= 0, \ldots$ **do**
  Reset system.
  Execute π_{θ_n} to generate a set of trajectories $\{\tau_i\}_{i=1}^N$.
  Reshape cost $c'(s_t, a_t) = c(s_t, a_t) + V^e_{t+1}(s_{t+1}) - V^e_t(s_t)$, for every $t \in [1, |\tau_i|]$ in every trajectory $\tau_i, i \in [N]$.
  Compute gradient:

$$\sum_{\tau_i} \sum_{t} \nabla_\theta (\ln \pi_\theta(a_t|s_t))|_{\theta=\theta_n} \hat{A}^{\pi_n,k}_{\mathcal{M}}(s_t, a_t) \tag{8}$$

  Update disadvantage estimator to $\hat{A}^{\pi_n,k}_{\mathcal{M}}$ using $\{\tau_i\}_i$ with reshaped cost c'.
  Update policy parameter to θ_{n+1}.
end for

Experiments

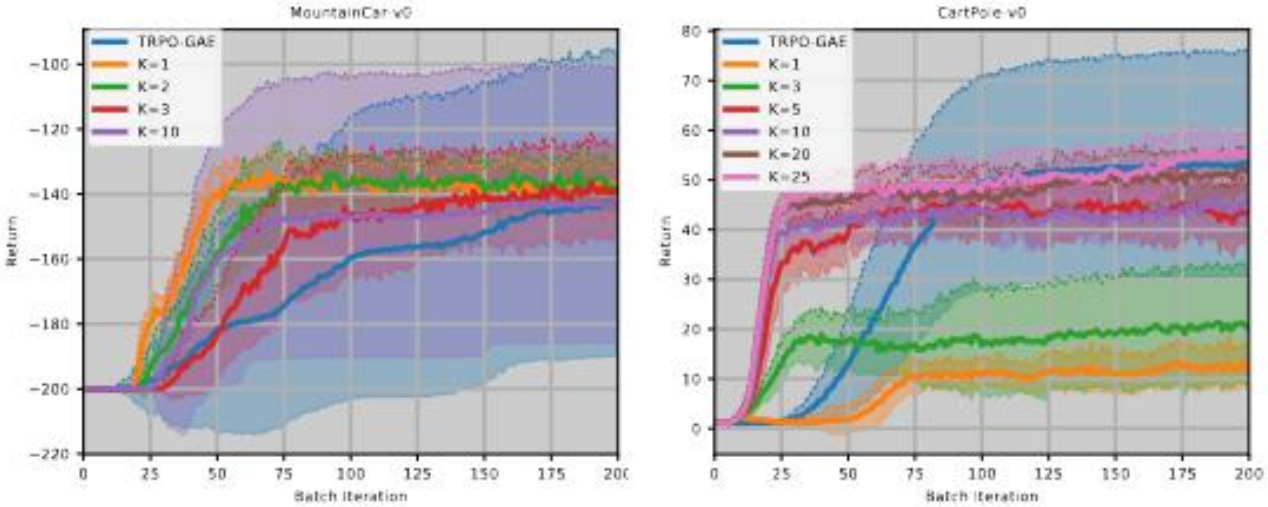
THOR was evaluated on robotics simulators from OpenAI Gym (Brockman et al., 2016) [9].

The baseline is TRPO-GAE (Schulmann et al., 2016) [8] and AggreVaTeD (Sum et al., 2017) [1].

For all methods the statistics (mean and std deviation) are from 25 seeds that are i.i.d. generated. k was the only tuned paramater

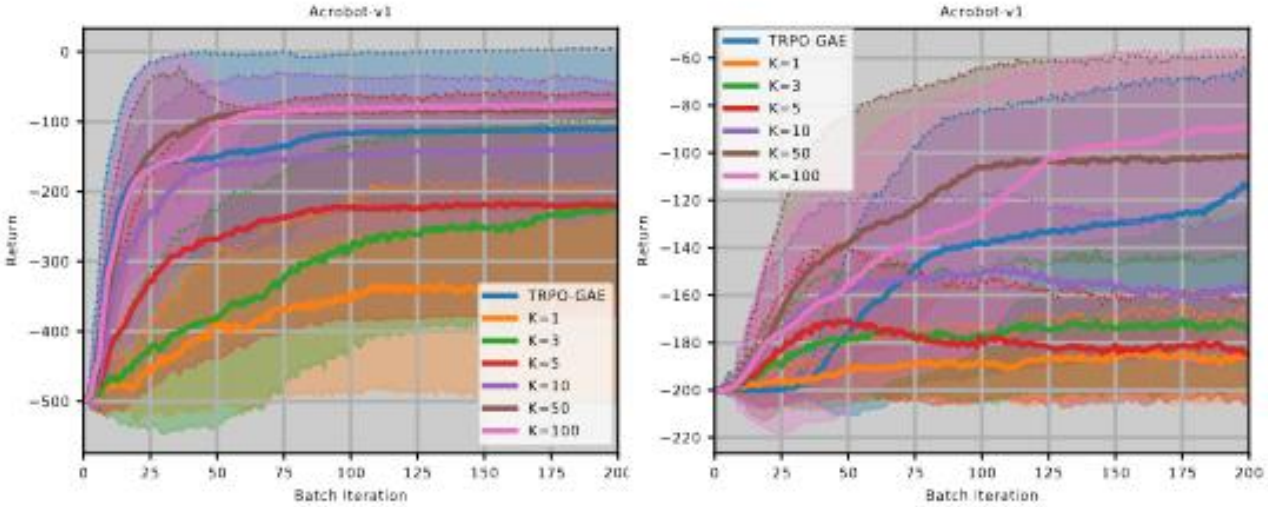
Discrete Action Control

All simulations have sparse rewards.



(a) Mountain Car with H = 200

(b) SR-CartPole with $H = 200$



(c) Acrobot with $H = 500$

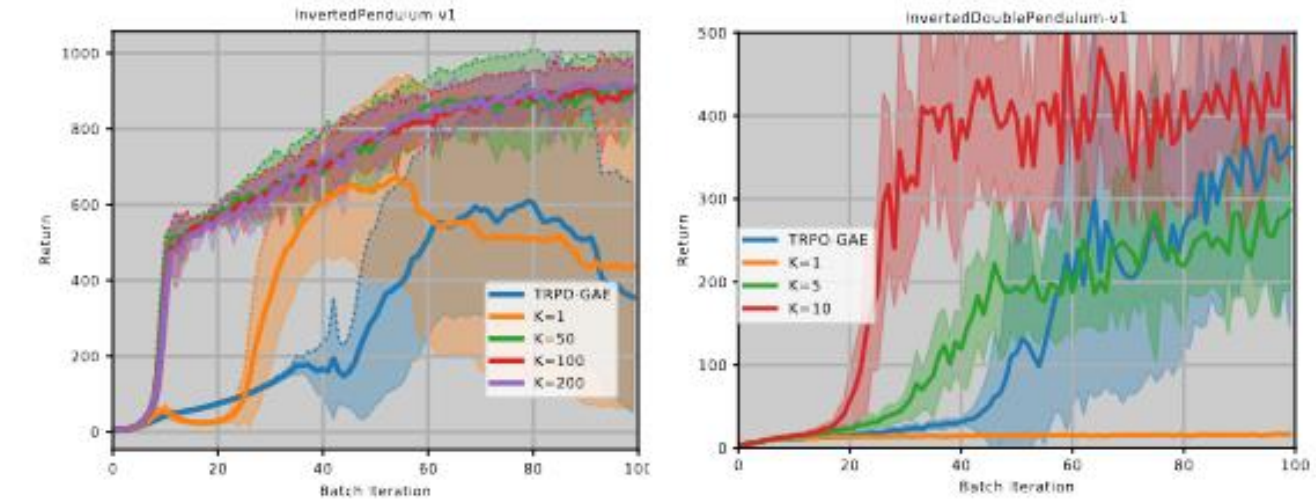
(d) Acrobot with $H = 200$

 Image taken from paper 10

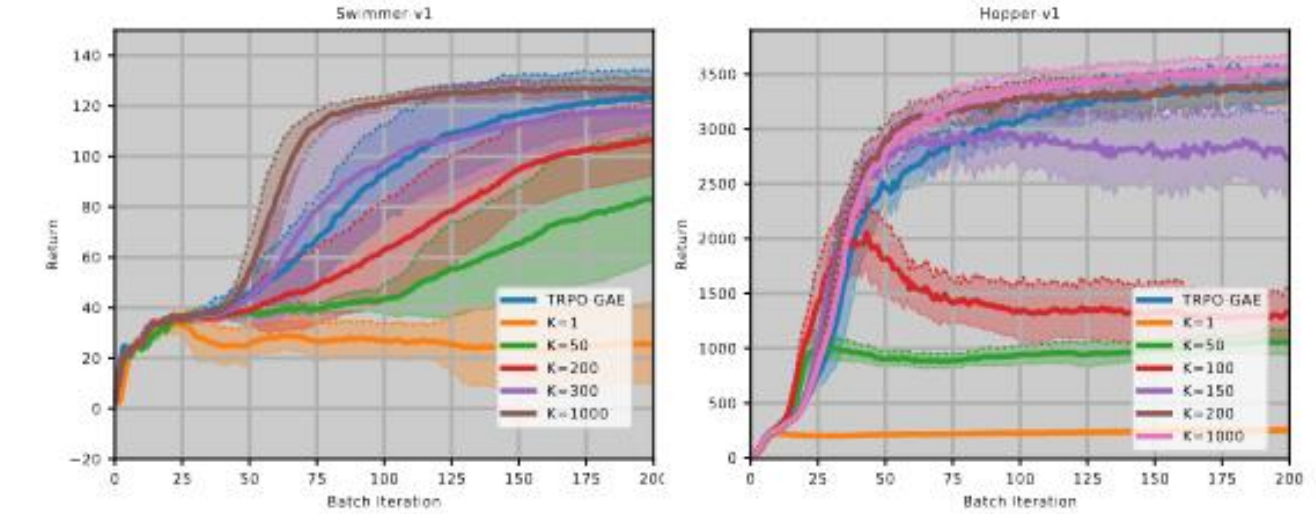
Continuous Action Control

Simulators with continuous state and action spaces from MuJoCo simulators.

Hopper and Swimmer do not have reward sparsity.



(a) SR-Inverted Pendulum (H=1000) (b) SR-Inverted Double Pendulum (H=1000)



(c) Swimmer (H=1000) (d) Hopper (H=1000)

 Image taken from paper 10

Conclusions

- A novel way to combine IL and RL
- Theoretical results on the connection between IL and RL
- Performance bound on AggreGaTe and THOR
- Strong performance

References

1. Wen Sun, Arun Venkatraman, Geoffrey J Gordon, Byron Boots, and J Andrew Bagnell. Deeply aggrevated: Differentiable imitation learning for sequential prediction. ICML, 2017.
2. St´ephane Ross, Geoffrey J Gordon, and J.Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In AISTATS, 2011.
3. Stephane Ross and J Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. arXiv preprint arXiv:1406.5979, 2014.
4. Kai-wei Chang, Akshay Krishnamurthy, Alekh Agarwal, Hal Daume, and John Langford. Learning to search better than your teacher. In ICML, 2015.
5. Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In ICML, volume 99, pp. 278–287, 1999.
6. Sham Kakade. A natural policy gradient. NIPS, 2002.
7. J Andrew Bagnell and Jeff Schneider. Covariant policy search. In IJCAI, 2003.
8. John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High- dimensional continuous control using generalized advantage estimation. ICLR, 2016.
9. Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.
10. Wen Sun, J. Andrew Bagnell, Byron Boots. Truncated Horizon Policy Search: Combining Reinforcement Learning & Imitation Learning, 2016

References

11. Volodymyr Mnih et al. Human-level control through deep reinforcement learning. Nature, 2015.
12. David Silver et al. Mastering the game of go with deep neural networks and tree search. Nature, 2016.
13. Matej Ve˘cer´ık, Todd Hester, Jonathan Scholz, Fumin Wang, Olivier Pietquin, Bilal Piot, Nicolas Heess, Thomas Roth¨orl, Thomas Lampe, and Martin Riedmiller. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. arXiv preprint arXiv:1707.08817, 2017.
14. Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. ICRA, 2018.
15. Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. IEEE, 2017.
16. Amir-massoud Farahmand, Daniel Nikolaev Nikovski, Yuji Igarashi, and Hiroki Konaka. Truncated approximate dynamic programming with task-dependent terminal value. In AAAI, pp. 3123–3129, 2016.
17. Andrew Y Ng. Shaping and policy search in reinforcement learning. PhD thesis, University of California, Berkeley, 2003.

ETH zürich

Thank you for you attention!