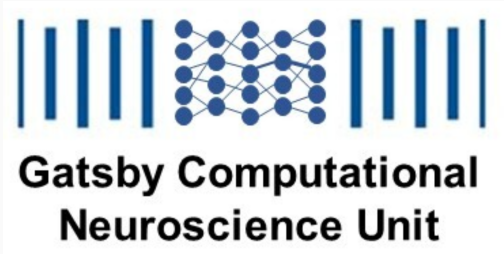
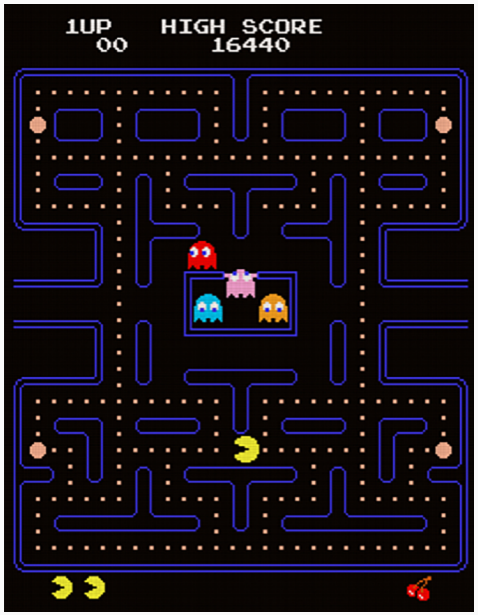
Tactical Optimism and Pessimism for Deep Reinforcement Learning

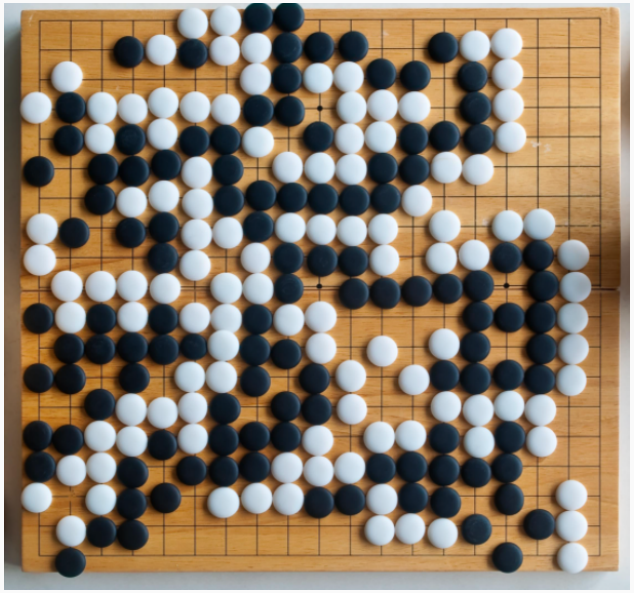
Ted Moskovitz, Jack Parker-Holder, Aldo Pacchiano,
Michael Arbel, Michael I. Jordan

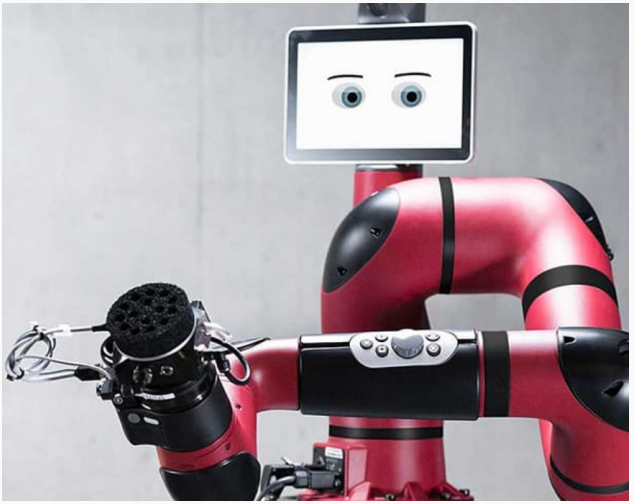
NeurIPS 2021



Motivation





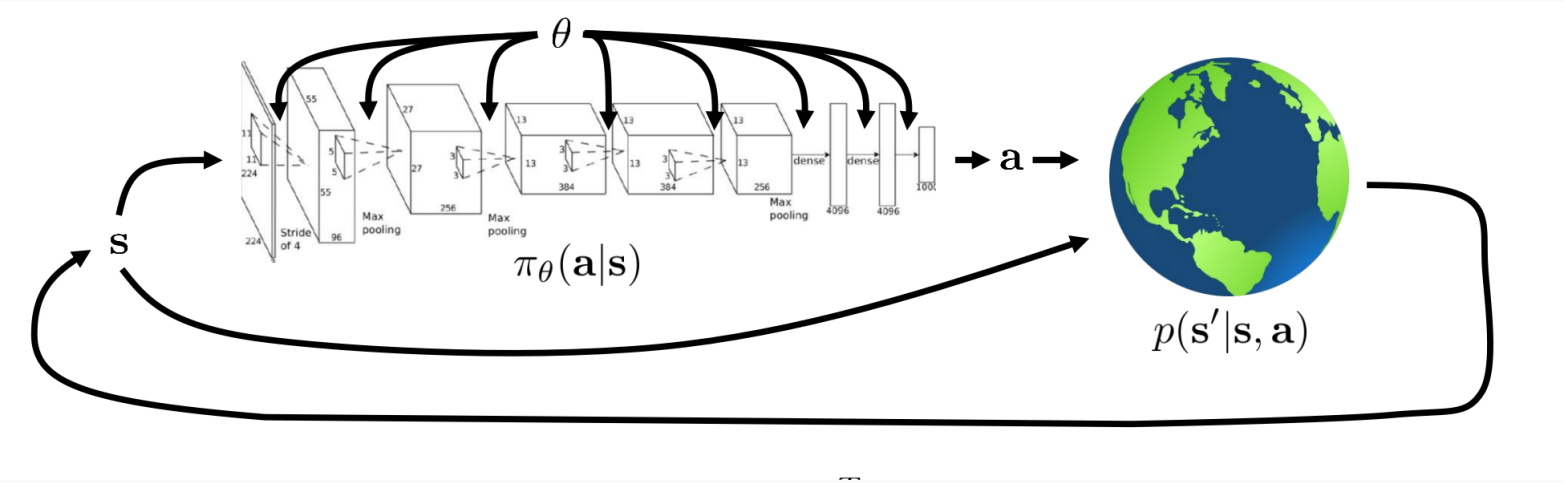


- Approximate value functions are key to the success of deep RL
- Optimism vs. pessimism: both have drawbacks + advantages
- How can we figure out which is best?

Summary + Contributions

- Demonstrate that the efficacy of optimism varies both across environments and over the course of training
- Introduce a novel framework for value estimation, *Tactical Optimism and Pessimism* (TOP)
- Adaptively updates its degree of optimism by modeling the choice as a multi-armed bandit problem
- Augmenting popular algorithms with TOP leads to state-of-the-art results

RL Background



- Assume an agent is acting in an MDP ($S; A; r; p;$)
- Running the policy in an episodic/finite horizon task of length T produces trajectories = $(s_1; a_1; r_1; : : : ; s_T; a_T; r_T)$
- Maximize $J(\ ) = E\ [Z(\ )] = E\ \sum_t {}^t r_t$

* figure adapted from http://rail.eecs.berkeley.edu/deeprlcourse/

The Actor-Critic Framework

- Actor, : a deterministic policy
- Critic, Q : evaluate actor,

$$Q \quad (s; a) := \mathrm{E} \quad [Z_t j s_t = s; a_t = s] \tag{0.1}$$

- How to train?
- Actor: $/ \quad r \quad J(\ ) = \mathrm{E} \quad r \ _a Q \quad (s; a) j_{a= \ (s)} r \quad (s)$
- Critic: given a transition $(s_t; a_t; r_t; s_{t+1})$,

$$/ \quad r \quad \frac{1}{2} k y_t \quad \underbrace{Q \quad (s_t; a_t)}_{:= \ t} k_2^2 \tag{0.2}$$

where $y_t = r_t + \quad Q \ _- (s_{t+1}; \quad (s_{t+1}))$

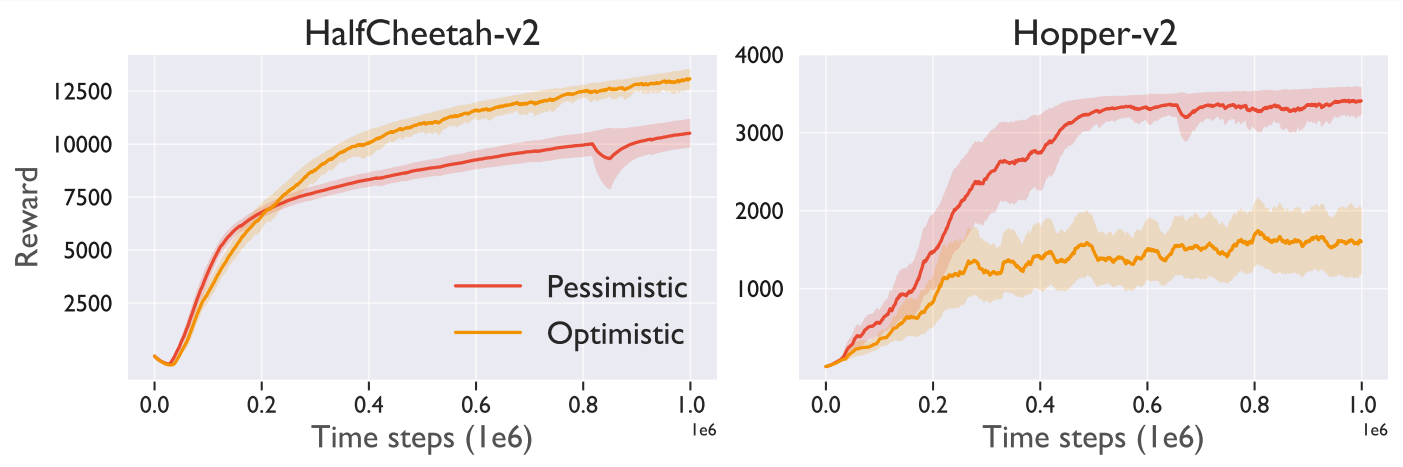
- Distributional RL: represent *distribution* of Z , not just mean [1]

Optimism vs. Pessimism

- Problem: Function approx. biases critic towards **overestimation** [5]
- Solution: build a **pessimistic** target using 2 critics [3]:

$$y_t = r_t + \min_{i \in \{1,2\}} Q_i(s_{t+1}; \quad (s) + \quad); \qquad \text{clip}(N(0; s^2); \quad c; c) \qquad (0.3)$$

- Problem: pessimistic critics can result in **underexploration** [2]
- Solution: use an **optimistic** upper-bound on the value [2]
- Confusing...



TOP: Tactical Optimism and Pessimism

- Distinguish between and represent two types of uncertainty:
- **Aleatoric uncertainty:** noise inherent to the world/task
-) represent using *distributional value estimation* Z Z $(s; a)$
- **Epistemic uncertainty:** noise due to lack of knowledge about the world
-) represent using an *ensemble* of $k = 1; : : : ; K$ critics
- We use these estimates to construct a *belief distribution* $\hat{Z}$ $(s; a)$:

$$q^{(k)}_{\hat{Z}\ (s;a)} = q^{(k)}_{Z\ (s;a)} + \quad q^{(k)}_{\hat{}(s;a)} \tag{0.4}$$

- 2 R then determines the degree of optimism

TOP: Tactical Optimism and Pessimism

- *Q: How to choose ?*
- A: Evaluate by its effect on performance!
- Model choice as a **bandit problem**:
  - Choose from $\{\beta_d\}_{d=1}^{D}$ by sampling a decision $d_m \in \{1, \ldots, D\}$ for episode m
  - $d_m \sim p_m(\cdot)$, where $p_m(d) \propto \exp(w_m(d))$
- Update arm weightings:

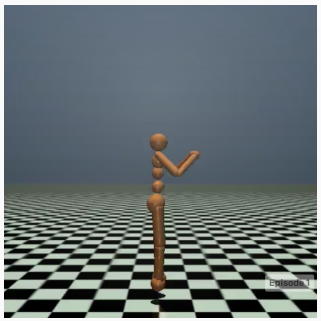
$$w_{m+1}(d) = \begin{cases} w_m(d) + \frac{f_m}{\mathbf{p}_m(d)} & \text{if } d = d_m \\ w_m(d) & \text{otherwise;} \end{cases} \tag{0.5}$$

- Feedback f_m is change in performance:

$$f_m = R_m \quad R_{m\ 1} \tag{0.6}$$

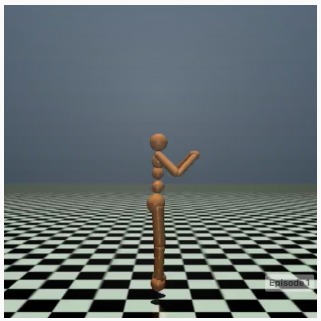
Experiments: State-based Control

TOP + TD3 [3]:



Experiments: State-based Control

TOP + TD3 [3]:



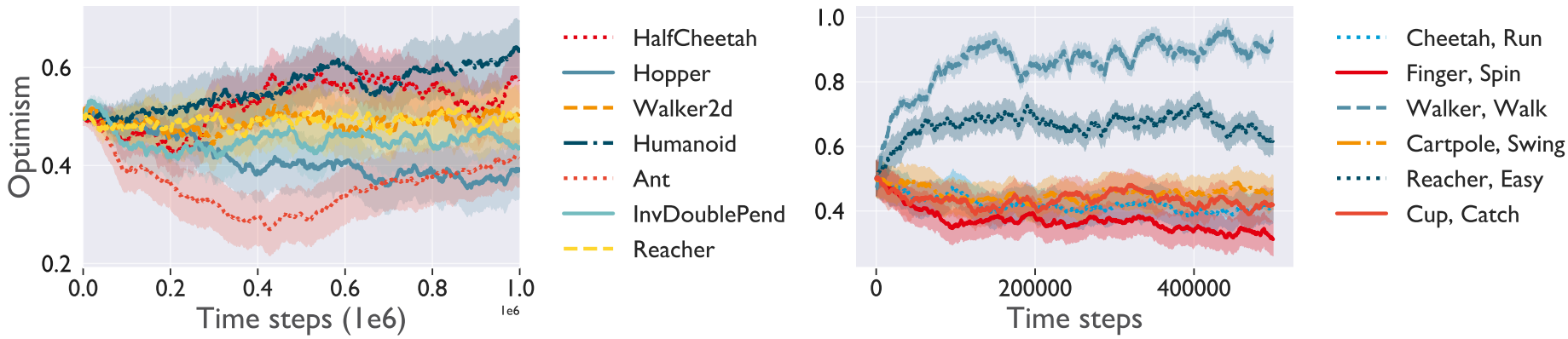
Experiments: Pixel-based Control

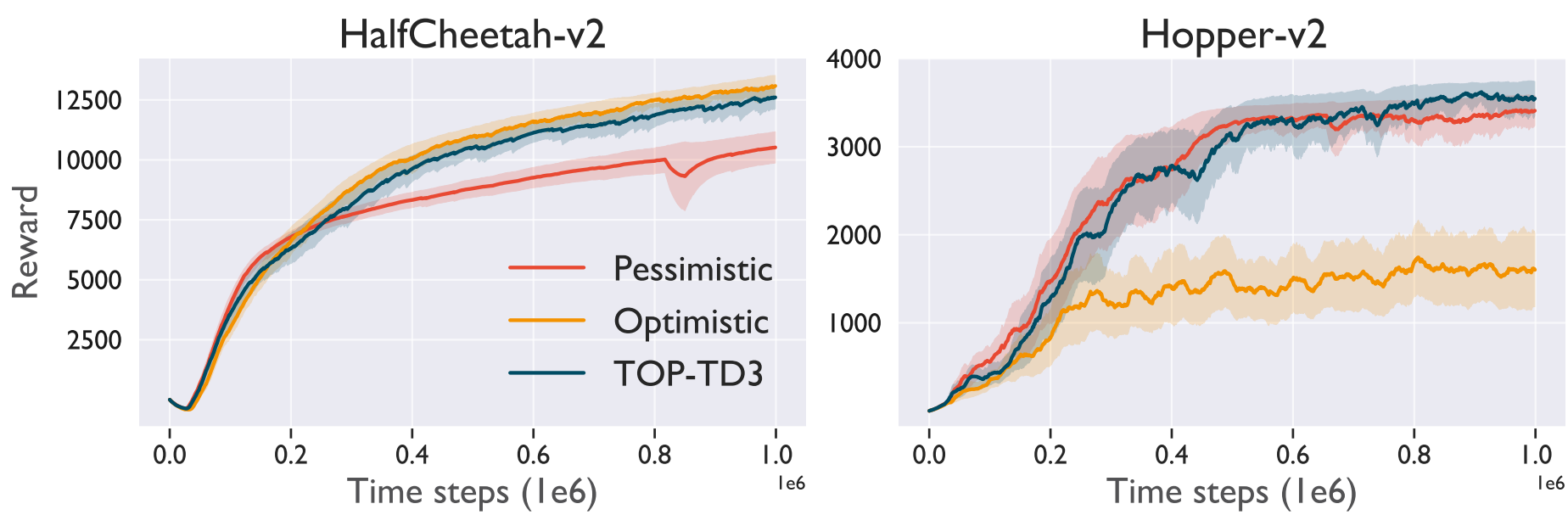
TOP + RAD [4]:

Experiments: Pixel-based Control

TOP + RAD [4]:

Experiments: The Impact of Adaptive Optimism





Summary

- It's difficult to know set the correct degree of optimism
- TOP is an adaptive, uncertainty-based method which does it for you
- TOP boosts SOTA performance on state- and pixel-based control
- Adding TOP requires only 10 lines of Python code

Thank you!

[1] M. G. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887*, 2017.

[2] K. Ciosek, Q. Vuong, R. Loftin, and K. Hofmann. Better exploration with optimistic actor-critic, 2019.

[3] S. Fujimoto, H. van Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. *CoRR*, abs/1802.09477, 2018.

[4] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas. Reinforcement learning with augmented data. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19884–19895. Curran Associates, Inc., 2020.

[5] S. Thrun and A. Schwartz. Issues in using function approximation for reinforcement learning. In *In Proceedings of the Fourth Connectionist Models Summer School*. Erlbaum, 1993.