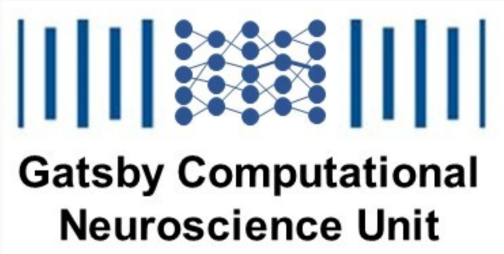
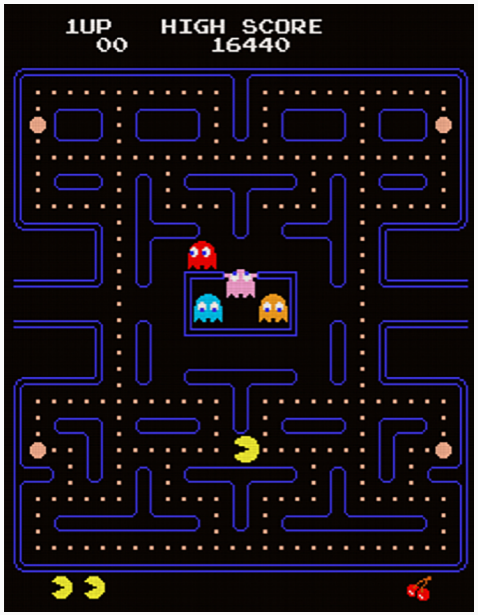
Tactical Optimism and Pessimism for Deep Reinforcement Learning

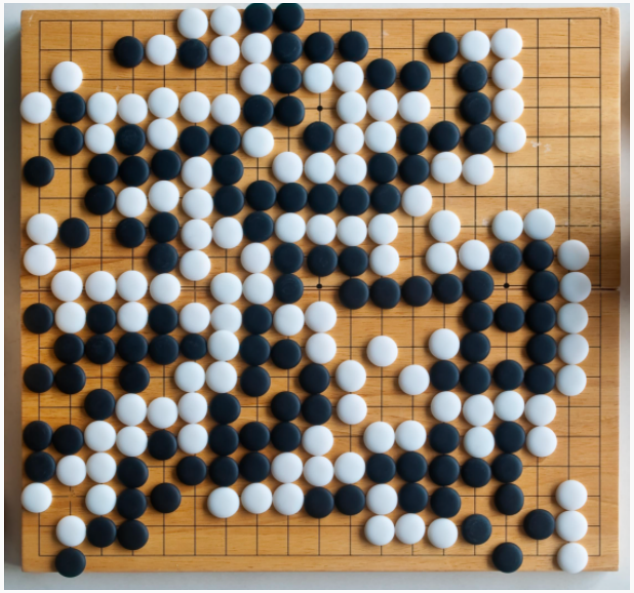
Ted Moskovitz, Jack Parker-Holder, Aldo Pacchiano,
Michael Arbel, Michael I. Jordan

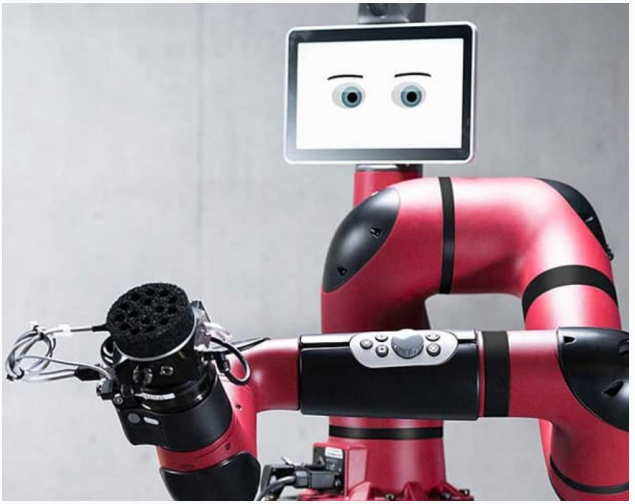
NeurIPS 2021



Motivation





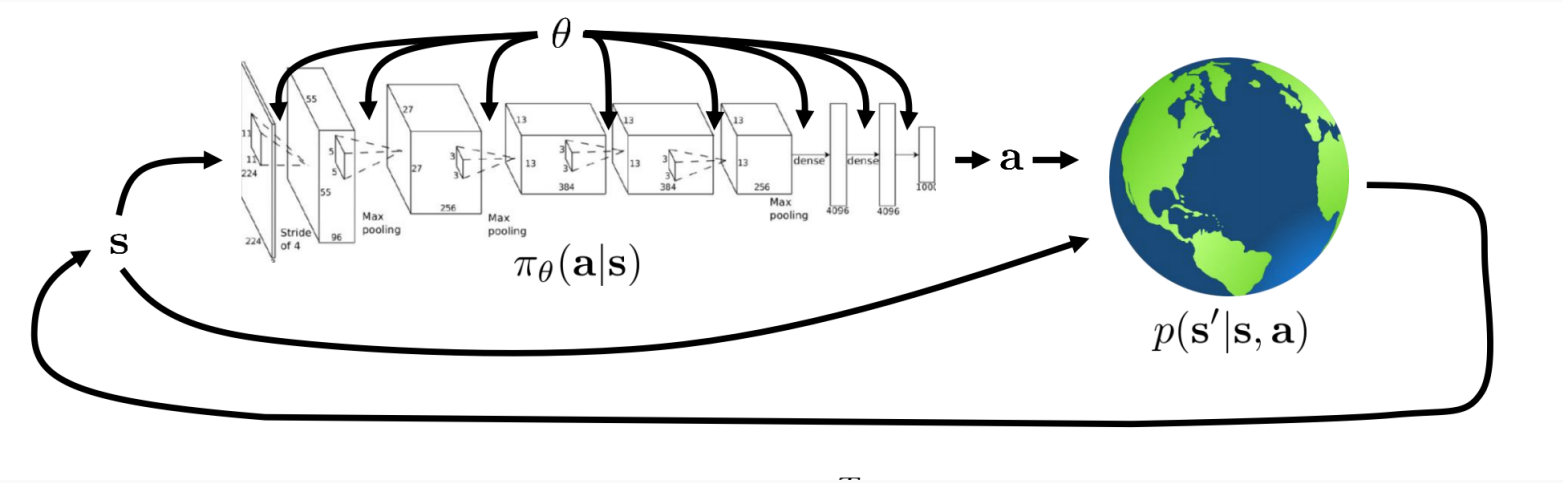


- Approximate value functions are key to the success of deep RL
- Optimism vs. pessimism: both have drawbacks + advantages
- How can we figure out which is best?

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 $(\mathcal{S}, \mathcal{A}, r, p, \gamma)$
- Running the policy in an episodic/finite horizon task of length T produces trajectories $\tau = (s_1, a_1, r_1, \ldots, s_T, a_T, r_T)$
- Maximize $J(\theta) = \mathbb{E}_\pi [Z(\tau)] = \mathbb{E}_\pi \left[\sum_t \gamma^t r_t\right]$

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

The Actor-Critic Framework

- Actor, π_θ: a deterministic policy
- Critic, $Q_\phi^{\pi_\theta}$: evaluate actor,

$$Q_\phi^{\pi_\theta}(s,a) := \mathbb{E}_\pi\left[Z_t | s_t = s, a_t = s\right] \tag{0.1}$$

- How to train?
- Actor: $\Delta\theta \propto \nabla_\theta J(\theta) = \mathbb{E}_\pi\left[\nabla_a Q_\phi^{\pi_\theta}(s,a)|_{a=\pi(s)} \nabla_\theta \pi_\theta(s)\right]$
- Critic: given a transition (s_t, a_t, r_t, s_{t+1}),

$$\Delta\phi \propto \nabla_\phi \frac{1}{2} \| \underbrace{y_t - Q_\phi^{\pi_\theta}(s_t, a_t)}_{:=\delta_t} \|_2^2 \tag{0.2}$$

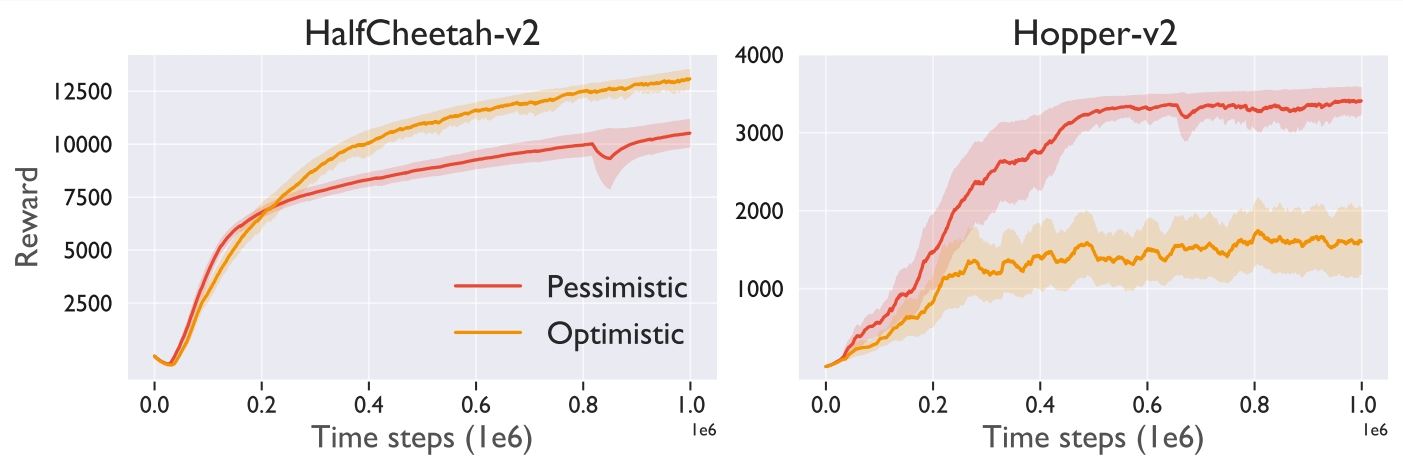
  where $y_t = r_t + \gamma Q_{\phi^-}^{\pi_\theta}(s_{t+1}, \pi_\theta(s_{t+1}))$
- Distributional RL: represent *distribution* of Z^π, not just mean [1]

Optimism vs. Pessimism

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

$$y_t = r_t + \gamma \min_{i \in \{1,2\}} Q^{\pi_\theta}_{\phi_i}(s_{t+1}, \pi_\theta(s) + \epsilon), \quad \epsilon \sim \mathrm{clip}(\mathcal{N}(0, s^2), -c, c) \tag{0.3}$$

- 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
- $\Rightarrow$ represent using *distributional value estimation* $Z \sim \mathcal{Z}^\pi(s,a)$
- **Epistemic uncertainty:** noise due to lack of knowledge about the world
- $\Rightarrow$ represent using an *ensemble* of $k = 1, \ldots, K$ critics
- We use these estimates to construct a *belief distribution* $\tilde{\mathcal{Z}}^\pi(s,a)$:

$$q^{(k)}_{\tilde{\mathcal{Z}}^\pi(s,a)} = q^{(k)}_{\bar{\mathcal{Z}}^\pi(s,a)} + \beta q^{(k)}_{\hat{\sigma}(s,a)} \tag{0.4}$$

- $\beta \in \mathbb{R}$ then determines the degree of optimism

TOP: Tactical Optimism and Pessimism

- *Q: How to choose β?*
- A: Evaluate β 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:

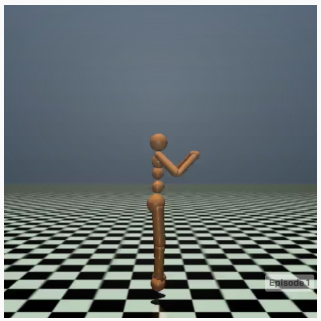
$$w_{m+1}(d) = \begin{cases} w_m(d) + \eta \frac{f_m}{\mathbf{p}_m(d)} & \text{if } d = d_m \\ w_m(d) & \text{otherwise,} \end{cases} \tag{0.5}$$

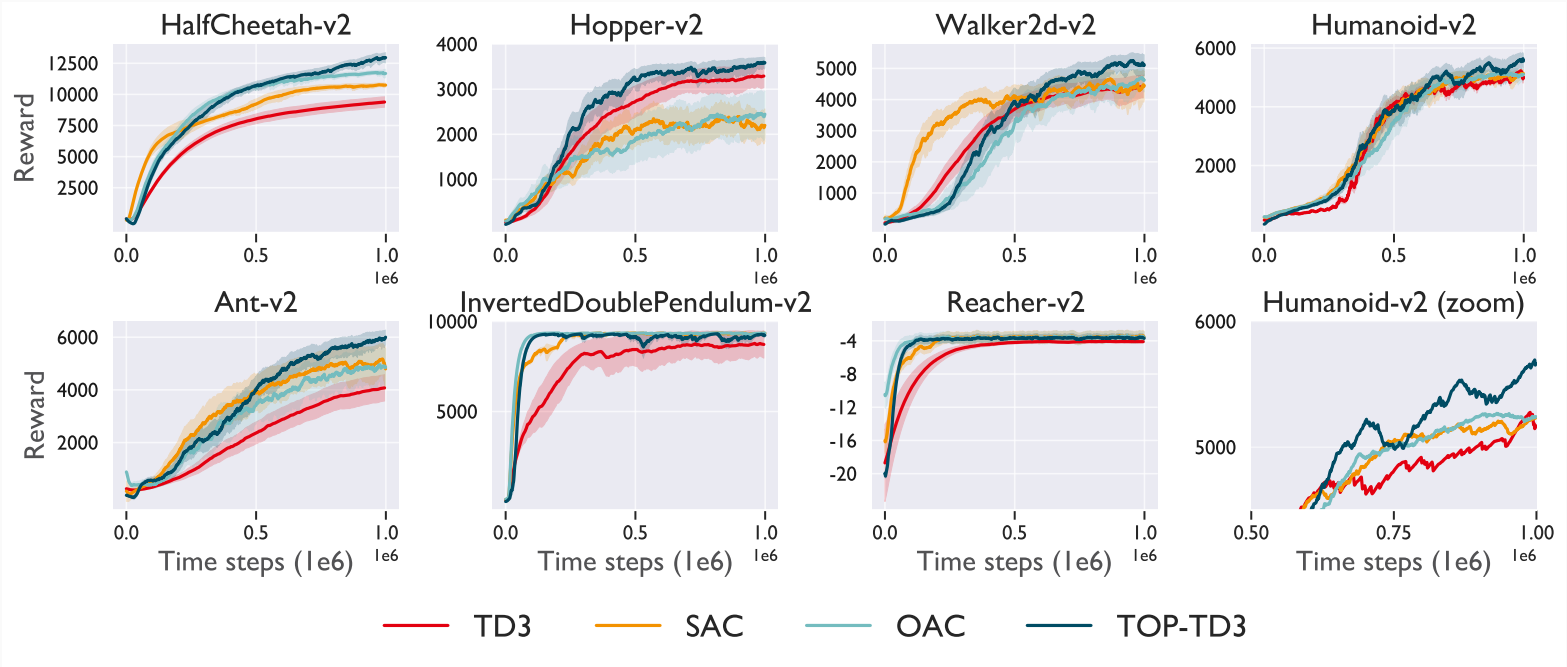
- Feedback f_m is change in performance:

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

Experiments: State-based Control

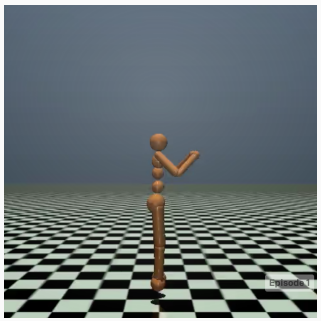
TOP + TD3 [3]:





Experiments: State-based Control

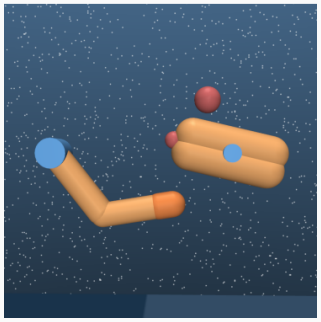
TOP + TD3 [3]:

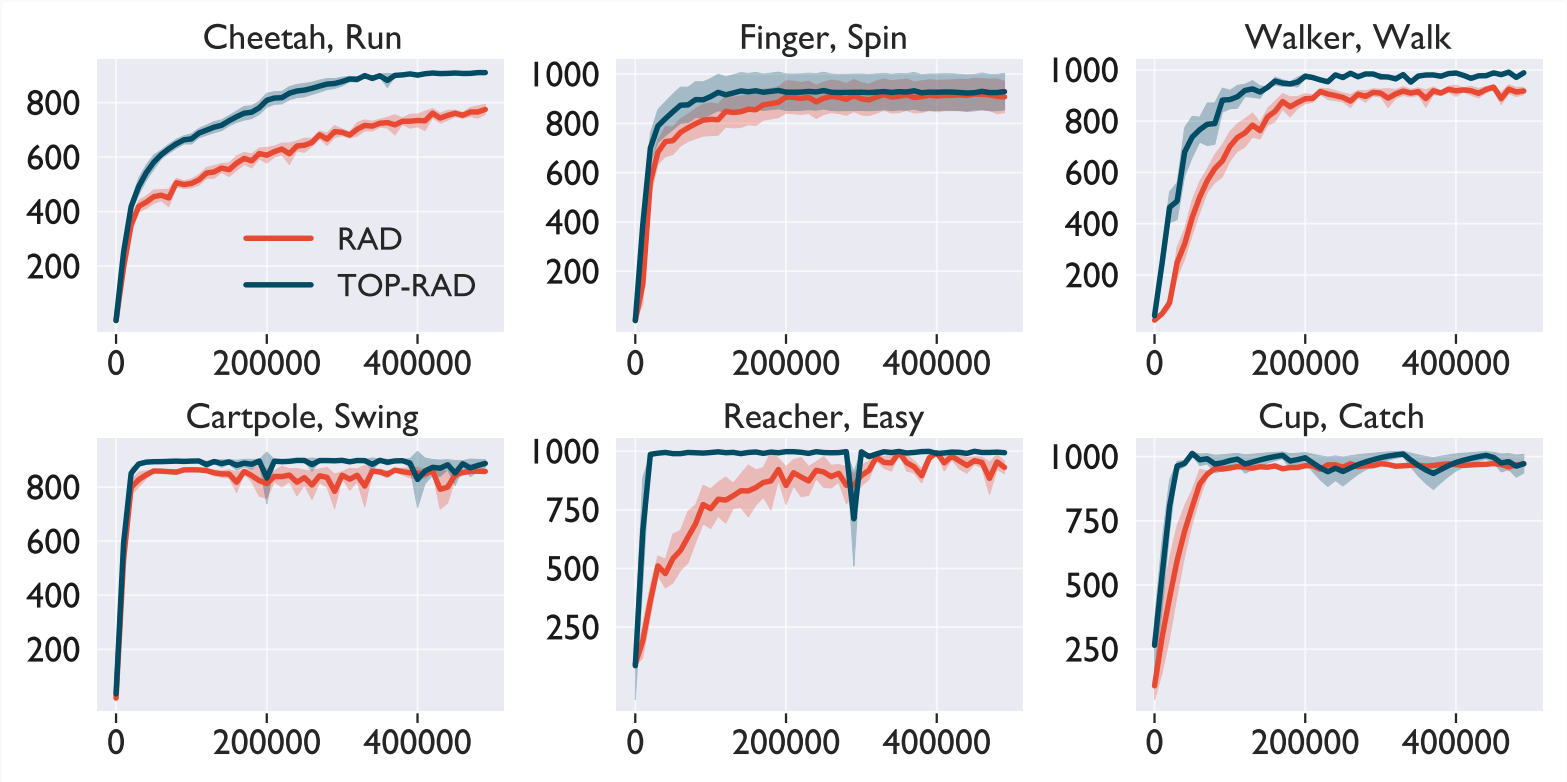


Task	TOP-TD3	ND TOP-TD3	QR-TD3	TD3	OAC	SAC
Humanoid	**5899±142***	5445	5003	5386	5349	5315
HalfCheetah	**13144 ± 701***	**12477**	11170	9566	11723	10815
Hopper	**3688 ± 33***	3458	3392	3390	2896	2237
Walker2d	**5111 ± 220***	4832	4560	4412	4786	4984
Ant	**6336 ± 181***	6096	5642	4242	4761	3421
InvDoublePend	**9337 ± 20***	**9330**	9299	8582	**9356**	**9348**
Reacher	**−3.85 ± 0.96**	**−3.91**	**−3.95**	**−4.22**	**−4.15**	**−4.14**

Experiments: Pixel-based Control

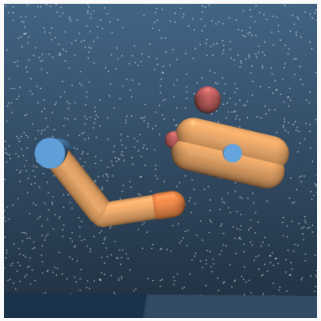
TOP + RAD [4]:





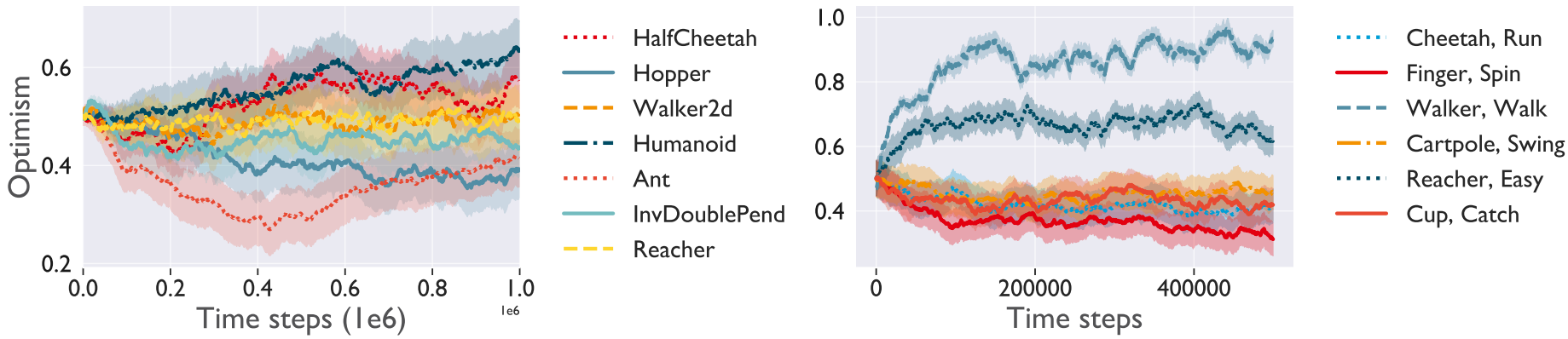
Experiments: Pixel-based Control

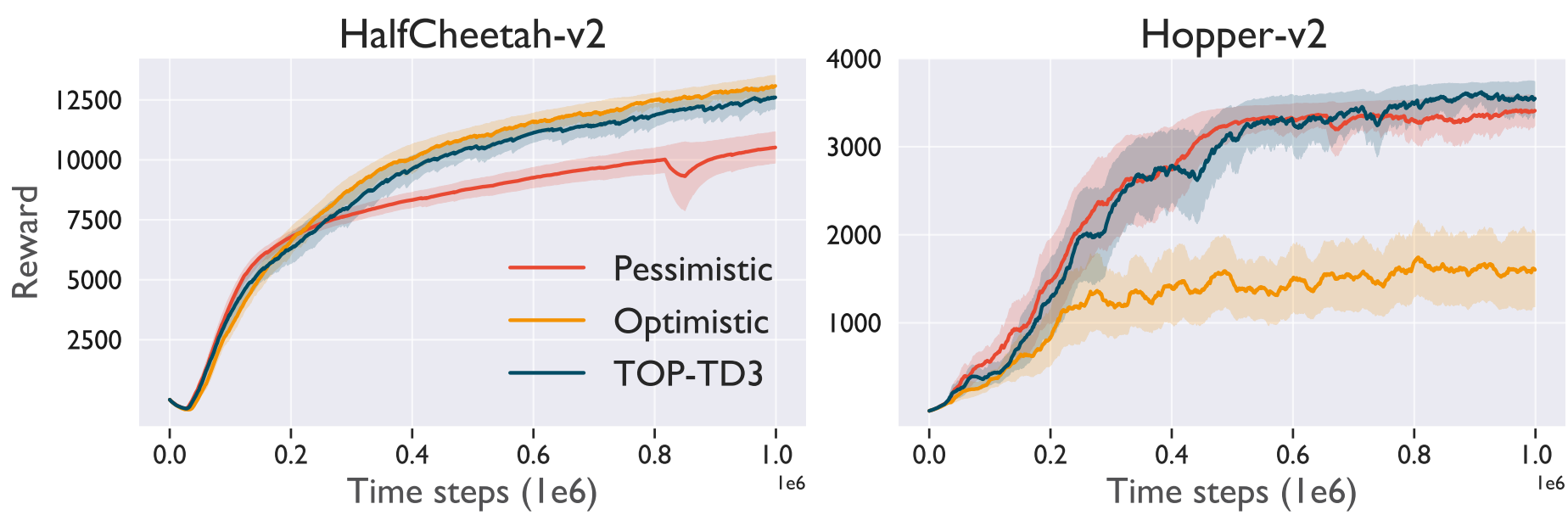
TOP + RAD [4]:



Task (100k)	TOP-RAD	RAD	DrQ	PI-SAC	CURL	PlaNet	Dreamer
Cheetah, Run	**674 ± 31***	499	344	460	299	307	235
Finger, Spin	873 ± 69	813	901	**957**	767	560	341
Walker, Walk	**862 ± 43***	644	612	514	403	221	277
Cartpole, Swing	**887 ± 13***	864	759	816	582	563	326
Reacher, Easy	**991 ± 3***	772	601	758	538	82	314
Cup, Catch	**970 ± 12***	950	913	933	769	710	246
Task (500k)	**TOP-RAD**	**RAD**	**DrQ**	**PI-SAC**	**CURL**	**PlaNet**	**Dreamer**
Cheetah, Run	**910 ± 4***	774	660	801	518	568	570
Finger, Spin	**928 ± 74**	**907**	**938**	**957***	**926**	718	796
Walker, Walk	**988 ± 4***	917	921	946	902	478	897
Cartpole, Swing	**890 ± 28***	858	**868**	816*	845	787	762
Reacher, Easy	**993 ± 5***	930	942	950	929	588	793
Cup, Catch	**972 ± 53***	**970**	**963**	**933***	**959***	**939**	879

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 $\sim$ 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.