Project #24

Optimistic Sampling for Exploration in Distributional RL

Alessio Russo Cristian Jensen Samuel Cestola

TA: Leo Widmer, ETH Zürich

ETH zürich

Background: Distributional Reinforcement Learning

Reinforcement learning (RL) is a subset of machine learning where an agent interacts with an environment to learn a policy that maximizes cumulative reward. Traditional RL methods typically estimate the expected return of a policy. In contrast, *distributional reinforcement learning* (DRL) captures the full distribution of returns, offering a better understanding of the variability and uncertainty inherent in the environment.

Distributional Bellman Equation

- Given a random variable $Z \sim \nu$ and a transformation $f : \mathbb{R} \to \mathbb{R}$, the distribution of $f(Z)$ is expressed as $f_{\#}\nu$, which is called the *pushforward distribution*. We then have $f_{\#}\nu(z) = \nu(f(z))$.
- *Distributional Bellman equation*: $\eta^\pi(s) = \mathbb{E}_\pi[(b_{R,\gamma})_{\#}\eta^\pi(S') \mid S = s]$, where $b_{r,\gamma}(z) \doteq r + \gamma z$ and $\eta^\pi(s)$ is the return distribution of state s under policy π.

Distribution Approximations

We cannot represent a full distribution with finite memory $\Rightarrow$ Approximate. (All models employ the same network as DQN with different output shapes.)

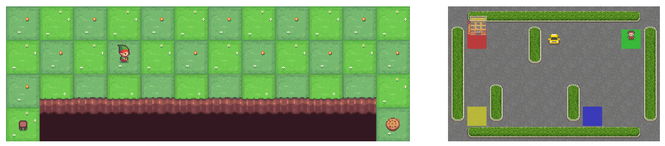
- C51 [BDM17] approximates the distribution as a categorical distribution, uniformly distributed between $V_{\min}$ and $V_{\max}$. It trains by minimizing the cross entropy between $\eta^\pi(s)$ and $\Pi((b_{r,\gamma})_{\#}\eta^\pi(s'))$, where Π projects the target distribution onto the predefined categories.
- QR-DQN [Dab+18a] solves the fixed support problem by approximating using quantiles with fixed value, effectively "transposing" the C51 representation. It trains by minimizing the quantile Huber loss.
- IQN [Dab+18b] extends QR-DQN by parametrizing the entire quantile function. The network learns to map quantiles to their value for a state-action pair, allowing the querying of any quantile. It also trains to minimize the quantile Huber loss.

Key Idea

Acting optimistically should make it possible to quickly discard obviously bad states, while still exploring actions that are potentially good. We can exploit the additional information that the return distribution provides to apply the principle of "Optimism in the face of uncertainty".

Background: Test Environments

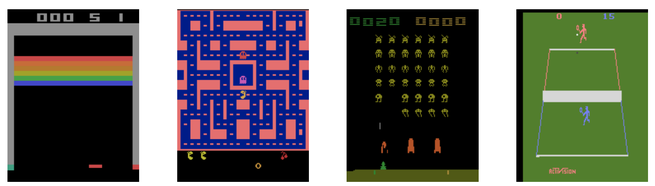
Gridworlds



(a) Cliff Walking (b) Taxi

Figure 1. Gridworld MDPs that we tested on [Tow+23].

Atari

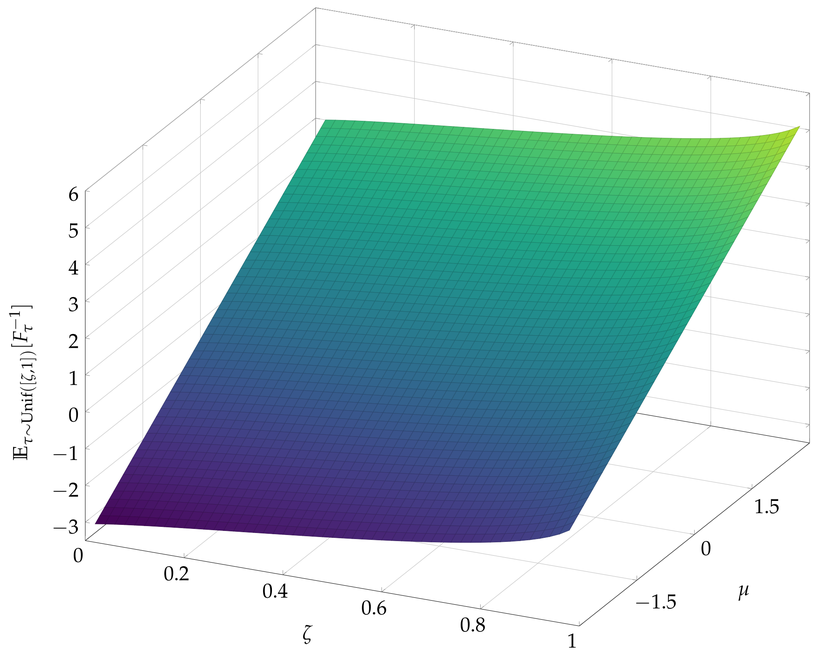


(a) Breakout (b) Ms. Pac-Man (c) Space Inv. (d) Tennis

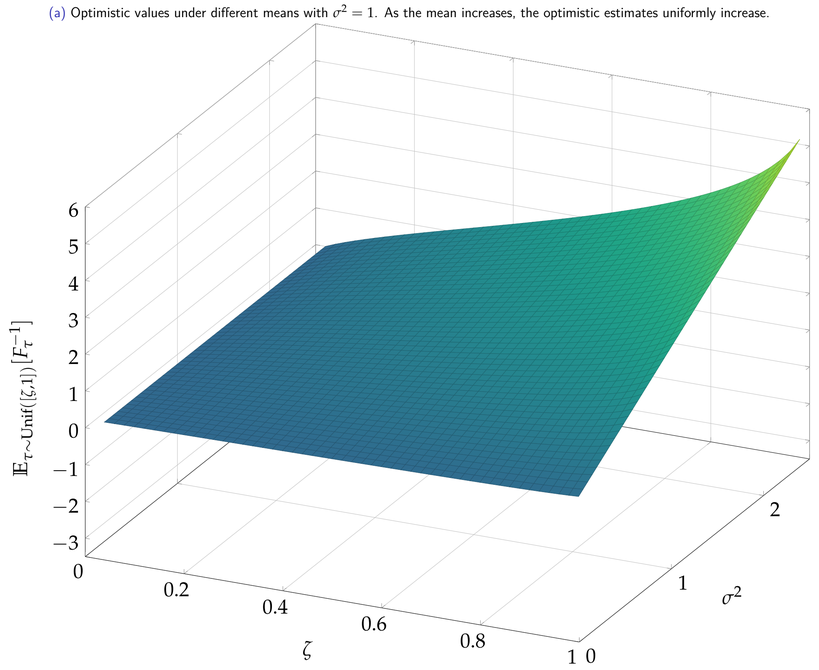
Figure 2. Atari games that we tested on [Bel+13].

Method 1: Optimistic Sampling

Intuition



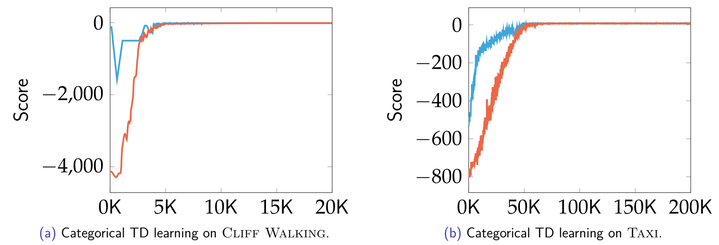
(a) Optimistic values under different means with $\sigma^2 = 1$. As the mean increases, the optimistic estimates uniformly increase.



(b) Optimistic values under different variances with $\mu = 0$. As the variance (uncertainty) grows, the more optimistic estimates are proportionally larger than the less optimistic estimates.

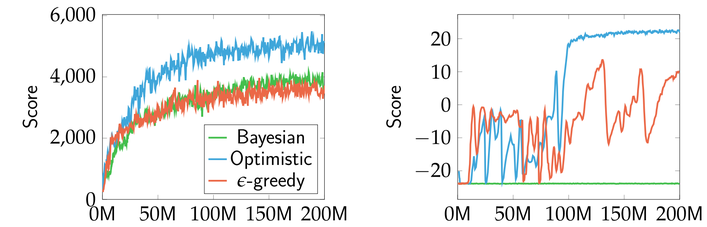
Figure 3. Upper quantile expectations of different Gaussian distributions.

Results



(a) Categorical TD learning on Cliff Walking. (b) Categorical TD learning on Taxi.

Figure 4. Performance during training on the Gridworld MDPs.



(a) IQN on Ms. Pac-Man. (b) C51 on Tennis.

Figure 5. Performance during training on the Atari games.

	Breakout	Ms. Pac-Man	Space Invaders	Tennis
O-C51	−13.7%	−2.8%	1.3%	126.3%
O-QR-DQN	−30.1%	5.3%	2.6%	21.3%
O-IQN	−9.3%	42.6%	−0.2%	944.4%
Mean	**−17.7%**	**15.0%**	**1.2%**	**364.0%**
BO-C51	17.1%	−23.8%	−26.3%	−341.4%
BO-QR-DQN	−3.5%	19.0%	−1.1%	−100.0%
BO-IQN	13.5%	3.9%	−20.7%	983.3%
Mean	**9.0%**	**−0.3%**	**−16.0%**	**180.7%**

Table 1. Performance increase of optimistic and optimistic Bayesian models, relative to their ϵ-greedy version, on Atari games. We report the relative increase in the mean score of the last 100 episodes of training.

Further work

- More experimentation with Bayesian neural networks.

References

[BDM17] Marc G Bellemare, Will Dabney, and Rémi Munos. "A distributional perspective on reinforcement learning". In: *International conference on machine learning*. PMLR. 2017, pp. 449–458.

[Bel+13] Marc G Bellemare et al. "The arcade learning environment: An evaluation platform for general agents". In: *Journal of Artificial Intelligence Research* 47 (2013), pp. 253–279.

[Blu+15] Charles Blundell et al. "Weight uncertainty in neural network". In: *International conference on machine learning*. PMLR. 2015, pp. 1613–1622.

[Dab+18a] Will Dabney et al. "Distributional reinforcement learning with quantile regression". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[Dab+18b] Will Dabney et al. "Implicit quantile networks for distributional reinforcement learning". In: *International conference on machine learning*. PMLR. 2018, pp. 1096–1105.

[Tow+23] Mark Towers et al. *Gymnasium*. Mar. 2023. DOI: 10.5281/zenodo.8127026. URL: https://zenodo.org/record/8127025 (visited on 07/08/2023).