Distributional Deep Reinforcement Learning (Part 1)

Adrian Meier

February 26, 2019

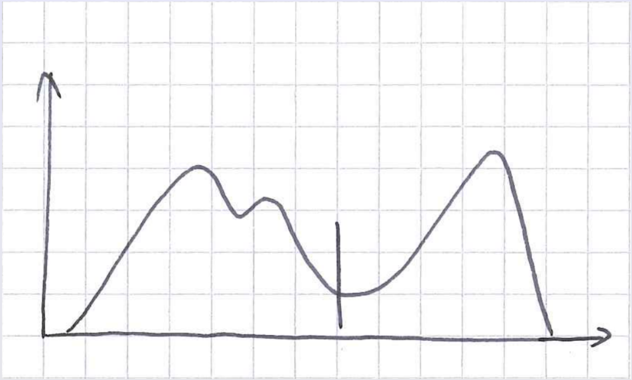
A Distributional Perspective on Reinforcement Learning
Marc G. Bellemare, Will Dabney, Rmi Munos. ICML 2017.

Consider Y to be the random variable that describes the result of rolling a six-sided die once.

$$R(Y) = \begin{cases} -2, & \text{for } Y = 1 \\ 1, & \text{otherwise} \end{cases}$$

$\mathbb{E}[R(Y)] = \frac{1}{6} \cdot (-2) + \frac{5}{6} \cdot 1 = \frac{1}{2}$

Expected return hides intrinsic randomness.



Key idea

Consider the **value distribution** Z of the random return, to capture randomness

$$\mathbb{E}(Z) = Q$$

- in the reward R
- in the transition P^{π}
- in the next-state value distribution $Z(X', A')$

Recall Bellman's equations

$$Q^*(x,a) = \mathbb{E}R(x,a) + \gamma \mathbb{E}_P \max_{a' \in A} Q^*(x',a')$$

Optimality operator T, repeated application to some Q_0 converges exponentially to fixed point Q^*.

$$TQ(x,a) := \mathbb{E}R(x,a) + \gamma \mathbb{E}_P \max_{a' \in A} Q(x',a')$$

How to write a Bellman equation using the value distribution?

$$Z(x,a) \overset{D}{=} R(x,a) + \gamma Z(X',A')$$

Find an operator which is a contraction mapping to show convergence.

Need a suitable measure of distance between distributions .

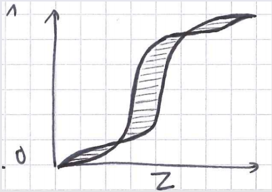
Distance functions between probability distributions

Kullback-Leibler divergence:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

Wassertein metric:

$$w_p(X, Y) = \left(\int_0^1 |F_X^{-1}(t) - G_Y^{-1}(t)|^p dt \right)^{\frac{1}{p}}$$



$$TZ(x,a) \overset{D}{=} R(x,a) + \gamma \max_{a' \in A} Z(X', A')$$

Converges to a sequence of optimal policies using the Wasserstein metric.

Distributional RL algorithm sketch

1. From x, a, sample a transition
2. Compute sample backup
3. Minimize Wasserstein loss

Modeling the value distribution

Discrete distribution, supported by N equispaced atoms between V_{min} and V_{max}.

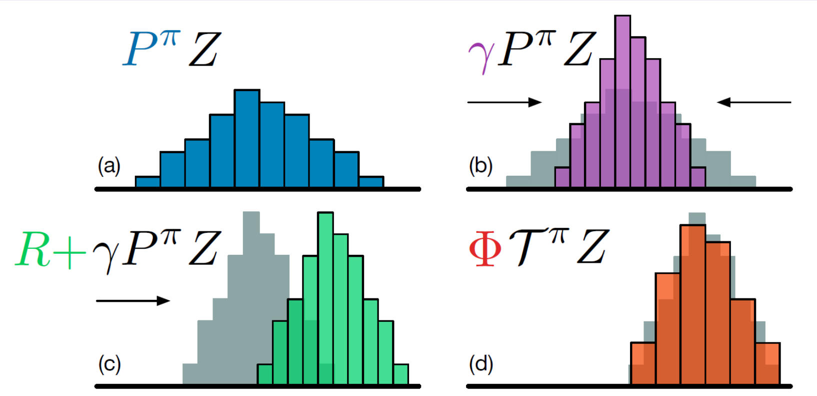
Bellman update is projected to the atoms by multiclass classification.

Choice of 51 atoms found to be highly expressive while remaining computationally friendly.

Categorical 'C51' Algorithm

1. From x, a, sample a transition
2. Compute sample backup
3. Project onto approximation support
4. Minimize KL-divergence

C51 iteration step visualized



Demo and discussion

Link to video

Conclusion

C51 drew lots of attention.

Can be further improved with known algorithmic optimizations.

Introduced a gap between theory and practice.