Distributional Reinforcement Learning

Hao Liang

The Chinese University of Hongkong, Shenzhen

March 11, 2019

Outline

Distributional Reinforcement Learning

- The traditional reinforcement learning (RL) is interested in maximizing the *expected return* so we usually work directly with those expectations.
- The main idea of *distributional* RL (M. G. Bellemare, Dabney, and Munos 2017) is to work directly with the full distribution of the return rather than with its expectation.
- Distributions rather than expectations are being optimized.

Setting

- Time-homogeneous Markov Decision Process $(\mathcal{X}, \mathcal{A}, R, P, \gamma)$.
- $\mathcal{X}$ and $\mathcal{A}$ are respectively the state and action spaces, P is the transition kernel $P(\cdot|x, a)$, $\gamma \in [0, 1]$ is the discount factor, and R is the reward function.
- We explicitly treat R as a random variable .
- A stationary policy π maps each state $x \in \mathcal{X}$ to a probability distribution over the action space $\mathcal{A}$.

Bellman's Equations

- The *return* Z^π is the sum of discounted rewards, which is also a random varaible(r.v.).
- The value function Q^π of a policy π describes the expected return from taking action $a \in \mathcal{A}$ from state $x \in \mathcal{X}$, then acting according to π:

$$Q^\pi(x,a) := \mathbb{E}Z^\pi(x,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(x_t, a_t)\right], \tag{1}$$

$$x_t \sim P(\cdot|x_{t-1}, a_{t-1}), a_t \sim \pi(\cdot|x_t), x_0 = x, a_0 = a.$$

- Bellman's equation for value function

$$Q^\pi(x,a) = \mathbb{E}R(x,a) + \gamma \mathbb{E}_{P,\pi} Q^\pi(x', a').$$

Bellman's Equations

- Bellman's optimality Equations

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

- A policy π^* is optimal if $\mathbb{E}_{a \sim \pi^*} Q^*(x,a) = \max_a Q^*(x,a)$.
- The *Bellman operator* $\mathcal{T}^\pi$ and *optimality operator* $\mathcal{T}$ are

$$\mathcal{T}^\pi Q(x,a) := \mathbb{E}R(x,a) + \gamma \mathbb{E}_{P,\pi} Q(x', a') \quad (2)$$

$$\mathcal{T} Q(x,a) := \mathbb{E}R(x,a) + \gamma \mathbb{E}_P \max_{a' \in \mathcal{A}} Q(x', a'). \quad (3)$$

- They are both contraction mappings (w.r.t. infinity norm), and their repeated application to some initial Q_0 converges exponentially to Q^π or Q^*.

Notations

- Probability space $(\Omega, \mathcal{F}, \Pr)$.
- $\|\mathbf{u}\|_p$: the L_p norm of a vector $\mathbf{u} \in \mathbf{R}^{\mathcal{X}}$ for $1 \leq p \leq \infty$; applies to vectors in $\mathbf{R}^{\mathcal{X}\times\mathcal{A}}$.
- The L_p norm of a random vector $U : \Omega \to \mathbf{R}^{\mathcal{X}}$ (or $\mathbf{R}^{\mathcal{X}\times\mathcal{A}}$) is $\|U\|_p := [\mathbb{E}[\|U(\omega)\|_p^p]]^{1/p}$.
- The c.d.f. of a random variable U by $F_U(y) := \Pr\{U \leq y\}$, and its inverse c.d.f. by $F_U^{-1}(q) := \inf\{y : F_U(y) \geq q\}$.
- A distributional equation $U \stackrel{D}{:=} V$ indicates that the distribution function of random variable U is the same as the distribution function of V.

The Wasserstein Metric

- The Wasserstein metric is defined between two c.d.fs F, G:

$$d_p(F, G) := \inf_{U,V} \|U - V\|_p,$$

where the infimum is taken over all pairs of random variables (U, V) with respective cumulative distributions F and G.

- Given two random variables U, V with c.d.fs F_U, F_V, we will write $d_p(U, V) := d_p(F_U, F_V)$.
- The metric d_p has the following properties:

$$d_p(aU, aV) \leq |a| d_p(U, V) \tag{P1}$$

$$d_p(A + U, A + V) \leq d_p(U, V) \tag{P2}$$

$$d_p(AU, AV) \leq \|A\|_p d_p(U, V). \tag{P3}$$

where a is a scalar and random variable A independent of U, V.

The Wasserstein Metric

- This metric can be extended to vectors of random variables, such as value distributions $Z(x, a)$, using the corresponding L_p norm.
- Let $\mathcal{Z}$ denote the space of value distributions with bounded moments. For two value distributions $Z_1, Z_2 \in \mathcal{Z}$ we will make use of a maximal form of the Wasserstein metric:

$$\bar{d}_p(Z_1, Z_2) := \sup_{x,a} d_p(Z_1(x, a), Z_2(x, a)).$$

- $\bar{d}_p$ will be used to establish the convergence of the distributional Bellman operators.

Lemma 1

$\bar{d}_p$ is a metric over value distributions.

The Wasserstein Metric

Proof.

The only nontrivial property is the triangle inequality. For any value distribution $Y \in \mathcal{Z}$, write

$$\begin{aligned}
\bar{d}_p(Z_1, Z_2) &= \sup_{x,a} d_p(Z_1(x,a), Z_2(x,a)) \\
&\overset{(a)}{\leq} \sup_{x,a} \left[d_p(Z_1(x,a), Y(x,a)) + d_p(Y(x,a), Z_2(x,a)) \right] \\
&\leq \sup_{x,a} d_p(Z_1(x,a), Y(x,a)) + \sup_{x,a} d_p(Y(x,a), Z_2(x,a)) \\
&= \bar{d}_p(Z_1, Y) + \bar{d}_p(Y, Z_2),
\end{aligned}$$

where in (a) we used the triangle inequality for d_p. □

Policy Evaluation

- We view the reward function as a random vector $R \in \mathcal{Z}$, and define the transition operator $P^\pi : \mathcal{Z} \to \mathcal{Z}$

$$P^\pi Z(x,a) \overset{D}{:=} Z(X', A') \tag{4}$$
$$X' \sim P(\cdot \,|\, x, a),\ A' \sim \pi(\cdot \,|\, X'),$$

  where capital letters are used to emphasize the random nature of the next state-action pair (X', A').
- The distributional Bellman operator $\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ is defined as

$$\mathcal{T}^\pi Z(x,a) \overset{D}{:=} R(x,a) + \gamma P^\pi Z(x,a). \tag{5}$$

Policy Evaluation

- Three sources of randomness define the compound distribution $\mathcal{T}^\pi Z$
  - The randomness in the reward R
  - The randomness in the transition P^π
  - The next-state value distribution $Z(X', A')$
- We make the usual assumption that these three quantities are independent.
- (5) is a contraction mapping whose unique fixed point is the random return Z^π.

Contraction in $\bar{d}_p$

- The distributional policy evaluation process $Z_{k+1} := \mathcal{T}^\pi Z_k$, starting with some $Z_0 \in \mathcal{Z}$ converges in the sense of $\bar{d}_p$.

Lemma 2

$\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ is a γ-contraction in $\bar{d}_p$.

Proof.

Consider $Z_1, Z_2 \in \mathcal{Z}$. By definition,

$$\bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) = \sup_{x,a} d_p(\mathcal{T}^\pi Z_1(x,a), \mathcal{T}^\pi Z_2(x,a)). \quad (6)$$

Contraction in $\bar{d}_p$

Proof.

By the properties of d_p, we have

$$\begin{aligned} & d_p(\mathcal{T}^\pi Z_1(x,a), \mathcal{T}^\pi Z_2(x,a)) \\ & \qquad = d_p(R(x,a) + \gamma P^\pi Z_1(x,a), R(x,a) + \gamma P^\pi Z_2(x,a)) \\ & \qquad \le \gamma d_p(P^\pi Z_1(x,a), P^\pi Z_2(x,a)) \\ & \qquad \le \gamma \sup_{x',a'} d_p(Z_1(x',a'), Z_2(x',a')), \end{aligned}$$

where the last line follows from the definition of P^π (see (4)). Combining with (6) we obtain

$$\begin{aligned} \bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) &= \sup_{x,a} d_p(\mathcal{T}^\pi Z_1(x,a), \mathcal{T}^\pi Z_2(x,a)) \\ &\le \gamma \sup_{x',a'} d_p(Z_1(x',a'), Z_2(x',a')) \\ &= \gamma \bar{d}_p(Z_1, Z_2). \end{aligned}$$

□

Contraction in $\bar{d}_p$

- Using Lemma 2, we conclude using Banach's fixed point theorem that $\mathcal{T}^\pi$ has a unique fixed point, which is Z^π as defined in (1).
- $\mathcal{T}^\pi$ is not a contraction in all metrics.
- Chung $\&$ Sobel (1987) have shown that $\mathcal{T}^\pi$ is not a contraction in total variation distance. Similar results can be derived for the Kullback-Leibler divergence.

Control

- Different from policy evaluation, we consider the *control* setting where we seek a policy π that maximizes value.
- While all optimal policies attain the same value Q^*, in general there are many optimal value distributions.
- We will show that the distributional Bellman optimality operator converges, in a weak sense, to the set of optimal value distributions. However, this operator is not a contraction in any metric between distributions.
- Let Π^* be the set of optimal policies. We can define the *optimal value distribution* .

Control

Definition 1 (Optimal value distribution)

An optimal value distribution is the v.d. of an optimal policy. The set of optimal value distributions is $\mathcal{Z}^* := \{Z^{\pi^*} : \pi^* \in \Pi^*\}$.

- Not all value distributions with expectation Q^* are optimal: they must match the full distribution of the return under some optimal policy.

Definition 2 (Set of greedy polices)

A greedy policy π *for* $Z \in \mathcal{Z}$ *maximizes the expectation of* Z*. The set of greedy policies for* Z *is*

$$\mathcal{G}_Z := \{\pi : \sum_a \pi(a \,|\, x)\mathbb{E}Z(x, a) = \max_{a' \in \mathcal{A}} \mathbb{E}Z(x, a')\}.$$

Control

- Recall that the expected Bellman optimality operator $\mathcal{T}$ is

$$\mathcal{T}Q(x,a) = \mathbb{E}R(x,a) + \gamma \mathbb{E}_P \max_{a' \in \mathcal{A}} Q(x',a'). \tag{7}$$

  The maximization at x' corresponds to some greedy policy implictly.

- We call a *distributional Bellman optimality operator* any operator $\mathcal{T}$ which implements a greedy selection rule

$$\mathcal{T}Z = \mathcal{T}^{\pi} Z \text{ for some } \pi \in \mathcal{G}_Z.$$

  Here we need to explicitly specify a optimal policy π for a given value distribution.

- We are interested in the behaviour of the iterates $Z_{k+1} := \mathcal{T}Z_k$, $Z_0 \in \mathcal{Z}$.

Control

Lemma 3 (Convergence of $\mathbb{E}Z_k$)

Let $Z_1, Z_2 \in \mathcal{Z}$. *Then*

$$\|\mathbb{E}\mathcal{T}Z_1 - \mathbb{E}\mathcal{T}Z_2\|_\infty \leq \gamma \|\mathbb{E}Z_1 - \mathbb{E}Z_2\|_\infty ,$$

and in particular $\mathbb{E}Z_k \to Q^*$ *exponentially quickly.*

Proof.

The proof follows by linearity of expectation. Write $\mathcal{T}_D$ for the distributional operator and $\mathcal{T}_E$ for the usual operator. Then

$$\begin{aligned}\|\mathbb{E}\mathcal{T}_D Z_1 - \mathbb{E}\mathcal{T}_D Z_2\|_\infty &= \|\mathcal{T}_E\mathbb{E}Z_1 - \mathcal{T}_E\mathbb{E}Z_2\|_\infty \\ &\leq \gamma \|\mathbb{E}Z_1 - \mathbb{E}Z_2\|_\infty .\end{aligned}$$

□

Control

- However, convergence of itself is not assured to reach a fixed point.

Definition 3

A nonstationary optimal value distribution Z^{**} is the value distribution corresponding to a sequence of optimal policies. The set of n.o.v.d. is $\mathcal{Z}^{**}$.

Theorem 1 (Convergence in the control setting)

Let $\mathcal{X}$ be measurable and suppose that $\mathcal{A}$ is finite. Then

$$\lim_{k\to\infty} \inf_{Z^{**}\in\mathcal{Z}^{**}} d_p(Z_k(x,a), Z^{**}(x,a)) = 0 \quad \forall x, a.$$

If $\mathcal{X}$ is finite, then Z_k converges to $\mathcal{Z}^{**}$ uniformly. Furthermore, if there is a total ordering $\prec$ on Π^*, such that for any $Z^* \in \mathcal{Z}^*$,

$$\mathcal{T} Z^* = \mathcal{T}^\pi Z^* \text{ with } \pi \in \mathcal{G}_{Z^*},\ \pi \prec \pi' \ \ \forall \pi' \in \mathcal{G}_{Z^*} \setminus \{\pi\}.$$

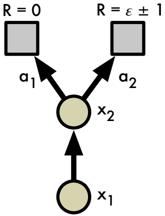
Then $\mathcal{T}$ has a unique fixed point $Z^* \in \mathcal{Z}^*$.

Control

Proposition 1

The operator $\mathcal{T}$ is not a contraction.

- Consider the following example (Figure 1, left).



	x_1	x_2, a_1	x_2, a_2
Z^*	$\epsilon \pm 1$	0	$\epsilon \pm 1$
Z	$\epsilon \pm 1$	0	$-\epsilon \pm 1$
$\mathcal{T}Z$	0	0	$\epsilon \pm 1$

Figure: Undiscounted two-state MDP for which the optimality operator $\mathcal{T}$ is not a contraction, with example. The entries that contribute to $\bar{d}_1(Z, Z^*)$ and $\bar{d}_1(\mathcal{T}Z, Z^*)$ are highlighted.

Control

- consider Z as given in Figure 1 (right), and its distance to Z^*:

$$\bar{d}_1(Z, Z^*) = d_1(Z(x_2, a_2), Z^*(x_2, a_2)) = 2\epsilon,$$

- When we apply $\mathcal{T}$ to Z, however, the greedy action a_1 is selected and $\mathcal{T}Z(x_1) = Z(x_2, a_1)$. But

$$\begin{aligned}\bar{d}_1(\mathcal{T}Z, \mathcal{T}Z^*) &= d_1(\mathcal{T}Z(x_1), Z^*(x_1)) \\ &= \tfrac{1}{2}|1-\epsilon| + \tfrac{1}{2}|1+\epsilon| > 2\epsilon\end{aligned}$$

  for a sufficiently small ϵ.

- Using a more technically involved argument, we can extend this result to any metric which separates Z and $\mathcal{T}Z$.

Control

Proposition 2

Not all optimality operators have a fixed point $Z^* = \mathcal{T} Z^*$.

To see this, consider the same example, now with $\epsilon = 0$, and a greedy operator $\mathcal{T}$ which breaks ties by picking a_2 if $Z(x_1) = 0$, and a_1 otherwise. Then the sequence $\mathcal{T} Z^*(x_1), (\mathcal{T})^2 Z^*(x_1), \ldots$ alternates between $Z^*(x_2, a_1)$ and $Z^*(x_2, a_2)$.

Proposition 3

That $\mathcal{T}$ has a fixed point $Z^* = \mathcal{T} Z^*$ is insufficient to guarantee the convergence of $\{Z_k\}$ to $\mathcal{Z}^*$.

Approximate Distributional Learning

- Full computation of the distributional Bellman operator on a return distribution function is typically either impossible (due to unknown MDP dynamics), or computationally infeasible.
- Several key approximations are required to produce a practical, scalable distributional RL algorithm
  - distribution parametrisation
  - stochastic approximation of the Bellman operator
  - projection of the Bellman target distribution
  - gradient updates via a loss function

Parametric Distribution

- We will approximate the value distribution using a discrete distribution parametrized by N and $V_{\text{MIN}}, V_{\text{MAX}}$, and whose support is the set of atoms $\{z_i = V_{\text{MIN}} + i\triangle z : 0 \leq i < N\}$, $\triangle z := \frac{V_{\text{MAX}} - V_{\text{MIN}}}{N-1}$.
- The atom probabilities are given by a parametric model $\theta : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^N$

$$Z_\theta(x,a) = z_i \quad \text{w.p.} \quad p_i(x,a) := \frac{e^{\theta_i(x,a)}}{\sum_j e^{\theta_j(x,a)}}.$$

Projected Bellman Update

- Using a discrete distribution may cause the Bellman update $\mathcal{T} Z_\theta$ and our parametrization Z_θ almost always have disjoint supports.
- It is natural to minimize the Wasserstein metric (viewed as a loss) between $\mathcal{T} Z_\theta$ and Z_θ, which is also robust to discrepancies in support.
- Evaluation of the distributional Bellman operator requires integrating over all possible next state-action-reward combinations, so stochastic approximation of Bellman operator which learns from sample transitions is needed.
- Combining them together, we project the sample Bellman update $\hat{\mathcal{T}} Z_\theta$ onto the support of Z_θ.

Projected Bellman Update

- Given a sample transition (x, a, r, x'), we compute the Bellman update $\hat{\mathcal{T}} z_j := r + \gamma z_j$ for each atom z_j, then distribute its probability $p_j(x', \pi(x'))$ to the immediate neighbours of $\hat{\mathcal{T}} z_j$.
- The next-state distribution as parametrized by a fixed parameter $\tilde{\theta}$. The sample loss $\mathcal{L}_{x,a}(\theta)$ is the cross-entropy term of the KL divergence

$$D_{\mathrm{KL}}(\Phi \hat{\mathcal{T}} Z_{\tilde{\theta}}(x, a) \,\|\, Z_\theta(x, a)),$$

which is readily minimized e.g. using gradient descent.
- This choice of distribution and loss is called the *categorical algorithm*

Algorithm 1 Categorical Algorithm

input A transition $x_t, a_t, r_t, x_{t+1}, \gamma_t \in [0, 1]$

$Q(x_{t+1}, a) := \sum_i z_i p_i(x_{t+1}, a)$

$a^* \leftarrow \arg\max_a Q(x_{t+1}, a)$

$m_i = 0, \quad i \in 0, \ldots, N-1$

for $j \in 0, \ldots, N-1$ **do**

  # Compute the projection of $\hat{\mathcal{T}} z_j$ onto the support $\{z_i\}$

  $\hat{\mathcal{T}} z_j \leftarrow [r_t + \gamma_t z_j]_{V_{\text{MIN}}}^{V_{\text{MAX}}}$

  $b_j \leftarrow (\hat{\mathcal{T}} z_j - V_{\text{MIN}})/\Delta z$ # $b_j \in [0, N-1]$

  $l \leftarrow \lfloor b_j \rfloor, u \leftarrow \lceil b_j \rceil$

  # Distribute probability of $\hat{\mathcal{T}} z_j$

  $m_l \leftarrow m_l + p_j(x_{t+1}, a^*)(u - b_j)$

  $m_u \leftarrow m_u + p_j(x_{t+1}, a^*)(b_j - l)$

end for

output $-\sum_i m_i \log p_i(x_t, a_t)$ # Cross-entropy loss

Figure: **Categorial Algorithm**

Arcade Learning Environment

- The categorical algorithm was applied to games from the Arcade Learning Environment. Five training games (Fig 3) and 52 testing games were used.

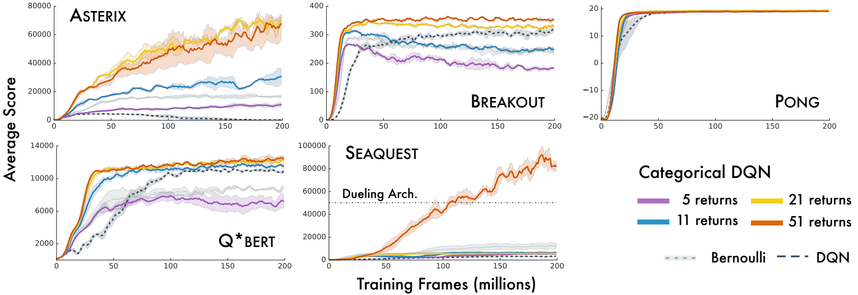


Figure: Categorical DQN: Varying number of atoms in the discrete distribution. Scores are moving averages over 5 million frames.

Arcade Learning Environment

- DQN architecture. Output the atom probabilities $p_i(x, a)$ instead of action-values, and chose $V_{\text{MAX}} = -V_{\text{MIN}} = 10$.
- Replace the squared loss $(r + \gamma Q(x', \pi(x')) - Q(x, a))^2$ by $\mathcal{L}_{x,a}(\theta)$ and train the network to minimize this loss.
- Figure 4 illustrates the typical value distributions we observed in our experiments.
- Three actions lead to the agent releasing its laser too early and eventually losing the game. The corresponding distributions assign a significant probability to 0.

Arcade Learning Environment

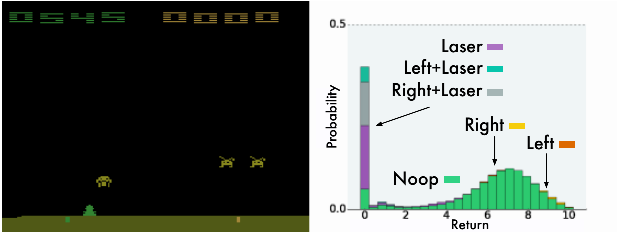


Figure: Learned value distribution during an episode of SPACE INVADERS. Different actions are shaded different colours. Returns below 0 (which do not occur in SPACE INVADERS) are not shown here as the agent assigns virtually no probability to them.

State-of-the-Art Results

- The performance of the 51-atom agent (C51) on the training games was compared with DQN ($\epsilon = 0.01$), Double DQN (van Hasselt et al., 2016) , the Dueling architecture (Wang et al., 2016), and Prioritized Replay (Schaul et al., 2016) , comparing the best evaluation score achieved during training.

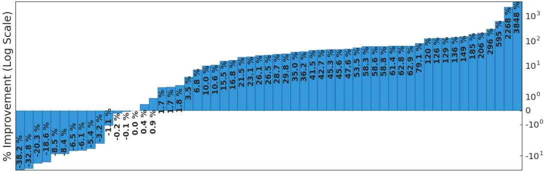


Figure: Percentage improvement, per-game, of C51 over Double DQN, computed using van Hasselt et al.'s method.

State-of-the-Art Results

	Mean	Median	> H.B.	> DQN
DQN	228%	79%	24	0
DDQN	307%	118%	33	43
Duel.	373%	151%	37	50
Prior.	434%	124%	39	48
Pr. Duel.	592%	172%	39	44
C51	**701%**	**178%**	**40**	**50**
UNREAL†	880%	250%	-	-

Figure: Mean and median scores across 57 Atari games, measured as percentages of human baseline (H.B., Nair et al., 2015)

Why Does Learning a Distribution Matter?

- **Reduced chattering.** The instability in the Bellman optimality operator combined with function approximation may prevent the policy from converging. The gradient-based categorical algorithm is able to mitigate these effects by effectively averaging the different distributions.
- **A richer set of predictions.** The distribution offers a richer set of predictions for learning, offerring a set of auxiliary tasks which is tightly coupled to the reward.

References

- Bellemare, M. G., Dabney, W., & Munos, R. (2017, August). A distributional perspective on reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 449-458). JMLR. org.
- Rowland, M., Bellemare, M. G., Dabney, W., Munos, R., & Teh, Y. W. (2018). An analysis of categorical distributional reinforcement learning. arXiv preprint arXiv:1802.08163.