MUNI
FI

Distributional RL

Vítek Unčovský

Faculty of Informatics, Masaryk University

September 29, 2024

Papers

Both by Bellamare et. al.

1. A Distributional Perspective on Reinforcement Learning
2. Distributional Reinforcement Learning with Quantile Regression

Distributional RL

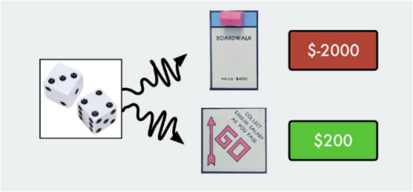
- Standard MDP formulation, but will use random state-action returns ($r(s, a)$ is a r.v.)
- The general idea - model and approximate the full return distribution $Z^\pi(s, a) = \sum_{i=0}^{\infty} r(s_i, a_i)$, where $a_i \sim \pi(\cdot \mid s_i)$ and $s_i \sim p(\cdot \mid a_{i-1}, s_{i-1})$ for all $i > 0$.
- Compare with standard RL : $Q^\pi(s, a) = \mathbb{E}(Z^\pi(s, a))$

Returns - example

Random immediate reward

Expected immediate reward

$$\mathbb{E}[R(x)] = \frac{1}{36} \times (-2000) + \frac{35}{36} \times (200) = 138.88$$



Random variable reward:

$$R(x) = \begin{cases} -2000 \text{ w.p. } 1/36 \\ 200 \text{ w.p. } 35/36 \end{cases}$$

Distributional RL - motivation?

Stability of learning:

- Multiple modes are preserved in the distributions - see the Monopoly example.
- Agent is able to learn from multiple predictions

Risk aware control, can base agents decisions off variance of $Z^{\pi}(s, a)$, etc.

Bellman Operators

$$\mathcal{T}^{\pi} Q(s,a) = \mathbb{E}\left[r(s,a) + \gamma \cdot Q(s',a')\right]$$

$$\mathcal{T} Q(s,a) = \mathbb{E}\left[r(s,a) + \gamma \cdot \max_{a'} Q(s',a')\right]$$

Lemma 1

The Bellman operators $\mathcal{T}^{\pi}$ and $\mathcal{T}$ are contractions in the max norm, i.e., for any two functions Q_1 and Q_2,

$$\|\mathcal{T}^{\pi} Q_1 - \mathcal{T}^{\pi} Q_2\|_{\infty} \leq \gamma \|Q_1 - Q_2\|_{\infty}$$

$$\|\mathcal{T} Q_1 - \mathcal{T} Q_2\|_{\infty} \leq \gamma \|Q_1 - Q_2\|_{\infty}$$

where $\|\cdot\|_{\infty}$ denotes the max norm and $0 \leq \gamma < 1$.

Distributional Bellman operator

$$\mathcal{P}^\pi Z(s,a) :\overset{D}{=} Z(S',A')^{1}$$
$$\mathcal{T}^\pi Z(s,a) \overset{D}{=} r(s,a) + \gamma \cdot \mathcal{P}^\pi Z(s,a)$$

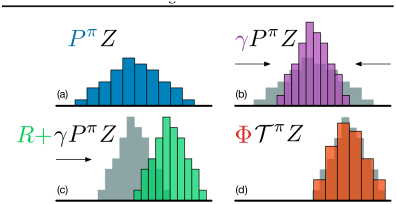


Figure 1. A distributional Bellman operator with a deterministic reward function: (a) Next state distribution under policy π, (b) Discounting shrinks the distribution towards 0, (c) The reward shifts it, and (d) Projection step (Section 4).

[1] Equality in distribution, r.v. on the left is distributed like the one on the right.

Is this operator a contraction?

And if so, in which metric?

- Q was much simpler than Z, $Z : S \times A \to \mathcal{D}(\mathbb{R})$
- The distance metric will reflect this

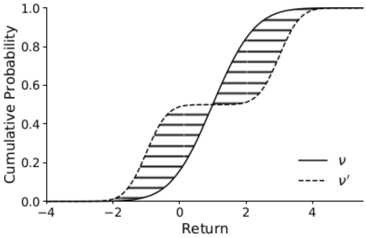
Not a contraction in d_{KL} or d_{TV} - do not reflect "closeness" of outcomes.
Consider a 1 state 1 action MDP with $r(s, a) = 0$, $Z_1 = \delta_{x=0}$, and $Z_2 = \delta_{x=1}$[2].

[2] $\delta_{x=a}$ - random variable that is a.s. equal to a.

The Wasserstein p-metric

Defined for $0 < p \leq \infty$ as:

$$d_{w_p}(Z_1, Z_2) = \left(\int_0^1 |F_{Z_1}^{-1}(u) - F_{Z_2}^{-1}(u)|^p \, du \right)^{\frac{1}{p}}$$



$\mathcal{T}^\pi$ is a contraction

Theorem 1

The distributional Bellman operator $\mathcal{T}^\pi$ is a γ-contraction in the maximum Wasserstein metric:

$$\overline{d_{w_p}}(Z_1, Z_2) = \max_{s,a} d_{w_p}(Z_1(s,a), Z_2(s,a))$$

i.e., for any two distributions $Z_1, Z_2 \in \mathcal{Z}$[a] and for any p:

$$\overline{d_{w_p}}(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) \leq \gamma \cdot \overline{d_{w_p}}(Z_1, Z_2)$$

[a] return distributions on S x A that have bounded moments

From Banach we get convergence to unique fixpoint $Z^\pi(s,a)$ as a limit of $(\mathcal{T}^\pi)^k Z_0$ for any Z_0.

Optimality operator

Notion of optimality remains the same - optimal policies maximize ***expected*** reward $Q(s, a)$. The goal is to find Z^{π^*} of an optimal (stationary) policy[3].

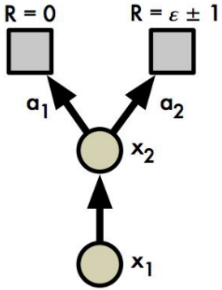
$$\mathcal{T}Z(s, a) \stackrel{D}{=} r(s, a) + \gamma \cdot Z(s', \pi^G(s'))$$

Where π^G is any greedy policy for Z. Note that the choice matters here, unlike in standard RL - possibly many different operators.

[3] While all optimal policies share the same Q, their return distributions may be different

Issues

The dist. opt. Bellman operator is not smooth



Consider distributions Z_ϵ

If $\varepsilon > 0$ we back up a bimodal distribution

If $\varepsilon < 0$ we back up a Dirac in 0

Thus the map $Z_\epsilon \mapsto TZ_\epsilon$ is not continuous

Properties of $\mathcal{T}$

1. Not a contraction w.r.t. any metric, see previous example.
2. May not have a fixed point - alternating between optimal actions.
3. Even if it does have a unique fixpoint corresponding to optimal stationary policy, you may not converge to it.

On a brighter note, since the update is greedy, you preserve the contractivity property of $\mathcal{T}Q$. This means that if optimal stationary policy is unique, you do have guaranteed convergence to its return.

C51

Algorithm 1 Categorical Algorithm

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

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

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

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

for $j \in 0, \dots, N-1$ **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

C51

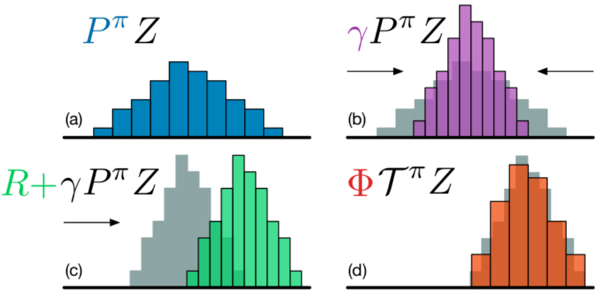
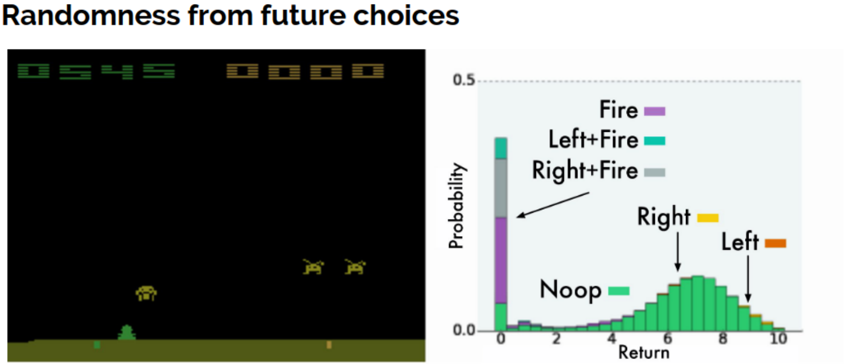


Figure 1. A distributional Bellman operator with a deterministic reward function: (a) Next state distribution under policy π, (b) Discounting shrinks the distribution towards 0, (c) The reward shifts it, and (d) Projection step (Section 4).

Learned distributions



Results

Beats DQN with several upgrades, without much tuning. Does really well on sparse reward tasks (still 0 reward on Montezuma's revenge though).

	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%	-	-

Why does C51 not optimize Wasserstein directly?

Property of unbiased sample gradients. [4] Sample gradients of Wasser. are biased, in a sense that $\mathbb{E}_{i \sim I} \nabla_\theta d_{w_p}(P_i, Q_\theta) \neq \nabla_\theta d_{w_p}(P_I, Q_\theta)$, where P_I is a mixture random variable.
In general, by optimizing the sample loss you get to a different minimum. The above makes Wasser. unfit for SGD optimisation.

[4] https://arxiv.org/abs/1705.10743

Quantile Regression Q-Learning

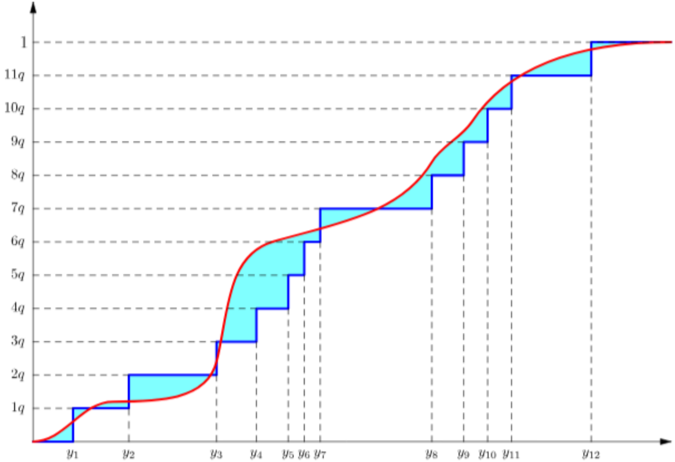
Instead of fixed support and optimized probabilities, consider the opposite. Approximate quantiles of Z:

$$Z_\theta(s,a) = \frac{1}{N} \cdot \sum_{i=0}^{N} \delta_{x=\theta_i(s,a)}$$

Avoids the tedious projection Φ onto support that was used in C51, also much more flexible. Will also enable us to approximately minimize Wasser. instead of KL as in C51.

Projection into quantile distribution

Finding a quantile distribution that minimizes d_{w_1} from target is pretty simple, since the quantile distributions F are step functions.



Solving the projection

Denote $\tau_i = \frac{i}{N}$, and (ordered) support of Z_θ as $\{\theta_1, \ldots, \theta_N\}$

$$d_W(Z_\theta, Z) = \sum_{i=1}^{N} \int_{\tau_{i-1}}^{\tau_i} \mid F_Z^{-1}(u) - \theta_i \mid$$

How does the projection onto the space of quantile distributions look like?

Solving the projection

Denote $\tau_i = \frac{i}{N}$, and (ordered) support of Z_θ as $\{\theta_1, \ldots, \theta_N\}$

$$d_{W_1}(Z_\theta, Z) = \sum_{i=1}^{N} \int_{\tau_{i-1}}^{\tau_i} \mid F_Z^{-1}(u) - \theta_i \mid$$

Lemma 2

The quantile distribution that minimizes 1-Wasserstein distance from Z, denoted $\Pi_W Z$ has the support $\theta_i = F_Z^{-1}(\overline{\tau_i})$, where $\overline{\tau_i} = \frac{\tau_{i-1}+\tau_i}{2}$. If F_Z is continuous, it is the unique minimizer.

This convenient form of the projection in terms of quantiles of Z will enable us to sidestep the biased gradient issues w. Wasserstein distance.

Quantile Regression

The minimum of the following loss is the τ quantile of Z:

$$\mathcal{L}_\tau(\theta) = \mathbb{E}_{\hat{Z} \sim Z} \left[\rho_\tau(\hat{Z} - \theta) \right]$$

where

$$\rho_\tau(u) = u \cdot (\tau - \delta_{x<0})$$

Penalize positive and negative distance of samples from θ by different weights, depending on the desired quantile.
Crucially, we can optimize this loss via SGD and take steps towards $\hat{\mathcal{T}} Z_\theta$ w.r.t. the Wasserstein metric.

Quantile Huber Loss

Actually, the final loss used in the algorithm incorporates the Huber loss, because of $x = 0$:

$$L_\kappa(x) = \begin{cases} \frac{1}{2}x^2, & \text{if } |x| \leq \kappa \\ \kappa(|x| - \frac{\kappa}{2}), & \text{otherwise} \end{cases}$$

Quantile Huber Loss is the asymmetric variant of the Huber loss:

$$\rho_\tau^\kappa(x) = |\tau - \delta_{x<0}| \cdot L_\kappa(x)$$

Quantile Huber Loss

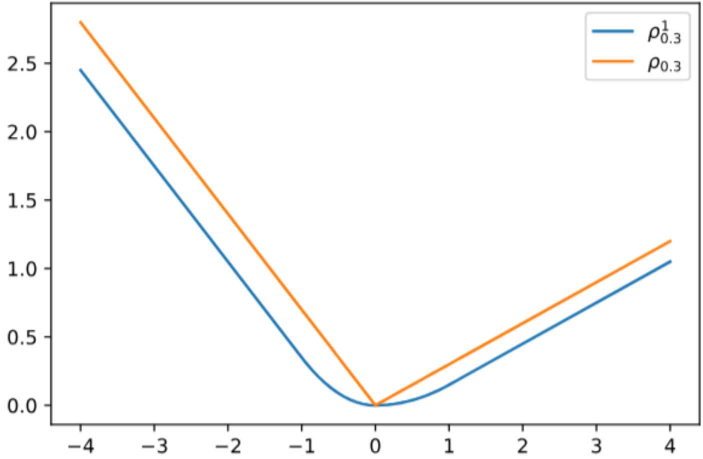


Figure 8.

Properties

Theorem 2

The distributional Bellman operator combined with projection to quantile distributions - $\Pi_W \mathcal{T}^\pi$ is a γ-contraction in the maximum Wasserstein metric for $p = \infty$:

$$\overline{d_{W_\infty}}(\Pi_W \mathcal{T}^\pi Z_1, \Pi_W \mathcal{T}^\pi Z_2) \leq \gamma \cdot \overline{d_{W_\infty}}(Z_1, Z_2)$$

So we again get guaranteed convergence to a fixed point - a quantile approximation of Z^π.

Algorithm

Algorithm 1 Quantile Regression Q-Learning

Require: N, κ

input $x, a, r, x', \gamma \in [0, 1)$

  # Compute distributional Bellman target

  $Q(x', a') := \sum_j q_j \theta_j(x', a')$

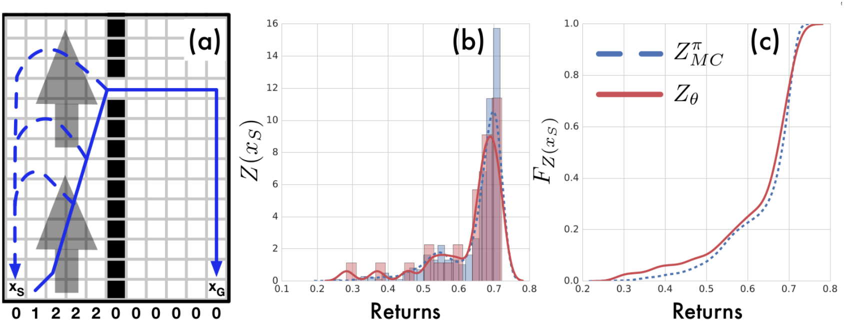
  $a^* \leftarrow \arg\max_{a'} Q(x, a')$

  $\mathcal{T}\theta_j \leftarrow r + \gamma\theta_j(x', a^*), \quad \forall j$

  # Compute quantile regression loss (Equation 10)

output $\sum_{i=1}^{N} \mathbb{E}_j \left[\rho^{\kappa}_{\hat{\tau}_i}(\mathcal{T}\theta_j - \theta_i(x, a))\right]$

Approximate Returns



Results

	Mean	Median	>human	>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
QR-DQN-0	881%	199%	38	52
QR-DQN-1	**915%**	**211%**	**41**	**54**