MATHEMATICAL TECHNIQUES FOR MACHINE LEARNING

A distributional perspective on Reinforcement Learning

By
Ahmed JAMOUSSI

Fall 2019

Abstract

This project consists in analyzing and formulating some critics to a recently published paper in the field of machine learning. We chose *A Distributional Perspective on Reinforcement Learning* [**?**], which aims at improving the learning of agents. The authors argue for the importance of the value distribution: the distribution of the random return received by a reinforcement learning agent. They give some theoretical results in the policy evaluation and in the control setting. Even though there is an instability in the distributional version of Bellman's optimality equation, they show that there are many benefits to learn an approximate value distribution rather than its usual approximated expectation.

1 Introduction

The reinforcement learning (RL) is a very active part of the learning algorithms field. It consists in building programs which learn how to predict and act in an environment, based on past experience. The training of the artificial agents is the main challenge in the confection of such algorithms. The most used approach for solving these problems is value-based RL, in which the agent predicts the *expected-return* in order to optimize its actions and its behaviour in the environment. One of the limitations of such an approach, is that an expected value based method outputs an average return and not the actual one. Hence, the motivation of studying the actual distribution of returns instead of restricting the learning on its expectation. Distributional approaches to value-based RL comes into the pictures to model the entire distribution of return, rather that just its expected value.

2 Setting

An artificial agent learns by acting in a possibly stochastic environment and receiving rewards corresponding to its acts. An agent must learn from his interactions with the environment in order to maximize his cumulative rewards.

Definition 2.1. *A **Markov Decision Process** is a 5-tuple* $(\mathcal{X}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$ *where:*

- $\mathcal{X}$ *is the state space of the environment.*
- $\mathcal{A}$ *is the action space of the agent.*
- $\mathcal{R} : \mathcal{X} \times \mathcal{A} \to R^{|\mathcal{X}|}$ *is the reward function, where* R *is a continuous set of possible rewards.*
- $\mathcal{P} : \mathcal{X} \times \mathcal{A} \to [0,1]^{|\mathcal{X}|}$ *is the transition function. It gives the conditional transition probabilities between states given an action.*
- $\gamma \in [0,1]$ *is the discount factor. It is used to encourage the agent to take actions early instead of postponing them.*

Definition 2.2. *A **stationary policy*** $\pi : \mathcal{X} \to [0,1]^{|\mathcal{A}|}$ *is a function that maps each state* $x \in \mathcal{X}$ *to a probability distribution over* $\mathcal{A}$ *and that is independent of time. The set of all stationary policies is* Π.

Definition 2.3. *Given a policy π along with a state x, the* ***return*** *Z^π is the sum of the discounted rewards along the agent's trajectory of interactions with the environment starting at x:*

$$Z^\pi(x) := \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t)$$
$$x_t \sim P(\cdot|x_{t-1}, a_{t-1}), x_0 = x, a_t \sim \pi(\cdot|x_t)$$

We can now easily see the importance of γ. Since $\gamma^t \xrightarrow{t\to\infty} 0$, the agent is forced to give more importance to the upcoming actions in contrast to actions further in time. The value function of a policy π is the expected return from taking action $a \in \mathcal{A}$ from state $x \in \mathcal{X}$ and then acting according to π:

$$Q^\pi(x, a) := \mathbb{E}[Z^\pi(x)] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(x_t, a_t)\right]$$
$$x_t \sim P(\cdot|x_{t-1}, a_{t-1}), x_0 = x, \forall t > 0\ a_t \sim \pi(\cdot|x_t), a_0 = a$$

We can also write it the following way:

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

Definition 2.4. *An* ***optimal policy*** *$\pi^\star$ is a stationary policy that maximizes the expected return. The set of all optimal policies is $\Pi^\star$.*

From the optimal value function $Q^\star$, we can directly find an optimal policy $\pi^\star$ from a given state x:

$$Q^\star(x, a) := \max_{\pi\in\Pi} Q^\pi(x, a) \implies \pi^\star(x) = \arg\max_{a\in\mathcal{A}} Q^\star(x, a)$$

The most common approach to reach an optimal policy $\pi^\star$ is via Q-learning which makes use of the Bellman operator.

Definition 2.5. *The* ***Bellman operator*** *$\mathcal{T}^\pi$ and the* ***Bellman optimality operator*** *$\mathcal{T}$ are defined as follows:*

$$(\mathcal{T}^\pi K)(x, a) := \mathbb{E}[R(x, a)] + \gamma \cdot \mathbb{E}_{P,\pi}[K(x', a')]$$
$$(\mathcal{T} K)(x, a) := \mathbb{E}[R(x, a)] + \gamma \cdot \mathbb{E}_P\left[\max_{a'\in\mathcal{A}} K(x', a')\right]$$

We note that $Q^\star$ is a fixed point to the Bellman operator. We can prove that this is the unique fixed point and that the Bellman operator is a contraction. The Q-learning algorithm applies repeatedly the Bellman operator to some initial $Q(x, a)$. The convergence to $Q^\star$ follows.

In this article, the authors want to make use of the the distributional version of the value function.

Definition 2.6. *The* ***value function*** *V^π of a policy $\pi \in \Pi$ is defined as follows:*

$$V^\pi(x) = \mathbb{E}_{a\sim\pi}[Q^\pi(x, a)] = \sum_{a\in\mathcal{A}} \pi(a|x)\mathbb{E}[Q^\pi(x, a)]$$

3 The Distributional Bellman Operator

As explained above, we are interested in the distributional perspective of the value function.

Definition 3.1. *The **value distribution** Z^π is a map from state-action pairs to a distribution $\widehat{\mathcal{R}}$ over the returns:*

$$Z^\pi : \mathcal{X} \times \mathcal{A} \to \widehat{\mathcal{R}}$$

*We will also need to use the **space of value distribution with bounded moments**, which we denote $\mathcal{Z}$.*

In this section, we will introduce the Wasserstein metric. Then, we will define the distributional analogues of the Bellman operators in order to show some theoretical results about the behaviour of those operators on the value distribution in the policy evaluation setting. The choice of metric is very important because the results we will show do not apply with any choice of metric (why the introduction about the Wasserstein metric). Furthermore, in the distributional version of the control setting, we will define the notion of an optimal value distribution, and show some instability in the search for an optimal value distribution using the distributional analogues of the Bellman operators.

3.1 The Wasserstein Metric

Before getting in our analysis (*Policy Evaluation* and *Control setting*), we need to define the Wasserstein metric, a major tool for our analysis. The Wasserstein metric is denoted as d_p and is defined as the following:
Let $p > 0$ and F, G two real valued cumulative distribution functions (c.d.f.). We define d_p as follows:

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

Note that the infimum is taken over all pairs of random variable (U, V) with c.d.fs F and G respectively. In addition, note that there's an equivalent definition for d_p using the inverse c.d.f transformations F^{-1}, G^{-1} from uniformly distributed random variable $\mathcal{U}$ on $[0, 1]$:

$$d_p(F, G) := \|F^{-1}(\mathcal{U}) - G^{-1}(\mathcal{U})\|_p$$

If p is finite we then can explicitly write d_p as the following:

$$d_p(F, G) := \left[\mathbb{E}|F^{-1}(\mathcal{U}) - G^{-1}(\mathcal{U})|^p\right]^{1/p} = \left[\int_{[0,1]} |F^{-1}(u) - G^{-1}(u)|^p du\right]^{1/p}$$

Given two random variables U, V with c.d.fs F_U, F_V respectively we can equivalently define d_p as

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

Finally, in order to study and analyse the behavior of the Bellman operator, we extend the metric d_p to vectors of random variables (such as value distributions) using the L_p norm. For two value distributions $Z_1, Z_2 \in \mathcal{Z}$, we use the maximal form of the Wasserstein metric to extend d_p:

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

Theorem 3.2. *$\overline{d}_p$ is a metric over value distributions.*

We begin by stating few helpful results to prove the theorem. Consider a scalar α and a random variable A independent of U and V. Note that d_p metric has the following properties:

$$d_p(\alpha U, \alpha V) \leq |\alpha| \; d_p(U, V) \tag{1}$$

$$d_p(A + U, A + V) \leq d_p(U, V) \tag{2}$$

$$d_p(AU, AV) \leq \|A\|_p \; d_p(U, V) \tag{3}$$

Lemma 3.3. *Let $A_1, A_2, A_3, \ldots$ be a set of random variables describing a partition of Ω , i.e $A_i(\omega) \in \{0, 1\}$ and for any ω there is exactly A_i with $A_i(\omega) = 1$.*
Let U, V be two random variables then

$$d_p(U, V) \leq \sum_{i \geq 1} d_p(A_i U, A_i U)$$

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

(i) (identity) Since d_p is a metric then $d_p(Z_1(x, (a), Z_1(x, a)) = 0, \; \forall x, a$. Thus,

$$\overline{d}_p(Z_1, Z_1) = \sup_{x,a} d_p(Z_1(x, a), Z_1(x, a)) = 0$$

(ii) (non-negativity) Since d_p is a metric then $d_p(Z_1(x, (a), Z_2(x, a)) \geq 0, \; \forall x, a$. Thus,

$$\begin{aligned} \overline{d}_p(Z_1, Z_2) &= \sup_{x,a} d_p(Z_1(x, a), Z_2(x, a)) \\ &\geq d_p(Z_1(x, a), Z_2(x, a)) \\ &\geq 0 \end{aligned}$$

(iii) (symmetry) Since d_p is a metric then $d_p(Z_1(x, a), Z_2(x, a)) = d_p(Z_2(x, a), Z_1(x, a)), \; \forall x, a$. Thus,

$$\begin{aligned} \overline{d}_p(Z_1, Z_2) &= \sup_{x,a} d_p(Z_1(x, a), Z_2(x, a)) \\ &= \sup_{x,a} d_p(Z_2(x, a), Z_1(x, a)) \\ &= \overline{d}_p(Z_2, Z_1) \end{aligned}$$

(iv) (triangle inequality) Let $Y \in \mathcal{Z}$.
Since d_p is a metric then $d_p(Z_1(x, a), Z_2(x, a)) \leq d_p(Z_1(x, a), Y) + d_p(Y, Z_2(x, a)), \; \forall x, a$. Thus,

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

□

3.2 Policy Evaluation

In this section, we are given a policy π, and we are interested in its value distribution Z^π.

Definition 3.4. *We define the* ***transition operator*** $P^\pi : \mathcal{Z} \to \mathcal{Z}$ *as follows:*

$$P^\pi Z(x,a) := Z(x',a')$$
$$x' \sim P(\cdot|x,a), a' \sim \pi(\cdot|X')$$

Definition 3.5. *From now on, we redefine* $\mathcal{T}^\pi$ *to be the* ***distributional Bellman operator*** $\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$*. We now view the reward function as a random vector* $\mathcal{R} \in \mathcal{Z}$ *and we define* $\mathcal{T}^\pi$ *as follows:*

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

It is worth to note that this operator is different from the usual Bellman operator. It contains 3 sources of randomness: the reward $\mathcal{R}$*, the transition operator* P^π*, and the next-state value distribution* $\mathcal{Z}(x',a')$*. We assume that these 3 quantities are independent.*

The most important result in this setting is coming from the following theorem:

Theorem 3.6. $\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ *is a* γ*-contraction in* $\overline{d}_p$*. In other words:*

$$\forall Z_1, Z_2 \in \mathcal{Z}, \overline{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) \leq \gamma \overline{d}_p(Z_1, Z_2)$$

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

$$\begin{aligned}
\overline{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) &= \sup_{x,a} d_p(\mathcal{T}^\pi Z_1(x,a), \mathcal{T}^\pi Z_2(x,a)) && \text{by definition of } \overline{d}_p \\
&= \sup_{x,a} d_p(\mathcal{R}(x,a) + \gamma P^\pi Z_1(x,a), \mathcal{R}(x,a) + \gamma P^\pi Z_2(x,a)) && \text{by definition of } \mathcal{T}^\pi \\
&\leq \gamma \sup_{x,a} d_p(P^\pi Z_1(x,a), P^\pi Z_2(x,a)) \\
&= \gamma \sup_{x',a'} d_p(Z_1(x',a'), Z_2(x',a')) && \text{by definition of } P^\pi \\
&= \gamma \overline{d}_p(Z_1, Z_2) && \text{by definition of } \overline{d}_p
\end{aligned}$$

□

Now, using Banach's fixed point theorem, we know that the contraction mapping $\mathcal{T}^\pi$ has a unique fixed point. By inspection, we find that this fixed point is the random return $\sum_{t=0}^\infty \gamma^t R(x_t, a_t)$. It follows that the sequence $\{Z_k\}_{k\in\mathbb{N}} \xrightarrow{k\to\infty} \sum_{t=0}^\infty \gamma^t R(x_t,a_t)$ in $\overline{d}_p$ where $\forall k \geq 1, Z_{k+1} = \mathcal{T}^\pi Z_k, Z_0 \in \mathcal{Z}$. This result is important since it allows us to obtain the value distribution corresponding to a given optimal policy.

Discussion

In the use of Banach's fixed point theorem, it is not stated that $(\mathcal{Z}, \overline{d}_p)$ is a complete metric space. However, this is an important condition because the failure to fulfill this condition can lead to not having a unique fixed point from the contraction mapping. We provide an example. Let

$0 < a < b \in \mathbb{R}$ and $X = ((a, b), d)$ with $d = |\cdot|$ the Euclidean distance. X is not complete since it is an open subset of $\mathbb{R}$. We consider the map $T : X \to X, x \mapsto \frac{x}{2}$. T is clearly a contraction since:

$$|T(x) - T(y)| = \frac{1}{2}|x - y|$$

However the fixed point of T is clearly at $0 \notin X$. At first, we thought that the fact that $(\mathcal{Z}, \overline{d}_p)$ is complete was trivial, but it is not. In the appendix, we give a proof of the completeness of $(\mathcal{Z}, \overline{d}_p)$, mainly inspired by [**?**] who proves it for $(\mathcal{Z}, d_p)$.

3.3 Control Setting

In this section, we are seeking a policy π that maximizes value. We will first define the notion of an optimal value distribution and redefine $\mathcal{T}$ to be the distributional optimality Bellman operator. Then, we will show that we cannot proceed as in the policy evaluation case in order to converge to an optimal distribution using repeatedly $\mathcal{T}$.

Definition 3.7. *An **optimal value distribution** is the value distribution of an optimal policy. The set of optimal value distributions is* $\mathcal{Z}^\star := \{Z^{\pi^\star} | \pi^\star \in \Pi^\star\}$.

It is important to understand the contrast between optimal policies and optimal value distributions. While all optimal policies attain the same value $Q^\star$, there are many optimal value distributions which all have the same expectation $Q^\star$. Also, not all value distribution with expectation $Q^\star$ are optimal: they must match the distribution of the return under some optimal policy.

Definition 3.8. *A **greedy policy** π for $Z \in \mathcal{Z}$ maximizes $\mathbb{E}[Z]$. The set $\mathcal{G}_Z$ of greedy policies for Z is defined as follows:*

$$\mathcal{G}_Z := \left\{ \pi | \sum_{a \in \mathcal{A}} \pi(a|x)\mathbb{E}[Z(x, a)] = \max_{a' \in \mathcal{A}} \mathbb{E}\left[Z\left(x, a'\right)\right] \right\}$$

Definition 3.9. *From now on, we redefine $\mathcal{T}$ to be the **distributional optimality Bellman operator** $\mathcal{T} : \mathcal{Z} \to \mathcal{Z}$. It is defined as any operator which implements a greedy selection rule. In other words, $\mathcal{T}$ is defined as follows:*

$$\mathcal{T}Z(x, a) = \mathcal{T}^\pi Z(x, a) \text{ for some } \pi \in \mathcal{G}_Z$$

The first result we present was expected.

Lemma 3.10. *Let $Z_1, Z_2 \in \mathcal{Z}$. Then:*

$$||\mathbb{E}[\mathcal{T}Z_1] - \mathbb{E}[\mathcal{T}Z_2]||_\infty \leq ||\mathbb{E}[Z_1] - \mathbb{E}[Z_2]||_\infty$$

This just means that the sequence $\{\mathbb{E}[Z_k]\}_{k \in \mathbb{N}} \xrightarrow{k \to \infty} Q^\star$ where $\forall k \geq 1, Z_{k+1} = \mathcal{T}Z_k, Z_0 \in \mathcal{Z}$. However, through an example, we will show multiple results that will demonstrate the instability in this control setting, and especially that $\{Z_k\}_{k \in \mathbb{N}}$ does not converge in $\overline{d}_p$ to a fixed point in $\mathcal{Z}^\star$.

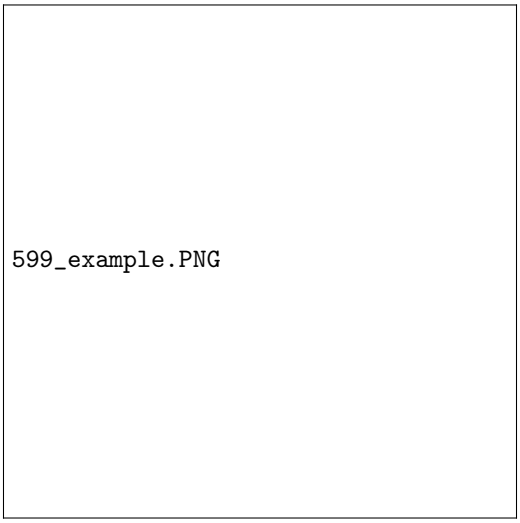

Figure 1: Undiscounted two-state Markov Decision Process. The rewards for choosing a_2 are of equal probability. The yellow entries are the one contributing to $\overline{d}_1(Z, Z^\star)$ and $\overline{d}_1(\mathcal{T}Z, \mathcal{T}Z^\star)$

Proposition 3.11. *$\mathcal{T}$ is not a contraction.*

Proof. We let $1 > \varepsilon > 0$. In this example from Figure 1, it is easy to see that there is a unique optimal policy $\pi^\star$, which is choosing action a_2, since the expected return would be of $\varepsilon > 0$, compared to an (expected) return of 0 for choosing action a_1. Hence, the unique optimal value distribution $Z^\star = Z^{\pi^\star}$ is well defined in the table. Now, since $0 > \frac{1}{2}(-\varepsilon + 1) + \frac{1}{2}(-\varepsilon - 1)$, we get that $\mathcal{T}Z(x_1) = 0$. We can now compute $\overline{d}_1(Z, Z^\star)$ and $\overline{d}_1(\mathcal{T}Z, \mathcal{T}Z^\star)$:

$$
\begin{aligned}
\overline{d}_1(Z, Z^\star) &= d_1(Z(x_2, a_2), Z^\star(x_2, a_2)) && \text{since } Z \text{ and } Z^\star \text{ only differ at } (x_2, a_2) \\
&= |-\varepsilon \pm 1 - (\varepsilon \pm 1)| \\
&= 2\varepsilon \\
\overline{d}_1(\mathcal{T}Z, \mathcal{T}Z^\star) &= d_1(\mathcal{T}Z, Z^\star) && \text{since } Z^\star \text{ is a fixed point under } \mathcal{T} \\
&= d_1(\mathcal{T}Z(x_1), Z^\star(x_1)) && \text{since } \mathcal{T}Z \text{ and } Z^\star \text{ only differ at } x_1 \\
&= \mathbb{E}[|\mathcal{T}Z(x_1) - Z^\star(x_1)|] && \text{by definition of } d_1 \\
&= \frac{1}{2}|1 - \varepsilon| + \frac{1}{2}|1 + \varepsilon| \\
&= 1 && \text{since } 0 < \varepsilon < 1
\end{aligned}
$$

It follows that $\forall \varepsilon < \frac{1}{2}$, and for $\gamma < 1$:

$$
\overline{d}_1(\mathcal{T}Z, \mathcal{T}Z^\star) > \overline{d}_1(Z, Z^\star) \implies \overline{d}_1(\mathcal{T}Z, \mathcal{T}Z^\star) > \gamma \overline{d}_1(Z, Z^\star)
$$

Which means that $\mathcal{T}$ is not a γ-contraction. □

Furthermore, it could be shown that $\mathcal{T}$ is not a contraction in any metric which separates Z and $\mathcal{T}Z$.

Proposition 3.12. *Not all distributional optimality Bellman operators $\mathcal{T}$ have a fixed point $Z^\star = \mathcal{T}Z^\star$.*

Proof. Let $\varepsilon = 0$. Then in terms of expected return, the actions a_1 and a_2 are equivalent. Hence the following selection rule from a value distribution Z: pick a_2 if $Z(x_1) = 0$, and a_1 otherwise, is a greedy selection rule because it maximizes the expectation of Z (which is always 0). We let $\mathcal{T}$ be an optimality operator implementing this greedy rule. Now, the sequence $\{(\mathcal{T})^k Z^\star(x_1)\}_{k\in\mathbb{N}}$ alternates between $Z^\star(x_2, a_1) = 0$ and $Z^\star(x_2, a_2) = \pm 1$. Hence, this operator does not have any fixed point. □

Proposition 3.13. *Even if a distributional optimality Bellman operators $\mathcal{T}$ has a fixed point $Z^\star = \mathcal{T}Z^\star$, it is not sufficient to guarantee the convergence of $\{Z_k\}_{k\in\mathbb{N}}$ to $\mathcal{Z}^\star$.*

While those results are disappointing, it is shown that the sequence $\{Z_k\}_{k\in\mathbb{N}}$ converges to a larger set than $\mathcal{Z}$. Indeed, under many conditions, we obtain that $\{Z_k\}_{k\in\mathbb{N}}$ converges to the set of non-stationary optimal value distributions.

4 Approximate Distributional learning

In this section we will present the proposed algorithm (*Categorical Algorithm*) based on the distributional Bellman operator introduced above. The algorithm requires choosing a predefined value distribution.

4.1 Parametric Distribution

In his paper, the authors chose to model the value distribution using a parametric discrete distribution. More specifically, the chosen distribution is parametrized by $N \in \mathbb{N}$ and $V_{min}, V_{max} \in \mathbb{R}$, and whose support is the the following set of points:

$$\mathcal{S} = \{z_i = V_{min} + i\Delta Z \ : \ i \in [N]\}, \text{ with } \Delta z = \tfrac{V_{max} - V_{min}}{N-1}.$$

Theoretically, the points in $\mathcal{S}$ are the *canonical returns* of our value distribution. The probability of the even $\{Z_\theta(x, a) = z_i\}$ is given by a parametric model $\theta : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^N$ as the the follows

$$p_i(x, a) = \mathrm{P}\{Z_\theta(x, a) = z_i\} = \frac{e^{\theta_i(x,a)}}{\sum_{j\in[N]} e^{\theta_j(x,a)}}$$

Note that a discrete distribution has the advantage of being very expressive and not computationally expensive at the same time.

4.2 Categorical Algorithm

Algorithm 1 Categorical Algorithm

Input: A transition $x_t, a_t, r_t, x_{t+1}, \gamma_t \in [0, 1]$
$Q(x_{t+1}, a) := \sum_{i=0}^{N-1} z_i p_i(x_{t+1}, a_t)$
$a^* \leftarrow \arg\max_a Q(x_{t+1}, a)$
$m_i = 0,\ i \in 0, \ldots, N-1$
for $j \in 0, \ldots, N-1$ **do**
  $\hat{\mathcal{T}} z_j \leftarrow [r_t + \gamma_t z_j]_{V_{min}}^{V_{max}}$ ▷ Compute the projection of $\hat{\mathcal{T}} z_j$ onto the support z_j
  $b_j \leftarrow \frac{\hat{\mathcal{T}} z_i - V_{min}}{\Delta Z}$ ▷ $b_j \in [0, N-1]$
  $l \leftarrow \lfloor b_j \rfloor,\ u \leftarrow \lceil b_j \rceil$
  $m_l \leftarrow m_l + p_i(x_{t+1}, a^*)(u - b_i)$ ▷ Distribute probability of $\hat{\mathcal{T}} z_j$
  $m_u \leftarrow m_u + p_i(x_{t+1}, a^*)(b_i - l)$
  end for
Output: $-\sum_{i \in N} m_i \log(p_i(x_t, a_t))$ ▷ Cross-entropy loss

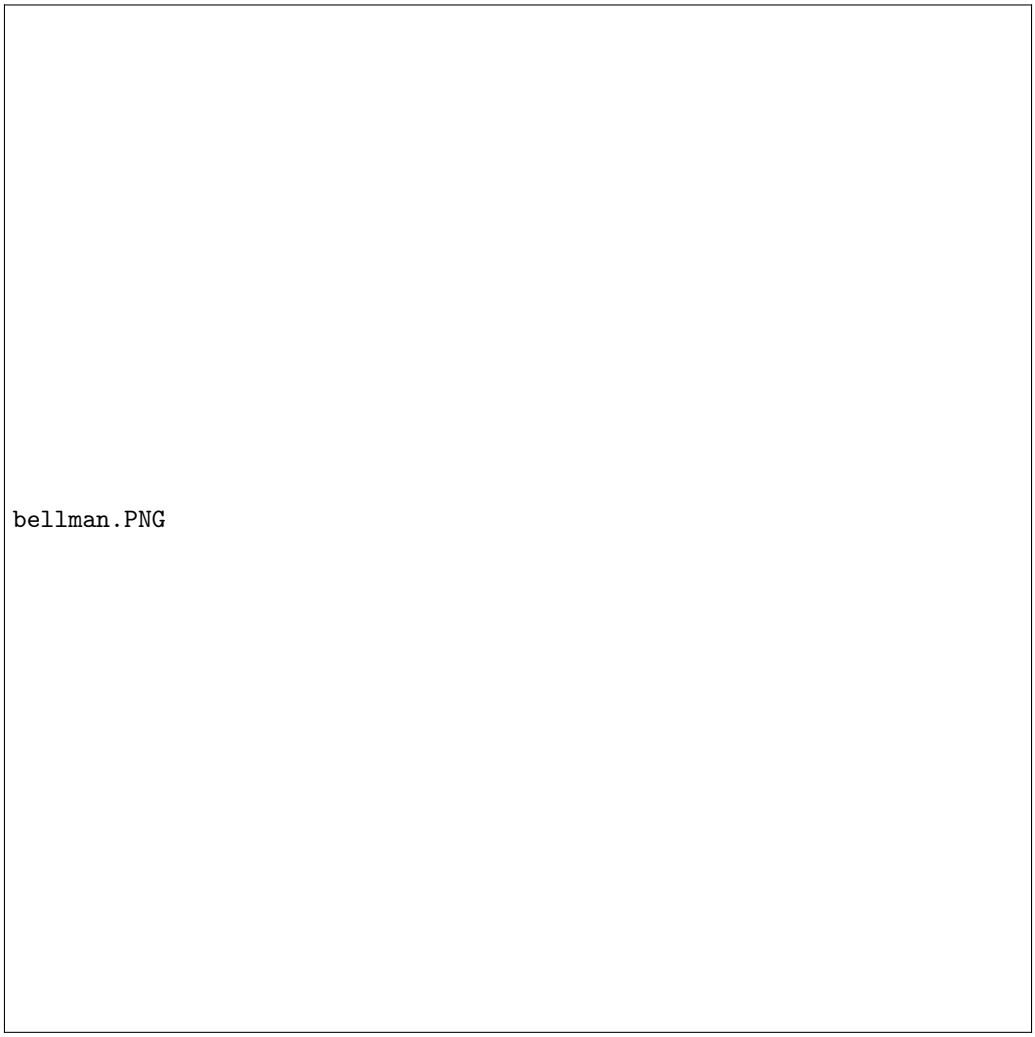

Figure 2: The steps realized by a distributional Bellman operator in the Categorical Algorithm

In this section we discuss and explain the *Categorical Algorithm* stated above.
Firstly, we describe the inputs:

- x_t : state at time t.
- x_{t+1}: next state at time $t+1$.
- a_t: next state at time $t+1$.
- r_t: return at time t.
- $p_i(x, a)$: represent the probability that the return of the action a at the state x is equal to z_i.

We assume that the possible returns are between V_{min} and V_{max} and then we choose $N \in \mathbb{N}$ to be the number of bins in the interval $[V_{min}, V_{max}]$. (i.e the $N+1$ points z_i). For each action $a \in \mathcal{A}$ we compute the expected reward $Q(x_{t+1}, a) = \sum_{i=0}^{N-1} z_i p_i(x_{t+1}, a_t)$. Afterwards, we define a^* to be the action with the maximal expected reward $Q(x_{t+1}, a^*) = \max_{a \in \mathcal{A}} Q(x_{t+1}, a)$. Hence, highlighting a greedy policy π. After choosing the best action a^* we apply the Bellman operator on the value distribution $Z(x_{t+1}, a^*)$. This is done by following steps:

1- Shrink the distribution by the discount factor $\gamma_t \in [0, 1]$. *(Figure 1,(b))*

2- Shift the distribution by the reward amount r_t. *(Figure 1,(c))*

Since the resulting bins $(\hat{\mathcal{T}} z_i, \forall i \in [N])$ do not line up with the original bins $(z_i, \forall i \in [N])$ then we cannot compute the loss function, which is the Wasserstein loss $l_{x,a}(\theta) = \overline{d}_p(\hat{\mathcal{T}} Z(x, a), Z_\theta(x, a))$ with $p > 0$. As a solution, we project the resulting distribution $(\hat{\mathcal{T}} z_i, \forall i \in [N])$ on the setting of the original distribution using an interpolation *(Figure 2(d))*. This is done by the following steps:

1- for each $i \in [N]$, we define b_i as $\frac{\hat{\mathcal{T}} z_i - V_{min}}{\Delta Z}$ with $\Delta Z = \frac{V_{max} - V_{min}}{N-1}$. Note that each $b_i \in [0, N_1]$.

2- for each $i \in [N]$, we define l to be the lower bound of b_i and u to be the upper bound of b_i. Note that since $b_i \in [0, N_1]$ then each $l, u \in \{0, \ldots, N-1\}$.

3- Finally, in order to distribute the probability of $\hat{\mathcal{T}} Z(x, a)$, we assign $m_l + p_i(x_{t+1}, a^*)(u - b_i)$ to m_l and $m_u + p_i(x_{t+1}, a^*)(b_i - l)$ to m_u for each $i \in [N]$.

Note that we can use other methods to obtain a better projection than the proposed interpolation. After computing the projection, we are now able to compute sample loss, chosen to be the cross entropy term of the KL divergence:

$$\mathcal{L}_{x,a}(\theta) = H(\hat{\mathcal{T}} Z(x, a), Z_\theta(x, a)) = -\sum_{i \in N} m_i \log(p_i(x_t, a_t))$$

We can now apply gradient descent in order to find the best parameters θ that minimize $\mathcal{L}_{x,a}(\theta)$ to get the best approximation $Z_\theta(x, a)$ of the actual value distribution $Z(x, a)$, for each $(x, a) \in \mathcal{X} \times \mathcal{A}$.

5 Conclusion

On the one hand, after the analysis of this paper, we see that most of the theoretical results about the distributional optimality Bellman operator $\mathcal{T}$ are not very promising. Indeed, it has been shown that $\mathcal{T}$ is not a contraction in any metric, which is a major drawback since this does not ensure the convergence to $Z^\star$, an optimal value distribution. However, the best result that we obtained so far is that $\mathcal{T}$ ensures convergence to the set of non-stationary optimal value distributions. Unfortunately, such a result is not an improvement since the usual Bellman operator provides the same result.
On the other hand, it has been shown in part 5 of the article that the use of C51, a specific version of the Caterogical algorithm (with N=51), outperforms a human player and DQN, one of the best RL algorithms nowadays.
Further improvements of the proposed approach can be made on the approximation method used to find the optimal parameters θ used to define the approximated distribution of $Z^\star$.
To sum up, the paper does not highlight any significant theoretical improvement over non-distributional approaches (usual Bellman operator). However, the use of the algorithm led to better performance on a practical side.

6 Appendix

Theorem 6.1. *$(\mathcal{Z}, \overline{d}_p)$ is complete.*

Proof. Denote $\mathcal{Z}_p$ to be the set of probability measures (value distribution) with finite p^{th} moment and let $(\mu_n^{(x,a)})_{n\in\mathbb{N}}$ be a Cauchy sequence of probability measures (probability distribution) in $\mathcal{Z}_p$, of taking action $a \in \mathcal{A}$ in state $x \in \mathcal{X}$. To make notation lighter, we will write μ_n instead of $\mu_n^{(x,a)}$.

Step 0: Show $(\mathbb{R}, |\cdot|)$ is separable
Since $\mathbb{Q}$ is a countable, dense subset in $\mathbb{R}$, we have that $(\mathbb{R}, |\cdot|)$ is separable.

Step 1: We prove that $(\mu_n)_n$ is uniformly tight for $p \geq 1$.

Let $\varepsilon > 0$. We note that $(\mu_n)_n$ is Cauchy in $(\mathcal{Z}_1, \overline{d}_1)$ since $\overline{d}_1 \leq \overline{d}_p$. Thus, $\exists N \in \mathbb{N}$ such that $\forall n, m \geq N$, we have $\overline{d}_1(\mu_n, \mu_m) \leq \varepsilon^2$ so that $\overline{d}_1(\mu_n, \mu_N) \leq \varepsilon^2$. Hence, $\forall n \in \mathbb{N}$, $\exists j \leq N$ such that:

$$\overline{d}_1(\mu_n, \mu_j) \leq \varepsilon^2 \tag{4}$$

By *Proposition 3.2* in [**?**], we obtain that $\forall n \in \mathbb{N}, \exists K_\varepsilon$ compact such that $\mu_n(\mathbb{R} \setminus K_\varepsilon) < \varepsilon$. In other words, $(\mu_n)_{n\in\mathbb{N}}$ is yy. Hence, by monotonicity of the measure, we have that:

$$\mu_n(\mathbb{R}) - \mu_n(\mathbb{K}_\varepsilon) < \varepsilon \iff \mu_n(K_\varepsilon) > 1 - \varepsilon$$

This means in particular that $(\mu_j)_{j\leq N}$ is uniformly tight and so $\exists K$ compact such that $\mu_j(K) < 1-\varepsilon$ for all $j \leq N$, since a finite union of compact sets remains compact. Now, since K is compact, any of its open cover has a finite subcover. Let $\{x_1, ..., x_M\}$ be a finite subcover of an open cover of K and define:

$$U := \bigcup_{k=1}^{M} B(x_k, \varepsilon) \supseteq K$$

$$f : \mathbb{R} \to \mathbb{R}^+, f(x) \mapsto \max\left\{0, 1 - \frac{d(x,U)}{\varepsilon}\right\}$$

where $d(x, U) = \inf_x\{|x - u||u \in U\}$. Note that f is $\frac{1}{\varepsilon}$-Lipschitz since:

$$|f(x) - f(y)| = \left|1 - \frac{d(x,U)}{\varepsilon} - 1 - \frac{d(y,U)}{\varepsilon}\right| = \frac{1}{\varepsilon}|d(y,U) - d(x,U)| \leq \frac{1}{\varepsilon}|x - y|$$

Hence, $\forall n \in \mathbb{N}, j \leq N$, if we let π be any joint measure with marginals μ_j and μ_n over $\mathbb{R}^2$, we have:

$$\begin{aligned}
\int_{\mathbb{R}} f(x)d\mu_j(x) - \int_{\mathbb{R}} f(y)d\mu_n(y) &= \iint_{\mathbb{R}^2} (f(x) - f(y))d\pi(x,y) && \text{by definition of } \pi \\
&\leq \frac{1}{\varepsilon}\iint_{\mathbb{R}^2} |x - y|d\pi(x,y) && \text{since } f \text{ is } \frac{1}{\varepsilon}\text{-Lipschitz} \\
&\leq \frac{1}{\varepsilon} \cdot d_1(\mu_j, \mu_n) && \text{by definition of } d_1 \\
&\leq \frac{1}{\varepsilon} \cdot \overline{d}_1(\mu_j, \mu_n) && \text{by definition of } \overline{d}_1
\end{aligned} \tag{5}$$

Now, note that $f \geq 0$, and if $x \in U$, then $f(x) = 1$. Also, if we let $U^\varepsilon = \{x || d(x, U) < \varepsilon\}$, we have $x \in U^\varepsilon \implies f(x) \leq 1$, and $x \notin U^\varepsilon \implies f(x) = 0$. It follows that $\mathbb{1}_U \leq f \leq \mathbb{1}_{U^\varepsilon}$, and so:

$$\mu_j(U) \leq \int_{\mathbb{R}} f(x) d\mu_j(x) \text{ and } \int_{\mathbb{R}} f(x) d\mu_n(x) \leq \mu_n(U^\varepsilon) \tag{6}$$

Hence, we have:

$$\begin{aligned} \mu_n(U^\varepsilon) &\geq \mu_j(U) - \frac{1}{\varepsilon} \cdot \overline{d}_1(\mu_j, \mu_n) && \text{using (5), (6)} \\ &\geq 1 - \varepsilon - \frac{\varepsilon^2}{\varepsilon} && \text{using (4), (5)} \\ &= 1 - 2\varepsilon \end{aligned}$$

which means that $\mu_n(\mathbb{R} \setminus U^\varepsilon) \leq 2\varepsilon$. Since $U^\varepsilon \subset \bigcup_{k=1}^M B(x_k, 2\varepsilon)$ by construction, we have proven that $\forall \varepsilon > 0, \exists x_1, ..., x_M \in \mathbb{R}$ such that:

$$\mu_n \left(\mathbb{R} \setminus \bigcup_{k=1}^{M} B(x_k, 2\varepsilon) \right) \leq 2\varepsilon$$

By replacing ε by $2^{-m-1}\varepsilon$, with $m \in \mathbb{N}$, we get that $\forall n \in \mathbb{N}, \exists x_1^m, ..., x_{M_m}^m \in \mathbb{R}$ such that:

$$\mu_n \left(\mathbb{R} \setminus \bigcup_{k=1}^{M_m} B(x_k^m, 2^{-m}\varepsilon) \right) \leq 2^{-m}\varepsilon \tag{7}$$

Now define S to be the smallest set U:

$$S := \bigcap_{m=1}^{\infty} \bigcup_{k=1}^{M_m} B(x_k^m, 2^{-m}\varepsilon) \implies \mathbb{R} \setminus S = \bigcup_{m=1}^{\infty} \bigcap_{k=1}^{M_m} B(x_k^m, 2^{-m}\varepsilon)$$

We have that, $\forall n \in \mathbb{N}$:

$$\begin{aligned} \mu_n(X \setminus S) &\leq \sum_{m=1}^{\infty} \mu_n(\mathbb{R} \setminus \bigcup_{k=1}^{M_m} B(x_k^m, 2^{-m}\varepsilon)) && \text{by subadditivity of } \mu_n \\ &\leq \sum_{m=1}^{\infty} 2^{-m}\varepsilon && \text{using (7)} \\ &= \varepsilon \end{aligned}$$

Now, we let $r > 0$, and choose $m^\star$ such that $2^{-m}\varepsilon \leq r$. It follows that, by construction, S is covered by the $M_{m^\star}$ open balls $B(x_k^m, 2^{-m}\varepsilon)$, each with radius $2^{-m}\varepsilon < r$. Hence, S is totally bounded. Since, $\mathbb{R}$ is complete, we have that $\overline{S}$ is compact. So we constructed a compact set $\overline{S}$ such that $\forall n \in \mathbb{N}, \forall \varepsilon > 0, \mu_n(\mathbb{R} \setminus \overline{S}) \leq \varepsilon$. This shows that $(\mu_n)_{n \in \mathbb{N}}$ is uniformly tight.

Step 2: We prove that $(\mu_n)_n$ converges in $(\mathcal{Z}_p, \overline{d}_p)$ $p \geq 1$.

By a corollary of Prohorov's theorem in $\mathbb{R}$, we have that since $(\mu_n)_{n\in\mathbb{N}}$ is uniformly tight, there exists a subsequence $(\mu_{n'})_{n'\in\mathbb{N}}$ of $(\mu_n)_{n\in\mathbb{N}}$ such that $(\mu_{n'})_{n'\in\mathbb{N}}$ converges weakly to a probability measure μ in $\mathcal{Z}_p$. We want to show that $\overline{d}_p(\mu_n, \mu) \xrightarrow{n\to\infty} 0$ as this would imply that $(\mathcal{Z}_p, \overline{d}_p)$ is complete. Let $\pi_{n',m'}$ be a joint probability measure with marginals $\mu_{n'}$ and $\mu_{m'}$ over $\mathbb{R}^2$ such that:

$$\iint_{\mathbb{R}^2} |x-y|^p d\pi_{n',m'}(x,y) = \overline{d}_p(\mu_{n'}, \mu_{m'})^p$$

Given some $m' \in \mathbb{N}$, since the marginal $\mu_{n'}$ of $\pi_{n',m'}$) is uniformly tight, it follows that $(\pi_{n',m'})_{n'\in\mathbb{N}}$ also is. Hence, by the same corollary of Prohorov's theorem, there exists a subsequence $(\pi_{n'',m'})_{n''\in\mathbb{N}}$ of $(\pi_{n',m'})_{n'\in\mathbb{N}}$ that converges weakly to a joint probability measure $\pi_{m'}$. Then:

$$\begin{aligned}
\iint_{\mathbb{R}^2} |x-y|^p d\pi_{m'}(x,y) &\leq \liminf_{n''\to\infty} \iint_{\mathbb{R}^2} |x-y|^p d\pi_{n'',m'}(x,y) && \text{by Fatou's lemma} \\
&= \liminf_{n''\to\infty} \overline{d}_p(\mu_{n''}, \mu_{m'})^p && \text{by definition of } \overline{d}_p \qquad (8)
\end{aligned}$$

Now, note that since $\pi_{n'',m'}$ has marginals $\mu_{n''}$ and $\mu_{m'}$, in the limit as $n'' \to \infty$, we have that $\pi_{m'}$ has marginals μ and $\mu_{m'}$. Hence:

$$\overline{d}_p(\mu, \mu_{m'})^p \leq \iint_{\mathbb{R}^2} |x-y|^p d\pi_{m'}(x,y) \qquad (9)$$

Also, since the sequence $(\mu_{n'})_{n'\in\mathbb{N}}$ remains Cauchy (since it is a subsequence of a Cauchy sequence), $\forall \varepsilon > 0, \exists n'', m'$ large enough so that:

$$\overline{d}_p(\mu_{n''}, \mu_{m'}) \leq \varepsilon \qquad (10)$$

Finally, we have that:

$$\begin{aligned}
\overline{d}_p(\mu, \mu_{m'})^p &\leq \iint_{\mathbb{R}^2} |x-y|^p d\pi_{m'}(x,y) && \text{using (9)} \\
&\leq \liminf_{n''\to\infty} \overline{d}_p(\mu_{n''}, \mu_{m'})^p && \text{using (8)} \\
&\leq \varepsilon^p && \text{using (10) and taking } n'' \text{ large enough}
\end{aligned}$$

Which means that $\overline{d}_p(\mu, \mu_{m'}) \leq \varepsilon$ by taking n'' large enough, since $p \geq 1$. Hence, we can choose the difference between the p^{th} moments of μ and $\mu_{m'}$ to be arbitrarily small. It follows that $\mu \in \mathcal{Z}_p$, and $\overline{d}_p(\mu, \mu_{n'}) \xrightarrow{n'\to\infty} 0$, which in turns implies that $\overline{d}_p(\mu, \mu_n) \xrightarrow{n\to\infty} 0$ since $(\mu_n)_{n\in\mathbb{N}}$ is Cauchy in $\mathcal{Z}_p$. Hence, $(\mathcal{Z}_p, \overline{d}_p)$ is complete. $\square$