# Summary of A Distributional Perspective on Reinforcement Learning

Ashutosh Kumar Jha, ME14B148
Mechanical Engineering, IIT Madras

***Abstract*—In this technical report, we summarize the contents and ideas presented in the technical paper titled** *A Distributional Perspective on Reinforcement Learning*[1]**. We also present the possible shortcomings of the work.**

***Keywords*—*Deep Reinforcement Learning, Return Distribution***

## I. Summary

The work in this paper tries to learn the Value Distribution instead of learning just the Value Function of the underlying MDP in a reinforcement learning problem. The work is seminal in the sense that it is the first of its kind to use this Value Distribution in the policy evaluation and the control.

In classical RL, the Bellman's equation is given by:

$$Q(x,a) = \mathbb{E}R(x,a) + \gamma \mathbb{E}Q(X',A')$$

where $x, a, X', A'$ is the transition of the agent. The paper aims to learn the random return defined using $Z$. There exists an distributional equivalent of the Bellman's equation for the random return which is given by:

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

The variable $Z$ is then equivalently referred to as Value Distribution. In classical RL, we try to maximize the value function $Q(x,a)$ to choose the action. This gives rise to the Bellman Optimality Equation:

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

In this context, the Bellman Operator($\mathcal{T}^\pi$) and the Bellman Optimality Operator($\mathcal{T}$) are defined as follows:

$$\begin{aligned}\mathcal{T}^\pi Q(x,a) &:= \mathbb{E}R(x,a) + \gamma \mathbb{E}_{P,\pi} Q(x',a') \\ \mathcal{T} Q(x,a) &:= \mathbb{E}R(x,a) + \gamma \mathbb{E}_P \max_{a' \in \mathcal{A}} Q(x',a')\end{aligned}$$

There exists theoretical guarantees with respect to these operators where the existence of a fixed point and convergence of algorithms which make use of these operators in a recursive fashion can be shown by using the Banach Fixed Point Theorem.

The paper then draws parallels in the distributional case and defines similar operators for the distributional case. The paper first introduces the reader to the Wasserstein Metric between two cumulative distributions as follows:

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

where the infimum is taken over all pairs of random variables (U,V) with respective cumulative distributions F and G. The paper then defines a maximal form of the Wasserstein Metric for two value distributions:

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

This maximal metric $\bar{d_p}$ is used as the metric in the Banach Space for the distributional operators to prove the convergence of the recurrent equations for the distributional operators.
For defining the operators for the policy evaluation, we note that the quantity we are interested in is the value distribution $Z^\pi$ analogous to the value function $V^\pi$ associated with a *given* policy $\pi$. The reward function can be seen as a random vector $R \in \mathcal{Z}$ where $\mathcal{Z}$ is the space of all the value distributions. We now define the transition operator $P^\pi : \mathcal{Z} \to \mathcal{Z}$

$$\begin{aligned}P^\pi Z(x,a) &\overset{D}{=} Z(X',A') \\ X' &\sim P(.|x,a) \\ A' &\sim \pi(.|X')\end{aligned}$$

The distributional Bellman operator $\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ as

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

The paper then goes ahead to show that $\mathcal{T}^\pi$ is a $\gamma$-contraction in $\bar{d_p}$. Thus, by using the Banach Fixed Point Theorem, we have that the operator thus has an unique fixed point.
Note here that the paper refers to the KL divergence also as an alternative metric and confirms that for the KL divergence, the operator is no more a contraction.

In the control setting in classical RL, the focus is on learning the optimal value function by using the Bellman Optimality Operator. It is far easier because we know that all the optimal policies(call the set of the same as $\Pi$) have the same value function(and thus an optimal value function can be defined).
In the Distributional RL case, we first define what we even mean by an optimal value distribution. An optimal value distribution is the value distribution of an optimal policy. The set of optimal value distributions is defined as $\mathcal{Z}^*$. An optimal value distribution is stricter than an optimal value function in the sense that the expectation of the value distribution being the optimal value function is only a necessary but not sufficient condition for optimality of the value distribution.
A greedy policy for a given value distribution is defined as that which maximizes the expectation of the value distribution. Denote the set of all greedy policies for a value distribution Z by using $\mathcal{G}_Z$.

We can finally now define the Bellman Optimality operator in the Distributional RL case as an operator which implements the following:

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

The paper then goes ahead to define Non-stationary Optimal Value Distribution. A non-stationary optimal value distribution $Z$ is the value distribution corresponding to a sequence of optimal policies. More simply, it's like picking actions at every different step by using one of the different optimal policies.

The paper then goes ahead to show that the Optimality operator is not a contraction. In fact it goes to show that not all optimality operators have a fixed point. Furthermore, as a lemma, the paper shows that even the existence of a fixed point doesn't guarantee the convergence of the iterative process.

The paper(finally!) gives out its algorithm in the form of Approximate Distributional Learning. The paper approximate the value distribution by using a discrete distribution. The work parametrizes the discrete distribution by $N$ and $V_{\text{MIN}}$ and $V_{\text{MAX}}$. $N$ is the number of "atoms" in the distribution. The support of the distribution is kept between $V_{\text{MIN}}$ and $V_{\text{MAX}}$. The support of the distribution is the set of atoms given by:

$$z_i = V_{\text{MIN}} + i\Delta z \qquad i \in [0, N)$$
$$\Delta z = \left(\frac{V_{\text{MAX}} - V_{\text{MIN}}}{N-1}\right)$$

The atom probabilities are then given by a parametric model:

$$Z_\theta(x,a) = z_i$$
$$p_i(x,a) = \frac{e^{\theta_i(x,a)}}{\sum_j e^{\theta_j(x,a)}}$$

Another important part of the algorithm is the Projected Bellman Update. In the way that the discretization is set up, when the recursive updates are applied, there is no guarantee that the support will remain the same at different iterations. As a result, the sample update from the Bellman Update is projected onto the fixed support of the distribution. Mathematically, the $i^{th}$ component of the projected update is:

$$(\Phi\hat{\mathcal{T}} Z_\theta(x,a))_i = \sum_{j=0}^{N-1} \left[1 - \frac{|[\hat{\mathcal{T}} z_j]_{V_{\text{MIN}}}^{V_{\text{MAX}}}| - z_i}{\Delta z}\right]_0^1 p_j(x', \pi(x'))$$

where $[.]_a^b$ binds the argument in range $[a, b]$. We view the next-state distribution as parametrized by a fixed parameter $\widetilde{\theta}$. The sample loss is defined as:

$$\mathcal{L}_{x,a}(\theta) = D_{KL}(\Phi\hat{\mathcal{T}} Z_{\widetilde{\theta}}(x,a) || Z_\theta(x,a))$$

The authors claim that the Wasserstein metric and the Cramer Metric(which is referred to in another arXiv paper) both don't work as well as the KL Divergence.

*Experiments and Results:*

The paper reports results on the Arcade Learning Environment. The authors use the DQN architecture but output the atom probabilities $p_i(x,a)$ instead of the action-values. The network is trained to minimize the loss $\mathcal{L}_{x,a}(\theta)$. In spite of the ALE's deterministic nature, the distributions are not concentrated on one or two values.
The authors claim that intuitively their algorithm makes better decisions because the distributional updates separate the low-valued "losing" event from the high-valued "survival" event which gets averaged out into an expectation otherwise.
The authors try training the DQN by varying the number of atoms. The best performing agent was found to have 51 atoms(and is referred to as C51). The C51 agent outperforms a fully trained DQN agent on 45 out of 57 games using just 50 mil. frames as compared to 200 mil. frames.
The recent version of ALE allows introducing explicit stochasticity in the environment of ALE by making the environment reject the agent's action by probability of 0.25. C51 still outperforms DQN showing the benefits of C51 in a stochastic setting as well.
The results also compare the C51 agent against improved DQN agents like Double DQN, Duelling DQN, Prioritized Replay and Prioritized Duelling DQN. The C51 agent outperforms all the variants by a margin in both the mean and median scores across the 57 Atari games.

## II. Shortcomings

In this section we make a note of possible short comings of this paper:

- The Categorical DQN algorithm makes use of the KL divergence as a metric despite the absence of contraction property for KL Divergence as a metric. This is clearly counter intuitive to me.
- No theoretical guarantees are available in the control setting and clearly presence of such guarantees would be interesting to have.
- The upper and lower bounds for the support are set to be hyper-parameters. They are hard to be known in new tasks where no results are available.

## III. Conclusion

The paper introduces and substantiates the importance of learning the full value distribution instead of just the value function and develops an algorithm on top of their theory. The algorithm clearly shows improvements in terms of both sample efficiency and the scores beating the state-of-the-art on ALE. Clearly one of the best papers I've come across this year. The only problem I see in the paper is the use of KL divergence for the loss function which tosses all the theory developed out of the window.

## References

[1] A Distributional Perspective on Reinforcement Learning, Marc Bellemare, Will Dabney, Remi Munos. International Conference on Machine Learning (ICML), 2017.