CS7011 : Topics in Reinforcement Learning
Summary of A Distributional Perspective Reinforcement Learning : Marc G. Bellemare, Will Dabney and Remi Munos

Sahana Ramnath : EE15B109

Abstract

This paper (Bellemare et al., 2017) argues for the fundamental importance of using the *value distribution* : the distribution of the random return received by the agent, as opposed to using just the expected value of the random return(the *value*). Though a lot of work has been done previously regarding distributional reinforcement learning, this paper's perspective of modelling the return itself as a distribution is completely new. The paper gives theoritical results about the convergence of these distributions in the policy evaluation and control settings. It also gives a new algorithm(**C51**) which applies Bellman's equations to the learning of (approximate) value distributions. Experiments are mainly conducted on ALE games where state-of-the-art results and meaningful state-action value distributions are obtained/visualized.

1. Benefits of using Value Distributions

Modelling the entire distribution rather than using a sample or the expectation of it has many benefits :

- Preserves the **mutimodality** of the value distribution : Leads to more stable learning and more informed choices by the agent.
- Helps to model the **risk-averseness** of the agent : Since the distribution encodes the variance of the return, it has been extensively used to design risk-sensitive algorithms/policies.
- Makes it easier to model **non-stationary policies** : As seen later in this report, the learning algorithm used is not a contraction in any metric, making it a favourable choice for algorithms that model the effects of non-stationary policies.

2. Related Work

A lot of work has been done in distributional reinforcement learning so far, and they can be split into the following three broad areas :

1. To model parametric uncertainty : The distributional perspective is useful to model uncertainty. One previous work is 'Bayesian Q-Learning' (Dearden et al., 1998) which considers a Normal-Gamma prior over the distribution of the reward $R_{s,a}$ and then updates these priors based on the agent's experience.
2. To design risk-sensitive algorithms : Extensive work has been done to use the variance of the return to model risk-aware agents/policies. One notable work is 'Learning the Variance of the Reward to-go' (Tamar et al., 2016) which uses linear function approximation to learn the variance of the return.
3. General Theoritic Analysis : This paper is based on many previous theoritical results. 'MDPs with a new optimality criterion' (Jaquette, 1973) showed that a moment optimality criterion which imposes a total ordering on distributions is achievable. 'Discounted MDPs: Distribution functions and exponential utility' (Chung & Sobel, 1987) showed that the operator(introduced in this paper) is not a contraction in the metric 'total variation distance'.

3. Setting and Notation

The setting here is similar to the standard fashion in which an agent will interact with the environment. The MDP is represented as $< \mathcal{S}, \mathcal{A}, R, P, \gamma >$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, P is the transition kernel, γ is the discount factor and R is the random reward, now explicitly treated as a random variable(distribution).

The return Z^π is the sum of discounted rewards along the agent's trajectory in the environment. The state-action value function Q^π is the expected return along the trajectory, i.e. $\mathrm{E}[\mathrm{Z}^\pi]$.

The space the value distributions belong to is denoted by $\mathcal{Z}$.

4. The Distributional Bellman Operator

The Bellman Operator $\mathcal{T}^B$:

$$\mathcal{T}^B Q^\pi(s,a) = \mathrm{E}[R(s,a)] + \gamma \mathrm{E}_{P,\pi}[Q^\pi(s',a')]$$

The Distributional Bellman Operator $\mathcal{T}^\pi$:

$$\mathcal{T}^\pi Z^\pi(s,a) :\overset{D}{=} R(s,a) + \gamma P^\pi Z^\pi(s,a)$$
$$\mathcal{T}^\pi Z^\pi(s,a) :\overset{D}{=} R(s,a) + \gamma Z^\pi(s',a'),$$

where $P^\pi : \mathcal{Z} \rightarrow \mathcal{Z}$ is the transition operator.

$$P^{\pi}Z(s,a) \overset{D}{:=} Z(S',A')$$
$$S' \sim P(.|s,a) \text{ and } A' \sim \pi(.|S')$$

While the two operators seem similar in form, they are fundamentally different. The distribution $\mathcal{T}^{\pi}Z$ is defined by three sources of randomness :

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

5. Contraction of the Operator

This paper has proved that, for a fixed policy π in the environment, the Distributional Bellman Operator $\mathcal{T}^{\pi}$ is a contraction over the **maximal form** of the **Wasserstein Metric**.The choice of metric matters since the operator is not a contraction in other distribution metrics such as KL Divergence, Komogorov Distance or total variation distance.

5.1. The Wasserstein Metric

5.1.1. Definition

The Wasserstein or Mallows metric d_p is between cumulative distribution functions. For F, G, two distribution functions over reals :

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

where infimum is taken over all pairs of random variables U, V with c.d.fs as F and G.
It is infeasible to calculate the above expression as such. Instead, for $p < \infty$, the metric can be calculated as :

$$d_p(F,G) := (\int_0^1 |F^{-1}(u) - G^{-1}(u)|^p du)^{1/p}$$

For convenience and greater legibility, the Wasserstein Metric between two distributions will be represented for the rest of this report as $d_p(U,V)$ instead of $d_P(F,G)$(following the paper itself).

5.1.2. Some properties

The metric d_p has the following properties (a is a scalar and A is a random variable independent of U, V):

$$d_p(aU,aV) \leq |a|d_p(U,V)$$
$$d_p(A+U,A+V) \leq d_p(U,V)$$
$$d_p(AU,AV) \leq ||A||_p d_p(U,V)$$

Let $A_1, A_2, ...$ be a set of random variables describing a partition of the sample space Ω, i.e., $A_i(w) \in \{0,1\}$ and for any w there is exactly one A_i with $A_i(w) = 1$. Then :

$$d_p(U,V) \leq \sum_i d_p(A_iU, A_iV)$$

5.1.3. Maximal Form of the Wasserstein Metric

For two value distributions $Z_1, Z_2 \in \mathcal{Z}$, the maximal form of the Wasserstein Metric $\bar{d}_p$ is defined as :

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

$\bar{d}_p$ is proven to be a metric over distributions.
It is used to establish the convergence of the distributional Bellman Operator $\mathcal{T}^{\pi}$.

5.2. Convergence in Policy Evaluation

Consider the process $Z_{k+1} := \mathcal{T}^{\pi}Z_k$ starting with an arbitrary $Z_0 \in \mathcal{Z}$.

$\mathcal{T}^{\pi} : \mathcal{Z} \rightarrow \mathcal{Z}$ is a $\underline{\gamma\text{-contraction}}$ in $\bar{d}_p$ with the unique fixed point Z^{π}.
Proof : Consider $Z_1, Z_2 \in \mathcal{Z}$.

$$\begin{gathered} d_p(\mathcal{T}^{\pi}Z_1(s,a), \mathcal{T}^{\pi}Z_2(s,a)) \\ = d_p(R(s,a) + \gamma P^{\pi}Z_1(s,a), R(s,a) + \gamma P^{\pi}Z_2(s,a)) \\ \leq d_p(\gamma P^{\pi}Z_1(s,a), \gamma P^{\pi}Z_2(s,a)) \\ \leq \gamma d_p(P^{\pi}Z_1(s,a), P^{\pi}Z_2(s,a)) \\ \leq \gamma sup_{s',a'} d_p(Z_1(s',a'), Z_2(s',a')) \end{gathered}$$

Now,

$$\begin{gathered} \bar{d}_p(\mathcal{T}^{\pi}Z_1, \mathcal{T}^{\pi}Z_2) = sup_{s,a} d_p(Z_1(s,a), Z_2(s,a)) \\ \leq \gamma sup_{s',a'} d_p(Z_1(s',a'), Z_2(s',a') \\ = \gamma \bar{d}_p(Z_1, Z_2) \end{gathered}$$

Hence,
$$\boxed{\bar{d}_p(\mathcal{T}^{\pi}Z_1, \mathcal{T}^{\pi}Z_2) \leq \gamma \bar{d}_p(Z_1, Z_2)}$$

5.3. Convergence in Control

5.3.1. Optimal Value Distributions

Let Π^* be the set of optimal policies. An optimal value distribution is the value distribution of an optimal policy. The set of optimal value distributions is

$$Z^* := \{Z^{\pi^*} : \pi^* \in \Pi^*\}$$

All optimal policies attain the same expected value, Q^*. However, not all value distributions with expectation Q^* are optimal. They must match the full distribution of the return under an optimal policy.

5.3.2. Set of greedy policies

A greedy policy π for Z maximizes the expectation of Z. The set of greedy policies for Z is

$$G_Z := \{\pi : \sum_a \pi(a|s) \, \mathrm{E}[Z(s,a)] = max_{a' \in A} \, \mathrm{E}[Z(s,a')]\}$$

5.3.3. Distributional Bellman Optimality Operator

Now, the Bellman Optimality Operator $\mathcal{T}^{B*}$ is

$$\mathcal{T}^{B*}Q(s,a) := \mathrm{E}[R(s,a)] + \gamma\,\mathrm{E}_P[max_{a' \in A} Q^*(s',a')],$$

where the maximization at s' corresponds to some greedy policy.
Similar to the above, any operator which implements a greedy selection rule is a Distributional Bellman Optimality Operator $\mathcal{T}$.

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

This paper proves that,

- The distributional Bellman optimality operator converges in a weak sense, to the set of optimal value distributions.
- This operator however, is not a contraction in any metric for distributions.

5.3.4. Convergence of the expected value

Let $Z_1, Z_2 \in Z$.
Then,$|| \mathrm{E}[\mathcal{T}Z_1] - \mathrm{E}[\mathcal{T}Z_2]||_\infty \leq \gamma || \mathrm{E}[Z_1] - \mathrm{E}[Z_2]||_\infty$, and in particular, $\mathrm{E}[Z_k] \to Q^*$ exponentially quickly.

5.3.5. Nonstationary Optimal Value Distributions

A nonstationary optimal value distribution Z^{**} is the value distribution corresponding to a sequence of optimal policies.
The set of nonstationary optimal value distributions is Z^{**}.
If the state space is finite, then Z_k converges to Z^{**} uniformly.
Also, if there is a total ordering $\prec$ on Π^* such that for any Z^*,

$$\mathcal{T}Z^* = \mathcal{T}^{\pi}Z^* \text{ with } \pi \in G_{Z^*}, \pi \prec \pi', \forall \pi' \in G_{Z^*}/\{\pi\}\ ,$$

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

5.3.6. Results about $\mathcal{T}$

While the mean of Z_k converges exponentially quickly to Q^*, its distribution need not be well behaved.

- The operator $\mathcal{T}$ is not a contraction.
- Not all optimality operators have a fixed point.
- The presence of a fixed point is insufficient to guarantee convergence.

6. Approximate Distributional Learning

6.1. Parametric Distribution

In the experiments, the value distribution is modelled using a discrete distribution parametrized by $N \in \mathbb{N}$ and $V_{MIN}, V_{MAX} \in \mathbb{R}$ and whose support is the set of atoms

$$\{z_i = V_{MIN} + i\Delta z : 0 \leq i < N\}, \Delta z := \frac{V_{MAX} - V_{MIN}}{N-1}.$$

The atom probabilities are given by a parametric model $\theta : SXA \to \mathbb{R}^N$

$$Z_\theta(s,a) = z_i \text{ with probability } p_i(s,a) := \frac{e^{\theta_i(s,a)}}{\sum_j e^{\theta_j(s,a)}}$$

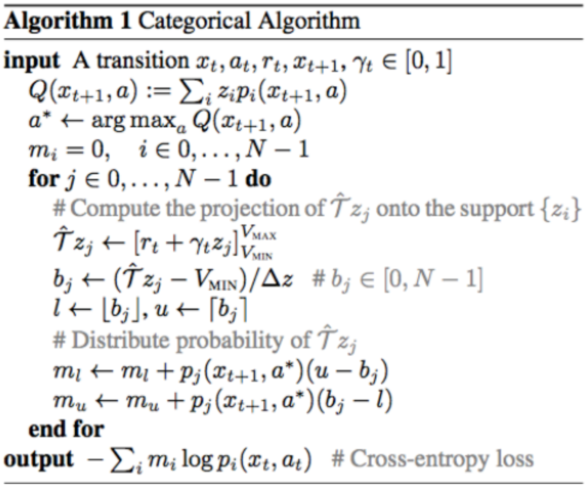
A discrete(rather than continuous) distribution is used, for easy computation and implementation purposes.

6.2. Projected Bellman Updates

Using discrete distributions poses one major problem : The Bellman update $\mathcal{T}Z_\theta$ and the original parametrization Z_θ almost always have disjoint supports.
This discrepancy in supports can be handled by the Wasserstein Metric(which can be viewed as a loss between $\mathcal{T}Z_\theta$ and Z_θ). However, the experiments are restricted to learning from sample transitions, which cannot be handled by the Wasserstein Metric.
Hence, the sample Bellman update $\hat{\mathcal{T}}Z_\theta$ is projected onto the support of Z_θ, effectively reducing the Bellman update to multiclass classification.

7. The Categorical C51 Algorithm

This is the main algorithm proposed by the paper to perform **iterative approximation** of the value distribution Z using the distributional Bellman equation.



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 $\{z_i\}$
    $\hat{\mathcal{T}}z_j \leftarrow [r_t + \gamma_t z_j]_{V_{\mathrm{MIN}}}^{V_{\mathrm{MAX}}}$
    $b_j \leftarrow (\hat{\mathcal{T}}z_j - V_{\mathrm{MIN}})/\Delta z$  # $b_j \in [0, N-1]$
    $l \leftarrow \lfloor b_j \rfloor, u \leftarrow \lceil b_j \rceil$
    # 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
```

Note : The number **51** represents the number of discrete values used to parametrize the value distribution (found using trial and error).

Basically, at each time step, a transition is sampled from the environment is used to compute the target distribution $R + \gamma Z'$. This is used to update the current distribution Z by minimizing the **cross-entropy** loss between the two.

8. Experiments Conducted

The C51 algorithm was applied to ALE games. While ALE is deterministic, stochasticity can occur due to state aliasing, learning from a non stationary policy and approximation errors.

A DQN was used to compute the atom probabilities $p_i(s, a)$ with $V_{MAX} = -V_{MIN} = 10$. The resulting architecture was named as the **'Categorical DQN'**. The squared error loss of the DQN is replaced by the cross-entropy loss. An ϵ-greedy policy was used to enable exploration.

8.1. Varying number of atoms

It was observed that more atoms always led to better performance and too few leads to poor behaviour. Exceedingly good results were obtained for the 51 atom version.

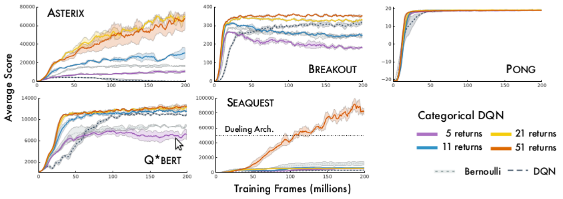
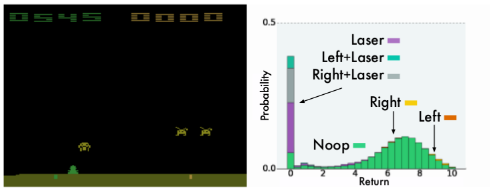


Figure 3. Categorical DQN: Varying number of atoms in the discrete distribution. Scores are moving averages over 5 million frames.

8.2. Meaningful Value Distributions

Below is the learned value distribution for the game 'Space Invaders.



Different actions are shaded different colours. In this episode, three actions (LASER, RIGHT/LEFT+LASER) lead to the agent releasing its laser too early and eventually losing the game. Their corresponding distributions depict this: they assign a significant probability to 0 (the terminal value).The safe actions have similar distributions (LEFT, which tracks the invaders' movement, seems to be slightly favoured).

The distribution keeps the low-value 'losing' event separated from the high-value 'survival' event, rather than averaging them into one unrealizable expectation.

Analysis of the value distributions of more games such as Seaquest, Pong and Q*Bert are given in the paper.

9. Intuition to learn Value Distributions

In this algorithm, the policy being used is still trying to maximize expected reward. But the difference here is that distributions are learnt to model uncertainty and approximation.

- **Reduced chattering :** By averaging the different distributions similar to conservative policy iteration, the oscillations/instability of convergence of the Bellman optimality operator is prevented.
- **State Aliasing :** By explicitly modelling the resulting distribution, a more stable learning target is provided.
- **A richer set of predictions :** Rich set of auxillary predictions. The accuracy of these auxillary predictions is closely tied to the agent's performance, so this leads to better learning.
- **Framework for inductive bias :** Treats all returns greater than V_{MAX} as equivalent. However, similar value clipping in DQNs significantly degrades the performance.
- **Well-behaved optimization :** KL Divergence is good to minimize for the discrete case, however it is not good for continuous distributions since it is insenstive to the value of its outcomes. Closer minimization of the Wasserstein metric would yield even better results.

References

Bellemare, Marc G, Dabney, Will, and Munos, Rémi. A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887*, 2017.

Chung, Kun-Jen and Sobel, Matthew J. Discounted mdp's: Distribution functions and exponential utility maximization. *SIAM journal on control and optimization*, 25(1):49–62, 1987.

Dearden, Richard, Friedman, Nir, and Russell, Stuart. Bayesian q-learning. In *AAAI/IAAI*, pp. 761–768, 1998.

Jaquette, Stratton C. Markov decision processes with a new optimality criterion: Discrete time. *The Annals of Statistics*, pp. 496–505, 1973.

Tamar, Aviv, Di Castro, Dotan, and Mannor, Shie. Learning the variance of the reward-to-go. *The Journal of Machine Learning Research*, 17(1):361–396, 2016.