A Distributional Perspective on RL

Marc G. Bellemare, Will Dabney, Remi Munos

***Abstract*—The paper portrays modelling *value distribution* instead of expectation of return or value as is done commonly. This can be used to incorporate *risk awareness* to the RL agent's behaviour. The authors make use of Bellman's equation inorder to learn the value distributions and provide theoretical results in policy evaluation as well as control, stating significant distributional instability in the latter.**

I. Main Concepts

Bellman Equation : Q^* is the optimal value function, corresponding to the set of optimal policies Π^* ; $Q^*(x,a) = \mathbb{E}\ R(x,a) + \gamma \mathbb{E}_P \max_{a' \in \mathcal{A}} Q^*(x',a')$. *Bellman operator* $\mathcal{T}^\pi$ and *Optimality operator* $\mathcal{T}$ are given by :
$\mathcal{T}^\pi Q(x,a) := \mathbb{E}R(x,a) + \gamma \underset{P,\pi}{\mathbb{E}}\, Q(x',a')$
$\mathcal{T} Q(x,a) := \mathbb{E}R(x,a) + \gamma \underset{P}{\mathbb{E}} \max_{a' \in \mathcal{A}} Q(x',a')$
Distributional Bellman Equation : Z is the random return with expectation Q. $Z(x,a) \overset{D}{=} R(x,a) + \gamma Z(X',A')$; reward R, the next state-action (X',A'), and its random return $Z(X',A')$
Wasserstein Metric : Wasserstein metric d_p between cumulative distribution functions F, G over the reals, is defined as $d_p(F,G) := \inf_{U,V} ||U - V||_p,$; $||U||_p := [\mathbb{E}[||U(\omega)||_p^p]]^{1/p}$
Infimum is taken over all pairs of random variables (U,V) with respective cumulative distributions F and $G : d_p(F,G) = ||F^{-1}(\mathcal{U}) - G^{-1}(\mathcal{U})||_p$; $F^1(q) := inf\{y : F_U(y) \geq q\}$
The *maximal form of Wasserstein metric* : $\bar{d}_p(Z_1,Z_2) := \sup_{x,a} d_p(Z_1(x,a), Z_2(x,a))$ for two value distributions $Z_1, Z_2 \in \mathcal{Z}$; $\mathcal{Z}$ denotes the space of value distributions with bounded moments and $\bar{d}_p$ is a metric over value distributions.
Contraction of the Policy Evaluation Bellman operator : Bellman operator $\mathcal{T}^\pi$ is a contraction over the value distribution (for a fixed policy) in a maximal form of the Wasserstein metric. The *distributiona Bellman operator* $\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ is defined as : $\mathcal{T}^\pi Z(x,a) :\overset{D}{=} R(x,a) + \gamma P^\pi Z(X',A')$
$P^\pi Z(x,a) :\overset{D}{=} Z(X',A')$; $X' \sim P(\cdot|x,a),\ A' \sim \pi(\cdot|X')$
The 3 sources of randomness associated with the distribution $\mathcal{T}^\pi Z$ are :

- randomness in the reward $R \in \mathcal{Z}$
- randomness in the transition $P^\pi : \mathcal{Z} \to \mathcal{Z}$
- next-state value distribution $Z(X',A')$

$\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ is a γ-contraction in $\bar{d}_p$.
$Z_{k+1} := \mathcal{T}^\pi Z_k$, starting with some $Z_0 \in \mathcal{Z}$, $\{Z_k\}$ converges to Z^π($\mathcal{T}^\pi$ has a unique fixed point) in $\bar{d}_p$ for $1 \leq p \leq \infty$ (assuming all moments are bounded) (all moments also converge exponentially quickly)
Note : $\mathcal{T}^\pi$ is not a contraction in total variation distance, Kullback-Leibler divergence, or Kolmogorov distance.
Note : $\mathcal{T}^\pi$ is a contraction in variance ($p = 2$), but it is not a contraction in the p^{th} centered moment, $p > 2$ (but the centered moments of the iterates $\{Z_k\}$ will still converge exponentially quickly to those of Z^π)
Instability in the control setting : The Bellman optimality operator $\mathcal{T}$ is not a contraction in any metric over the value function distribution; although it is a contraction in expected value. $\mathcal{T} Q(x,a) := \mathbb{E}R(x,a) + \gamma \underset{P}{\mathbb{E}} \max_{a' \in \mathcal{A}} Q(x',a')$
Multiple optimal policies exist.
The distributional analogue of the Bellman optimality operator converges, in a weak sense, to the set of optimal value distributions(value distribution of an optimal policy)
$Z^* := Z^{\pi^*} : \pi^* \in \Pi^*$; Z^{π^*}-set of optimal value distributions and Π^*- set of optimal policies
Note : All value distributions with expectation Q^* are NOT optimal; full distribution of return under some optimal policy should be the same.
For $Z_{k+1} := \mathcal{T} Z_k; Z_0 \in \mathcal{Z}$, mean of $\{Z_k\}$ converges exponentially quickly to Q^* ($\mathbb{E}Z_k \to Q^*$), but its distribution need not.

- Optimality operator $\mathcal{T}$ is not a contraction
- Not all optimality operators have fixed point $Z^* = \mathcal{T} Z^*$
- $\mathcal{T}$ having a fixed point $Z^* = \mathcal{T} Z^*$ is insufficient to guarantee the convergence of $\{Z_k\}$ to Z^*

Approximate Distributional Learning : The value distribution is modelled using a discrete distribution (computationally efficient): $Z_\theta(x,a) = z_i$ with atomic probability (parametric model θ) $p_i(x,a) := \frac{e^{\theta_i(x,a)}}{\sum_j e^{\theta_j(x,a)}}$; where $z_i = V_{\text{MIN}} + i\Delta z$ for $0 \leq i < N$ and $\Delta z := \frac{V_{\text{MAX}} - V_{\text{MIN}}}{N-1}$
Since it's not really possible to learn from sample trajectories under Wasserstein loss (minimize Wasserstein metric between $\mathcal{T} Z_\theta$ and Z_θ), the sample Bellman update $\hat{\mathcal{T}} Z_\theta$ is projected onto the support of Z_θ, effectively reducing the Bellman update to a multiclass classification problem. The sample loss $L_{x,a}(\theta)$ is the cross-entropy term of the KL divergence between projected update $\Phi \hat{\mathcal{T}} Z_\theta(x,a)$ and $Z_\theta(x,a)$. This is referred to as the *categorical algorithm* and when $N = 2$, it is called *Bernoulli algorithm*
Better approximating the full distribution : The distributional Bellman operator preserves multimodality in value distributions, lead to more stable learning and this approximation mitigates the effects of learning from a non-stationary policy. Other benefits include reduced chattering, state aliasing, richer set of predictions and well-behaved optimization.

II. Experiments

They used a *categorical DQN* architecture (output the atom probabilities $p_i(x,a)$ instead of action-values, and chose $V_{\text{MAX}} = -V_{\text{MIN}} = 10$). The 51-atom version is called C51.

References

[1] Marc G Bellemare, Will Dabney, and Rémi Munos. "A distributional perspective on reinforcement learning". In: *arXiv preprint arXiv:1707.06887* (2017).