A Note on An Analysis of Categorical Distributional Reinforcement Learning

Sophia Lien
Department of Computer Science
National Chiao Tung University
sophia.cs07g@nctu.edu.tw

1 Introduction

The Paper I chose is "An Analysis of Categorical Distributional Reinforcement Learning". The main research challenges tackled by this paper is even though the distributional approaches to value-based reinforcement learning model the entire distribution of returns have recently been shown to yield state-of-the-art empirical performance, the theoretical properties of this algorithms are not well understood. Take the C51 paper for example, they introduced a contractive distributional Bellman operator in real distribution yet implementing with a projected operator. In this paper they use Cramer distance between probability distributions as the center of the framework to give a proof of convergence for sample-based categorical distributional reinforcement learning algorithms. The main contribution of the paper is to provide a theoretical framework for the analysis of distributional algorithms demonstrated their convergence properties. From my personal perspective, it is important to know that even though in C51 they use KL-divergence the convergence is only guaranteed under the soft update version rather than KL loss.

2 Problem Formulation

Consider a Markov decision process $(\mathcal{X}, \mathcal{A}, p, \gamma, \mathcal{R})$ with a finite state space $\mathcal{X}$, a finite action space $\mathcal{A}$, a transition kernel $p : \mathcal{X} \times \mathcal{A} \to \mathcal{P}(\mathcal{X})$ which defines a distribution over next state given a current state-action pair, a discounted rate $\gamma \in [0, 1)$, and a reward distribution$\mathcal{R}(x, a) \in \mathcal{P}(\mathcal{R})$ $for\ each$ $(x, a) \in \mathcal{X} \times \mathcal{A}$. Hence if an agent is at state $X_t \in \mathcal{X}$ at time t $\in \mathbb{N}_0$, and an action$A_t \in \mathcal{A}$ is taken, the agent transitions to a state $X_{t+1} \sim p(\cdot|X_t, A_t)$ and receives a reward $R_t \sim R(X_t, A_t)$.

Consider a given stationary Markov policy $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ which defines a probability distribution over the action space given a current state. The return of a policy π is defined as the random variable given by the sum of rewards:

$$\sum_{t=0}^{\infty} \gamma^t R_t | X_0 = x, A_0 = a \tag{1}$$

The major two tasks in RL are : (1) evaluation task: evaluation of π is to compute the expected return $E_\pi[\sum_{t=0}^{\infty} \gamma^t R_t | X_0 = x, A_0 = a]$, where E_π indicates that at each time step $t \in \mathbb{N}_0$, the agent's action A_t is sampled from $\pi(\cdot|X_t)$. (2) control task: this task is to find a policy $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ that maximizes the expected reward.

Preprint. Under review.

For each policy π, we consider the collection of distributions

$$\eta_\pi = \left\{ \eta_\pi^{(x,a)} | (x,a) \in \mathcal{X} \times \mathcal{A} \right\} \tag{2}$$

where $\eta_\pi^{(x,a)}$ is the distribution of the return of policy π at initial state-action $(x,a) \in \mathcal{X} \times \mathcal{A}$

$$\eta_\pi^{(x,a)} = \sum_{t=0}^{\infty} \gamma^t R_t | X_0 = x, A_0 = a, X_{t+1} \sim p(\cdot|X_t, A_t), A_t \sim \pi(\cdot|X_t) \tag{3}$$

This means the return distribution functionη_π maps each state-action pair (x,a) to a distribution $\eta_\pi^{(x,a)}$, and in Bellemare et al.[2017a], they showed this function η_π satisfies a distributional variant of the Bellman equation. In order to express the distributional Bellman equation, we need the notion of pushforward measures:

Definition 1. Given a probability distribution $\nu \in \mathcal{P}(R)$ and a measurable function $f : \mathbb{R} \to \mathbb{R}$, we let $f_\# \nu$ denote the pushforward measure, where $f_\# \nu \in \mathcal{P}(R)$ is defined by $f_\# \nu(A) = \nu(f^{-1}(A))$, for all Borel sets $A \subseteq \mathbb{R}$. In particular,given $r, \gamma \in \mathbb{R}$, we let$(f_{r,\gamma})_\# \nu$ be the pushforward measure, where $(f_{r,\gamma})_\# \nu(x) \triangleq r + \gamma x$.

We consider the distributional Bellman operator given by $T^\pi : \mathcal{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}} \to \mathcal{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}}$

$$\begin{aligned} \eta_\pi^{(x,a)} &= (T^\pi \eta_\pi)^{(x,a)} \\ &= \int_{\mathbb{R}} \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} (f_{r,\gamma})_\# \eta_\pi^{(x',a')} \pi(a'|x') p(dr, x'|x, a), \end{aligned} \tag{4}$$

for all $\eta_\pi \in \mathcal{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}}$ and the operator T^π maps functions into functions.

For the control version of the distributional Bellman operator:

$$\begin{aligned} (Teta)^{(x,a)} = \int_{\mathbb{R}} \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} (f_{r,\gamma})_\# \eta^{(x', a^*(x'))} \pi(a'|x') p(dr, x'|x, a), \\ where\ a^*(x') \in \underset{a' \in A}{arg\ max} \mathbb{E}_{R \sim \eta(x',a')}[R] \end{aligned} \tag{5}$$

To measure the distance between distributions, we need more notion:

Definition 2. The p-Wasserstein distance d_p, for $p \leq 1$is defined on $P_p(\mathbb{R}$, the set of probability distributions with finite p^{th}moments:

$$d_p(\nu_1, \nu_2) = \left(\underset{\lambda \in \Lambda(\nu_1 \nu_2)}{inf} \int_{\mathbb{R}^2} |x - y|^p \lambda(dx, dy) \right)^{1/p} \tag{6}$$

for all $\nu_1, \nu_2 \in \mathcal{P}_p(\mathbb{R})$, where$\Lambda(\nu_1, \nu_2)$is the set of probability distributions on $\mathbb{R}^2$ with marginals $\nu_1$and ν_2.

Then for the distance between two collection of distributions, we can use the supremum-p-Wasserstein metrics$\bar{d}_p$:

$$\bar{d}_p(\eta, \mu) = \underset{(x,a) \in \mathcal{X} \times \mathcal{A}}{sup} d_p(\eta^{(x,a)}, \mu^{(x,a)}), \tag{7}$$

for all $\eta, \mu \in \mathcal{P}_p(\mathbb{R})^{\mathcal{X} \times \mathcal{A}}$.

3 Theoretical Analysis

3.1 Categorical policy evaluation and categorical Q-learning

The first contribution of this paper is to make the parametrizations, approximations and assumptions in the categorical distributional reinforcement learning(CDRL) clear.

3.1.1 Distributional Approximations in CDRL

In Bellemare et al. [2017a], they prove that the distributional Bellman operator T^π is a $\gamma -$ *contraction* in $\bar{d}_p$, for all $p \leq 1$. Hence we have, for any initial set of distributions $\eta \in P(\mathbb{R})^{(\mathcal{X}\times\mathcal{A})}$:

$$(T^\pi)^m \eta \to \eta_\pi \; in \; \bar{d}_p, \; as \; m \to \infty \tag{8}$$

Under distributional RL, we initiate an estimation of the return distributions and update the estimates iteratively. However, it is impossible to represent the full space of probability distributions with a finite collection of parameters. In CDRL, they select some fixed set of supports, and use this supports to represent the categorical distributions:

$$\mathcal{P} = \left\{ \sum_{i=1}^{K} p_i \delta_{z_i} | p_1, ..., p_K \leq 0, \sum_{k=1}^{K} p_k = 1 \right\}, \tag{9}$$

where supports $z_1, ..., z_K$ is a set of equally-space supports.

3.1.2 Stochastic Approximations in CDRL

In CDRL, they follow the popular approximation way by using the transition $(x_t, a_t, r_t, x_{t+1}, a^*)$ to update the distributional Bellman operator $^\pi$. This defines a stochastic Bellman operator:

$$(\hat{T}^\pi \eta_t)^{(x_t,a_t)} = (f_{r_t,\gamma})_{\#_t}^{(x_{t+1},a^*)}, \tag{10}$$

where the randomness in$\hat{T}^\pi$ comes from the randomly sampled transition $(x_t, a_t, r_t, x_{t+1}, a^*)$. However $(\hat{T}^\pi \eta_t)^{(x_t,a_t)}$ typically do not lie in the parametric family $\mathcal{P}$ as shown in (9) after the transformation by the affine map $f_{r,\gamma}$. Hence we need a projection operator $\Pi : P(\mathbb{R}) \to \mathcal{P}$ to map the distribution back into the parametric family. In CDRL, they use a mixture of Diracs as their heuristic projection operator Π_c. After getting a stochastic approximation$\hat{\eta}_t^{(x_t,a_t)} = \left(\Pi_c \hat{T}^\pi \eta_t\right)^{(x_t,a_t)}$, they use one step of gradient descent on the Kullback-Leibler dibergence of the prediction $\eta_t^{(x_t,a_t)}$from the target $\hat{\eta}_t^{(x_t,a_t)}$:

$$KL(\hat{\eta}_t^{(x_t,a_t)} || \eta_t^{(x_t,a_t)}), \tag{11}$$

3.2 Convergence analysis

After understanding the approximations, parametrisations, and heuristics in CDRL, we are going to discuss the theoretical guarantees. In Bellemare et al. [2017a], they established the distributional Bellman operator to an initial return distribution function guarantees convergence to the true set of return distributions in the supremum-Wasserstein metric. However in 3.1.1 we know that actually we need a projection Π_c which could break the contractivity under Wasserstein distances.

Lemma 1. The operator $\Pi_c T^\pi$ is in general not a contraction in $\bar{d}_p$, for $p \leq 1$.

Intuitively, the reason why this could break the contractivity is the cost of using the parametrisation $\mathcal{P}$ rather than the fully non-parametric probability distributions. Fortunately, under Cramer distance [Szekely, 2002], the combined operator $\Pi_c T^\pi$ is a contraction.

Definition 3. The Cramer distance l_2 between two distributions $\nu_1, \nu_2 \in \mathcal{P}(\mathbb{R})$, with cumulative distribution function F_{ν_1}, F_{ν_2} respectively:

$$l_2(\nu_1, \nu_2) = \left(\int_{\mathbb{R}} (F_{\nu_1}(x) - F_{\nu_2}(x))^2 dx \right)^{1/2} \tag{12}$$

And the supremum-Cramer metric $\bar{l}_2$ between two distribution functions $\eta, \mu \in \mathcal{P}(R)^{(\mathcal{X}\times\mathcal{A})}$:

$$\bar{l}_2(\eta, \mu) = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} l_2(\eta^{(x,a)}, \mu^{(x,a)}). \tag{13}$$

The Cramer distance provide a useful geometric structure on the space of probability measures. This provides a new interpretation of the heuristic projectionΠ_c connected with the Cramer distance.

Proposition 1. The Cramer metric l_2 endows a particular subset of $\mathcal{P}(\mathbb{R})$with a notion of orthogonal projection, and the orthogonal projection onto the subset $\mathcal{P}$ is exactly the heuristic projection Π_c. Consequently, Π_c is a non-expansion with repect to l_2 which means it satisfy the Pythagoran theorem.

This helps us to establish the proof of contractibity of the operator $\Pi_c T^\pi$:

Proposition 2. The operator $\Pi_c T^\pi$ is a $\sqrt{\gamma}$-contraction in $\bar{l}_2$. Further, there is a unique distribution function $\eta_c \in \mathcal{P}^{\mathcal{X}\times\mathcal{A}}$ such that given any initial distribution function η_0, we have:

$$(\pi_c T^\pi)^m \eta_0 \to \eta_c \; in \; \bar{l_2}, \; as \; m \to \infty \tag{14}$$

In order to provide the convergence guarantees for sample-based distributional RL, we update the return distribution softly by learning rate $(\alpha_t(x,a)|(x,a) \in \mathcal{X}\times\mathcal{A}, t \leq 0)$, rather than use KL loss:

$$\eta_{t+1}^{(x,a)} \leftarrow (1-\alpha_t(x,a))\eta_t^{(x,a)} + \alpha_t(x,a)\hat{\eta}_t^{(x,a)} \tag{15}$$

3.2.1 Convergence of categorical policy evaluation

Under the mixture update rule shown as (15), categorical policy evaluation is guaranteed to converge to the fixed point of the projected Bellman operator $\Pi_c T^\pi$:

Theorem 1 In the context of policy evaluation for some policy π, suppose that:
(1) the stepsize $(\alpha_t(x,a)|(x,a) \in \mathcal{X}\times\mathcal{A}, t \leq 0$ satisfy the Robbins-Monro conditions:

$$\sum_{t=0}^{\infty} \alpha_t(x,a) = \infty \tag{16}$$

$$\sum_{t=0}^{\infty} \alpha_t(x,a)^2 < \infty \tag{17}$$

(2) we have initial estimates $\eta_0^{(x,a)}$ of the distribution of returns for each state-action pair $(x,a) \in \mathcal{X}\times\mathcal{A}$, each with support contained in $[z_1, z_K]$.

Then for the updates of mixture update rule shown in (15), we have almost sure convergence in $\bar{l}_2$.

3.2.2 Convergence of categorical Q-learning

Theorem 2. Suppose the assumption in Theorem 1(1) holds shown as (16)(17), and all unprojected target distributions are supported within $[z_1, z_K]$ almost surely. Assume further that there is a unique optimal policy π^* for the MDP. Then, for the mixture updates in the control case, we have almost sure convergence of $(\eta_t^{(x,a)})_{(x,a)\in\mathcal{X}\times\mathcal{A}}$ in $\bar{l}_2$ to some limit η_c^*, and furthermore the greedy policy with respect to η_c^* is the optimal policy π^*

4 Conclusion

In the paper An analysis of Categorical Distributional Reinforcement Learning, they provide a framework for distributional RL and convergence analysis. The main problem in C51 is they use KL loss for the update. However the convergence only in the true distribution in supremum-Wasserstein distance. In that algorithm, because the support assumption, we need consider the projected operator in order to fit the support assumption. In that case, the algorithm could break the convergence. In this paper, they prove that if we use a soft update for the algorithm can converge in both categorical policy evaluation and categorical Q-learning.

The issue of how function approximation in distributional RL remains an important question,for example, whether the algorithm convergence results hold for the KL update remains open. And also any other empirical parametrisation assumption is an interesting future direction.