IFT 6085 - Lecture 16
Distributional RL

This version of the notes has not yet been thoroughly checked. Please report any bugs to the scribes or instructor.

Scribes **Instructor:** Ioannis Mitliagkas
Winter 2019: [Gabriel Laberge & Leonard Sauvé & Jeremie Zumer & Chin-Wei Huang]

1 Review of Previous Lecture

In the previous lecture we discussed reinforcement learning (RL) algorithms in the finite horizon case to solve a Markov Decision Process (MDP).
An MDP over a finite horizon is defined as the tuple $(\mathcal{S}, \mathcal{A}, r, P)$ where:

- $\mathcal{S}$ is the state space;
- $\mathcal{A}$ is the action space;
- r is the reward function;
- P is the state transition function.

An RL algorithm optimizes a policy π in order to maximize the expected sum of future rewards (the "return").
Learning in RL is defined by the state value function, or the state-action value function, which respectively map states or state-action tuples to putative values, i.e. they model the expected return at some state, possibly depending on the action to take. Training is then performed by iterating two main steps:

1. Compute the value function under the current policy;
2. Update the policy based on the new value function.

Where the details of the iteration depends on the specific algorithm and the policy can be implicit (such as in the case of value-iteration-based algorithms).

Definition 1 (finite-horizon State Value Function)**.** *We define the* ***state value function*** $V_T^\pi(s)$ *of a policy* π *as the expected sum of rewards over t steps:*

$$V_T^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{T} r_t \Big| s\right]$$

Similarly, we define the state-value function in infinite-horizon by taking 1 and by multiplying the summation terms with a factor γ^t. Informally, this parameter will set how the value function V_T^π takes into account future steps for computing the reward.

Definition 2 (infinite-horizon State Value Function)**.** *We define the infinite-horizon* ***state value function*** $V_T^\pi(s)$ *of a policy* π *as the expected sum of rewards over T steps, and parameter* $\gamma < 1$ *as:*

$$V_T^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t \Big| s\right]$$

Since the sate value function is also dependent on the sequence of actions a taken at each time step, we can develop definition 1 further with a state-action value function $Q(s, a)$ also tied to the policy π and the time-steps T.

Definition 3 (State Action Value Function)**.** *We define the* ***state-action value function*** $Q_T^\pi(s, a)$ *of a policy as the expected sum of rewards over T steps:*

$$Q_T^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{T} r_t \Big| s, a\right]$$

Note that the state-action value function $V_T^\pi(s)$ is simply the state action value $Q_T^\pi(s, a)$ function where actions a are marginalized over.
Similarly to def 2, we get the following form for an infinite-horizon setup:

Definition 4 (infinite-horizon State Action Value Function)**.** *We define the infinite-horizon* ***state-action value function*** $Q_T^\pi(s, a)$ *of a policy as the expected sum of rewards over T steps, and parameter* $\gamma < 1$ *as:*

$$Q_T^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t \Big| s, a\right]$$

Consequently, optimization of the expected reward can be seen as solving the bellman's equation for an utility function V, that is described below. We have that for an infinite-horizon decision problem, the value function $V(s_0)$ reaches equilibrium when the Bellman's equation is satisfied.

Definition 5 (Bellman's equation)**.**

$$V(s_0) = \max_{a_{t=0}^\infty} \sum_{t=0}^{\infty} \gamma^t F(s_t, a_t)$$

Where $F(a_t, s_t)$ *defines the expected reward of a state-action pair, submitted to constraints* $s_{t+1} \sim \pi(s_t, a_t)$ *and discount factor* $0 < \gamma < 1$.

Note that V is a function of the optimal sequence of actions $a_{t=0}^\infty$ **as well as** the optimal/maximal value of the reward allowed by the environment. We will see that this is equivalent to solving a function with a singular fixed point. Additionally, this equation is an algorithm in itself to compute the reward and can be easily translated into code.

In this lecture, we will analyze what happens when the whole distribution of rewards is considered rather than just the expected reward, a setting called distributional RL.

2 Banach Fixed Point Theorem

In this section we discuss a central theorem in the analysis of RL algorithms: the Banach fixed point theorem[1].
The Banach fixed point theorem gives convergence guarantees to a unique fixed point under iterated contraction mappings. Thus, if the update equation of an RL algorithm can be shown to be a contraction mapping, it will eventually converge. Moreover, regardless of initial values, convergence leads to the same fixed point (per uniqueness).

Definition 6 (Contraction Mapping)**.** *Let* (X, d) *be a metric space with metric* d *on space* X*. Then a function* $T : X \to X$ *is a contraction mapping on* X *if there exists* $q \in [0, 1)$ *such that*

$$\forall x, y \in X : d(T(x), T(y)) \leq q d(x, y)$$

Definition 6 is reminiscent of the definition of L-Lipschitz functions. Indeed, a contraction mapping is simply an L-Lipschitz function (where the image is a subset of the domain) for some $L \in [0, 1)$. That is, beyond restricting the maximum growth rate of the function, we also have that the function must grow arbitrarily slower as the function is repeatedly applied to its images. Figure 1 illustrates iteration of the contraction mapping $f(x) = 0.5x$.

[1]The derivations and proofs are mostly taken from Wikipedia [2]

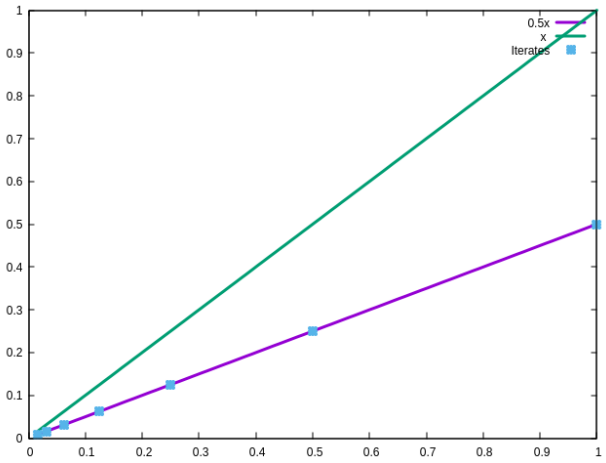

Figure 1: Iterates of the simplest contraction mapping $f(x) = \alpha x$ with $\alpha \leq 1$ (here, $\alpha = 0.5$) at starting point $x_0 = 1$ tend toward the fixed point $f(x) = x$ at $(0, 0)$. Green: the line $f(x) = x$. Blue: the contraction mapping $f(x) = 0.5x$, respecting $d(0.5x, 0.5y) \leq 0.5d(x, y)$ with $d(x, y) = |y - x|$. Points: iterates of $f(x)$ with $x_0 = 1$.

Theorem 7 (Banach Fixed Point Theorem)**.** *Let (X, d) be a non-empty complete metric space with a contraction mapping $T : X \to X$, then T admits a unique fixed point $x^* \in X$. Moreover, for any $x_0 \in X$, $T^n(x_0) \to x^*$ as $n \to \infty$*

Proof. Let $x_0 \in X$ be an arbitrary point and let $\{x_n\}$ be the sequence of iterates such that $\forall n \in \mathcal{N}^+, x_n = T(x_{n-1})$. Then we have

$$\begin{aligned} d(x_{n+1}, x_n) &\leq q d(x_n, x_{n-1}) \\ &\leq \dots \\ &\leq q^n d(x_1, x_0) \end{aligned}$$

Let $m, n \in \mathcal{N}^+$ such that $m > n$, then, because (X, d) is a metric space and T is a contraction mapping on X,

$$\begin{aligned} d(x_m, x_n) &\leq d(x_m, x_{m-1}) + \dots + d(x_{n+1}, x_n) \\ &\leq q^{m-1} d(x_1, x_0) + \dots + q^n d(x_1, x_0) \\ &= q^n d(x_1, x_0) \sum_{k=0}^{m-n-1} q^k \\ &\leq q^n d(x_1, x_0) \sum_{k=0}^{\infty} q^k \\ &= \frac{q^n}{1-q} d(x_1, x_0) \end{aligned}$$

For some arbitrary ϵ, since $q \in [0, 1)$, there is some N such that

$$q^N < \frac{\epsilon(1-q)}{d(x_1, x_0)}$$

Choosing $m, n > N$ gives

$$d(x_m, x_n) \leq \frac{q^n}{1-q} d(x_1, x_0) < \frac{\epsilon(1-q) d(x_1, x_0)}{(1-q) d(x_1, x_0)} = \epsilon$$

Thus, $\{x_n\}$ is a Cauchy sequence with a limit $x^* \in X$, and it is a fixed point of T because

$$\begin{aligned} x^* &= \lim_{n\to\infty} x_n \\ &= \lim_{n\to\infty} T(x_{n-1}) \\ &= T(\lim_{n\to\infty} x_{n-1}) && \text{Valid because } T \text{ is continuous} \\ &= T(x^*) \end{aligned}$$

Finally, it can be shown by contradiction that the fixed point is unique. Suppose there exists $p_1, p_2 \in X$ such that $p_1 \neq p_2$ and p_1 and p_2 are fixed points of T. Then

$$\begin{aligned} d(T(p_1), T(p_2)) = d(p_1, p_2) &\leq q d(p_1, p_2) \\ &\iff q = 0 && q \in [0, 1) \\ &\implies d(p_1, p_2) \leq 0 \end{aligned}$$

Since $d(p_1, p_2) \geq 0$ by definition, this implies $d(p_1, p_2) = 0 \implies p_1 = p_2$, which is a contradiction. □

Theorem 7 tells us that given any point $x_0 \in X$, repeated application of the contraction mapping not only leads to convergence, but also converges to the unique fixed point $x^* \in X$. This will be used in the next section to demonstrate convergence of distributional RL.

3 Distributional RL

Distributional Reinforcement Learning was first introduced in Bellemare et al. [1]. We will discuss it in the context of infinite horizon. First we define The following are all the random variables involved in the MDP:

1. $r_t, s_t \sim \mathrm{P}(\cdot, \cdot \,|s_{t-1}, a_{t-1})$ (non-deterministic reward and transition)
2. $a_t \sim \pi(\cdot \,|s_{t-1})$ (non-deterministic policy)

Using these variables, we can define what is known as the return:

Definition 8 (Return)**.** *Let $(s_t, a_t)_{t=1}^{\infty}$ be the state-action pairs of an infinite horizon MDP with a discount factor $0 < \gamma \leq 1$. The **return** is defined as*

$$\Phi^\pi(s, a) = r_0 + \sum_{t=1}^{\infty} \gamma^t r_t \quad \textit{with } s_0 = s \textit{ and } a_0 = a \tag{1}$$

which is a function that takes an initial state and action pair as input and output a random variable on the initial state and action.

As return, $\Phi^\pi(s, a)$, is a random variable, it has a law, or a distribution. Let $\mathcal{D}$ denote **the space of all distributions of return**. We define the value distribution as a mapping from the state-action space to the space of distributions of return below.

Definition 9 (Value Distribution)**.** *Then we define the value distribution as*

$$Z^\pi : \mathcal{S} \times \mathcal{A} \to \mathcal{D}$$

such that

$$\Phi^\pi(s, a) \sim Z^\pi(s, a)$$

$\mathcal{D}^{|\mathcal{S}|\times|\mathcal{A}|}$*. Since $Z^\pi(a, s)$ is a distribution, we can write in the discrete case*

$$P(\Phi = \phi) = Z^\pi(a, s)(\phi)$$

It is important to understand that for given a and s, $\Phi^\pi(s,a)$ is a random variable and $Z^\pi(s,a)$ is the distribution of that random variable. The return and value distributions are linked to the state-action value function described last lecture via the formula

$$Q^\pi(s,a) := \mathbb{E}_{Z^\pi(s,a)}[\Phi(s,a)] \tag{2}$$

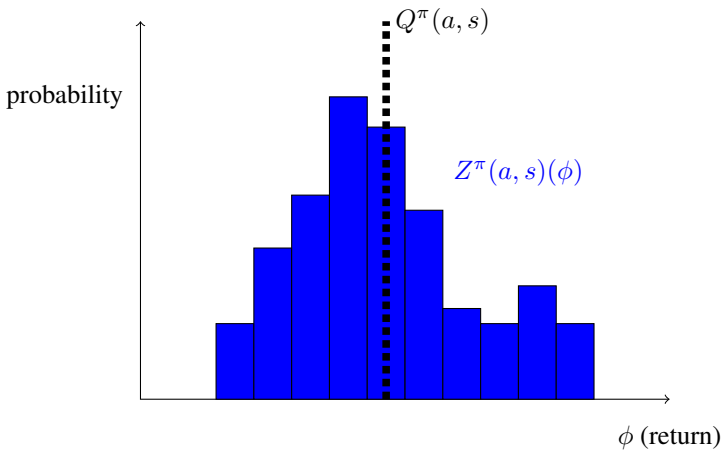

Figure 2: example of distribution $Z^\pi(a,s)(\phi)$

But why would we want to use the distribution instead of its expectation. The problem with expectation is that we lose information about the random variables.

For example lets take two dice (A, B) and play a game. The user must choose a die and throw it once. If die A lands on an even number, the player gets 1 dollar and he loses 1 otherwise. If die B land on $1, 2, 3, 4, 5$ he gets 20 dollars and loses 100 dollars if it lands on 6. Both dice have an expectation of 0, but the two choices involve different levels of risk as they have different distributions. This risk cannot be characterized by simply looking at the expectation because both dice are indistinguishable in the sense of expectation, which does not reflect the other statistics of the distribution (such as variance).

Before anything else, we define some algebraic operations between distributions:

Definition 10. *Let Z_1, Z_2 be two independant univariate distributions and let $\gamma \in \mathbb{R}$ be a constant We define:*

1. $(Z_1 + Z_2)(x) := (Z_1 * Z_2)(x)$ *(convolutionnal product)*
2. $(\gamma\, Z_1)(x) := \frac{1}{\gamma} Z_1(\gamma x)$ *(contracting the function horizontally)*
3. $(\gamma + Z_1)(x) := Z_1(x - \gamma)$ *(horizontal translation)*

Now we turn to policy evaluation using distributional RL. The policy π is fixed and we want to find the value distribution $Z^\pi(s,a)$ for a given (s,a) pair. First we define the transition operator:

Definition 11 (Transition Operator)**.**

$$P^\pi : \mathcal{D} \to \mathcal{D}$$
$$P^\pi Z(s,a) = Z(S', A') \tag{3}$$

where $S' \sim P(.|s,a)$ and $A' \sim \pi(.|S')$. Capital letter are used to emphasize the randomness of the new state and action.

The output of this operator can be seen as a mixture distribution with weight $P(S' = s', A' = a'|a,s)$ for the each component distribution $Z^\pi(s',a')$. In what follows, we assume the randomness in the reward and the transition P^π are independent. Another operator that we define below is the distributional Bellmann operator:

Definition 12 (Distributional Bellmann operator)**.**

$$\mathcal{T}^\pi : \mathcal{D} \to \mathcal{D}$$

$$\mathcal{T}^\pi Z(s,a) = Distr(r) + \gamma \mathcal{P}^\pi Z(s,a) \tag{4}$$

where Dist(r) is the distribution of rewards at a given state s and by doing the action a. We could also change the input and output spaces of the operator so it acts on random variables instead of their distributions

$$\mathcal{T}^\pi \Phi^{\pi'}(s,a) \overset{D}{=} r + \gamma \mathcal{P}^\pi \Phi^{\pi'}(s,a) \tag{5}$$

These two definitions are equivalent (in the sense of distribution).

Here is a graphical explanation of the operator with deterministic reward:

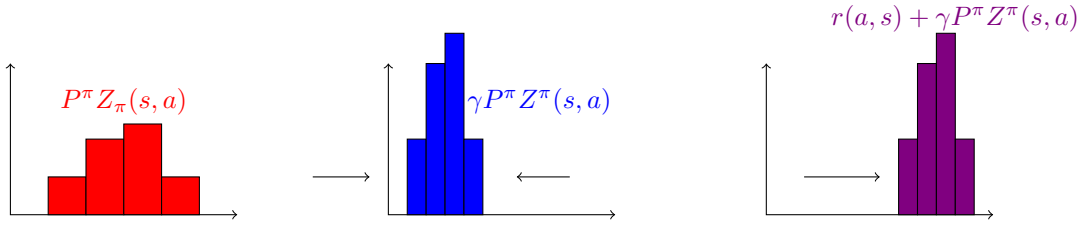

Figure 3: Illustration of how $\mathcal{T}^\pi$ operates on a distribution

The final operation is simply a translation because r is deterministic in that example. In a more general case, the rightmost distribution would have been obtained with a convolutional product. It is important to note that $\mathcal{T}^\pi$ resembles the Bellman equations for expected reward. In the spirit of policy evaluation, we want to show that the operator is a contraction mapping with respect to some metric. Let's first introduce the Wasserstein metric between distributions.

Definition 13. *Let Z_1 and Z_2 be two distributions; i.e. $Z_1, Z_2 \in \mathcal{D}$. For $p \geq 1$, the p-Wassertein metric is defined as*

$$d_p(Z_1, Z_2) = \inf_{D \in \prod(Z_1, Z_2)} \mathbb{E}_{(z_1,z_2)\sim D}[||z_1 - z_2||_p^p]^{\frac{1}{p}} \tag{6}$$

where $\prod(Z_1, Z_2)$ is the set of all joint distributions with marginals Z_1 and Z_2. The metric has the following properties (for $\gamma \in \mathbb{R}$ and a, a random variable independent of z_1 and z_2 that follows distribution A):

1. $d_p(\gamma Z_1, \gamma Z_2) \leq |\gamma| d_p(Z_1, Z_2)$
2. $d_p(A + Z_1, A + Z_2) \leq d_p(Z_1, Z_2)$
3. $d_p(AZ_1, AZ_2) \leq ||a||_p d_p(Z_1, Z_2)$

Note that the Wasserstein metrics are metrics of distributions, whereas value distributions are mappings (from the space of state-action pairs to the space of distributions), the former are not yet metrics of the latter. Let $\mathcal{Z}$ denote the **space of value distributions** (with bounded moments). Let us define a uniform form of the Wasserstein distance as

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

for $Z_1, Z_2 \in \mathcal{Z}$. Then we can establish the following result.

Lemma 14. *$\bar{d}_p$ is a metric over value distributions.*

The only nontrivial part to prove is triangle inequality of a metric.

Proof. For $Y \in \mathcal{Z}$, we have

$$\begin{aligned}
\bar{d}_p(Z_1, Z_2) &= \sup_{s,a} d_p(Z_1(s,a), Z_2(s,a)) \\
&\leq \sup_{s,a} d_p(Z_1(s,a), Y(s,a)) + d_p(Y(s,a), Z_2(s,a)) \\
&\leq \sup_{s,a} d_p(Z_1(s,a), Y(s,a)) + \sup_{s,a} d_p(Y(s,a), Z_2(s,a)) \\
&= \bar{d}_p(Z_1, Y) + \bar{d}_p(Y, Z_2)
\end{aligned}$$

where the first inequality is because d_p is a metric which admits triangle inequality over the space of distributions. □

Now consider the metric space $(\mathcal{Z}, \bar{d}_p)$. Considering the iterative process $Z_{k+1} := \mathcal{T}^\pi Z_k$ with some initial value distirbution $Z_0 \in \mathcal{Z}$, we now show that "distributional" Bellman operator is a contraction mapping.

Lemma 15. *$\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ is a γ-contraction in $\bar{d}_p$.*

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

$$\begin{aligned}
\bar{d}_p(\mathcal{T}^\pi(Z_1), \mathcal{T}^\pi(Z_2)) &= \sup_{s,a} d_p(\mathcal{T}^\pi(Z_1), \mathcal{T}^\pi(Z_2)) \\
&= \sup_{s,a} d_p(Distr(r; s, a) + \gamma P^\pi Z_1(s,a), Distr(r; s, a) + \gamma P^\pi Z_2(s,a)) \\
&\leq \sup_{s,a} \gamma d_p(P^\pi Z_1(s,a), P^\pi Z_2(s,a)) \\
&\leq \sup_{s,a} \gamma \sup_{s',a'} d_p(Z_1(s',a'), Z_2(s',a')) \\
&= \gamma \bar{d}_p(Z_1, Z_2)
\end{aligned}$$

where the first two lines are just the definitions of $\bar{d}_p$ and $\mathcal{T}^\pi$; the third line is due to the properies of Wasserstein distance (Definition 13) and the (conditional) independence of reward and the transition; the fourth line is due to taking the sup rather than taking a random next state-action pair. □

By construction, Z^π is a fixed point of the Bellman equation, and by the Banach fixed point theorem, the sequence $(Z_k)_{k \geq 1}$ will converge in $\bar{d}_p$ to Z^π.

4 Summary

In this lecture, we saw the infinite horizon variant of RL. We introduced the Banach fixed point theorem, a central theorem to many convergence results in and outside RL algorithms, and used it to demonstrate convergence for distributed RL methods.

References

[1] M. G. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 449–458. JMLR.org, 2017.

[2] Wikipedia. Wikipedia. `https://en.wikipedia.org/wiki/Banach_fixed-point_theorem`, April 2019.